Accepted for publication in the Journal of Automata, Languages and Combinatorics
Final version created on May 29, 2021

THE PREFIX AUTOMATON

Sabine Broda[(A)] Eva Maia[(B)] Nelma Moreira[(A)] Rogério Reis[(A)]

[(A)] *CMUP& DCC, Faculdade de Ciências da Universidade do Porto*
Rua do Campo Alegre, 4169-007 Porto, Portugal
{sabine.broda,nelma.moreira,rogerio.reis}@fc.up.pt

[(B)] *GECAD - Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development, ISEP*
Rua Dr. António Bernardino de Almeida, 431, 4249-015 Porto, Portugal
egm@isep.ipp.pt

ABSTRACT

There are many different constructions when converting regular expressions to finite automata. In this paper we focus on the prefix automaton, $\mathcal{A}_{\mathrm{Pre}}$, introduced by Yamamoto in 2014. We present two different methods for the construction of $\mathcal{A}_{\mathrm{Pre}}$. First, an inductive one, based on a system of expression equations. A second one using an iterative function for computing the states and transitions. We establish relationships between $\mathcal{A}_{\mathrm{Pre}}$ and other constructions, such as the position automaton, partial derivative automaton and their double reversal (dual) counterparts. We study the average size of these constructions, both experimentally and from an analytic combinatorics point of view. Finally, we extend the construction of the prefix automaton to regular expressions with intersection and show that the relationships with the other automaton constructions also hold for these expressions.

Keywords: regular expressions, nondeterministic finite automata, prefix automata, average complexity, regular expressions with intersection

1. Introduction

Conversions from regular expressions to equivalent nondeterministic finite automata can be with or without spontaneuos (ε) transitions. The classic construction with ε transitions is the Thompson construction ($\mathcal{A}_{\varepsilon\text{-T}}$) [20], while the Glushkov/position automaton is a standard ε-free construction ($\mathcal{A}_{\mathrm{POS}}$) [14]. It is well known that if ε transitions are eliminated from the Thompson automaton, the result is the Glushkov automaton [13]. In 2014, Yamamoto [21] presented a new construction of an ε-free automaton starting from the Thompson automaton. For that, each state s of $\mathcal{A}_{\varepsilon\text{-T}}$ was labelled with two regular expressions, one corresponding to the left language of the state, $\mathsf{LP}(s)$, and the other to its right language, $\mathsf{LS}(s)$. Merging states with

[(A)] Work partially supported by CMUP (UIDB/00144/2020), which is funded by FCT- Fundação para Ciência e Tecnologia, I.P.

the same LP label leads to the prefix automaton, and with the same LS leads to the suffix automaton. While the suffix automaton corresponds to the partial derivative automaton ($\mathcal{A}_{\mathrm{PD}}$) which has been well studied [18, 1, 9, 10, 5], the prefix automaton was not studied before (as far as we know). Yamamoto's final automaton was obtained as follows: first constructing one of these automata; then for the states of the original $\mathcal{A}_{\varepsilon\text{-}\mathrm{T}}$ that were not joined, i.e., their equivalence class w.r.t that labelling was a singleton, the possible mergings w.r.t. the other labelling were taken in consideration.

In this paper we further study the prefix automaton ($\mathcal{A}_{\mathrm{Pre}}$) and consider relationships between this automaton and other ε-free constructions, such as the position automaton, the partial derivative automaton and their double reversal (dual) counterparts. We also study the average size of these constructions, experimentally and from an analytic combinatorics point of view. Finally, we extend the prefix automaton construction to regular expressions with intersection and show that the relationships with the other automaton constructions also hold for these expressions. This paper expands and revises some results that appeared in Maia et al. [17]. In particular, most results in Sections 3.1, 4.3, 4.4, and 5 have been stated therein without proofs, which are provided here as well as some new results. The extension of the prefix automaton for expressions with intersection in Section 6 is completely new. Broda et al. [3] presented a taxonomy of conversions from (standard) regular expressions to equivalent deterministic and nondeterministic finite automata. In particular, the determinisation of the prefix automaton and its double reversal ($\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}$) were placed in the conversion's taxonomy. Noticeable the determinisation of $\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}$ is the smallest automaton among the studied constructions in that paper.

We now summarise our contributions and the structure of the paper. The next section recalls some basic notions on regular expressions and finite automata. In Section 3, we first define the inductive construction of the prefix automaton using a system of left expression equations from [17] and prove its correctness. This parallels with the Mirkin's construction for the partial derivative automaton and a similar construction for the right-partial derivative automaton ($\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$) (which are recalled in Appendix A and Appendix B, following Maia [16]). Note that these constructions are essential to obtain average case results using analytic combinatorics. Section 3.2 presents a new iterative definition of $\mathcal{A}_{\mathrm{Pre}}$ which is shown to coincide with the inductive definition given before. In Section 4 the prefix automaton $\mathcal{A}_{\mathrm{Pre}}$ is shown to be a quotient of the position automaton $\mathcal{A}_{\mathrm{POS}}$ and it is related with the position automaton dual $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}$, as well as with $\mathcal{A}_{\mathrm{PD}}$ and $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ [17]. In Sections 4.1, 4.2 and 4.3 we start by reviewing these four constructions. In particular, we relate $\mathcal{A}_{\mathrm{PD}}$ with $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ and $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}$. The average size complexity of $\mathcal{A}_{\mathrm{Pre}}$ is studied in Section 5 [17]. First some experimental results are presented that compare the sizes of $\mathcal{A}_{\mathrm{POS}}$, $\mathcal{A}_{\mathrm{PD}}$, $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ and $\mathcal{A}_{\mathrm{Pre}}$ obtained from uniformly random generated regular expressions. Using the framework of analytic combinatorics, we estimate a lower bound for the number of mergings of states that arise when computing $\mathcal{A}_{\mathrm{Pre}}$ from $\mathcal{A}_{\mathrm{POS}}$. Recently, the partial derivative automaton and the position automaton were extended to regular expressions with the intersection operator [2, 8]. In Section 6 we extend the prefix automaton for those expressions and show that it is also in this case a quotient of the position automaton. Section 7 concludes with some final remarks.

2. Preliminaries

Given an alphabet $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_k\}$ of size k, the set RE of *regular expressions* α over Σ is defined by the following grammar:

$$\alpha := \emptyset \mid \varepsilon \mid \sigma_1 \mid \cdots \mid \sigma_k \mid (\alpha + \alpha) \mid (\alpha \cdot \alpha) \mid (\alpha^\star), \tag{1}$$

where the $\cdot$ is often omitted. If two α and β are syntactically equal, we write $\alpha \doteq \beta$. We denote by Σ_α the alphabet consisting of the letters that occur in α. The *size* of a regular expression α, $|\alpha|$, is its number of symbols, disregarding parentheses, and its *alphabetic size*, $|\alpha|_\Sigma$, is its number of letters from Σ. The language represented by a regular expressions α is denoted by $\mathcal{L}(\alpha)$. Given a set S of regular expressions let $\mathcal{L}(S) = \cup_{\alpha \in S}\mathcal{L}(\alpha)$. Two regular expressions α and β are *equivalent* if $\mathcal{L}(\alpha) = \mathcal{L}(\beta)$, and we write $\alpha = \beta$. We define the function ε by $\varepsilon(\alpha) = \varepsilon$ if $\varepsilon \in \mathcal{L}(\alpha)$ and $\varepsilon(\alpha) = \emptyset$, otherwise. This function can be naturally extended to sets of regular expressions and languages. If a set of expressions is a singleton $\{\alpha\}$, the parenthesis may be omitted. We consider regular expressions reduced by the following rules:

$$\begin{aligned}
&\varepsilon\alpha = \alpha = \alpha\varepsilon,\\
&\emptyset + \alpha = \alpha = \alpha + \emptyset,\\
&\emptyset\alpha = \emptyset = \alpha\emptyset,\\
&\emptyset^\star = \varepsilon.
\end{aligned}$$

In particular, $\emptyset$ does not occur in any expression but $\emptyset$ itself. The same rules apply if α is substituted by a set of expressions. Given a language $\mathcal{L} \subseteq \Sigma^\star$ and $w \in \Sigma^\star$, the *left quotient* of $\mathcal{L}$ w.r.t. w is $w^{-1}\mathcal{L} = \{\, x \mid wx \in \mathcal{L}\,\}$, and the *right quotient* of $\mathcal{L}$ w.r.t. w is the language $\mathcal{L}w^{-1} = \{\, x \mid xw \in \mathcal{L}\,\}$. The *reversal* of a word $w = \sigma_1\sigma_2\cdots\sigma_n$ is $w^R = \sigma_n\cdots\sigma_2\sigma_1$. The *reversal* of a language $\mathcal{L}$, denoted by $\mathcal{L}^R$, is the set of words whose reversal is in $\mathcal{L}$. We have $\mathcal{L}w^{-1} = ((w^R)^{-1}\mathcal{L}^R)^R$. The reversal of a regular expression α is denoted by α^{R}, and is inductively defined by: $\alpha^{\mathrm{R}} = \alpha$ for $\alpha \in \Sigma \cup \{\varepsilon, \emptyset\}$, $(\alpha+\beta)^{\mathrm{R}} = \beta^{\mathrm{R}} + \alpha^{\mathrm{R}}$, $(\alpha\beta)^{\mathrm{R}} = \beta^{\mathrm{R}}\alpha^{\mathrm{R}}$ and $(\alpha^\star)^{\mathrm{R}} = (\alpha^{\mathrm{R}})^\star$. The reversal expression α^{R} describes $\mathcal{L}(\alpha)^{\mathrm{R}}$.

A *nondeterministic finite automaton* (NFA) is a five-tuple $A = (Q, \Sigma, \delta, I, F)$ where Q is a finite set of states, Σ is a finite alphabet, $I \subseteq Q$ is the set of initial states, $F \subseteq Q$ is the set of final states, and $\delta : Q \times \Sigma \to 2^Q$ is the transition function. The transition function can be extended to words and to sets of states in the natural way. When $I = \{q_0\}$, we use $I = q_0$. If $|Q| = n$ we can consider $Q = [0, n-1]$. Given a state $q \in Q$, the *right language* of q is $\mathcal{L}_q(A) = \{\, w \in \Sigma^\star \mid \delta(q,w) \cap F \neq \emptyset\,\}$, and the *left language* is $\overleftarrow{\mathcal{L}}_q(A) = \{\, w \in \Sigma^\star \mid q \in \delta(I,w)\,\}$. The language accepted by A is defined by $\mathcal{L}(A) = \bigcup_{q\in I}\mathcal{L}_q(A)$. Two NFAs are *equivalent* if they accept the same language. If two NFAs A and B are isomorphic, we write $A \simeq B$.

Given an automaton $A = \langle Q, \Sigma, \delta, I, F\rangle$ its reversal is $A^{\mathrm{R}} = \langle Q, \Sigma, \delta^{\mathrm{R}}, F, I\rangle$, where for $q \in Q$, $\sigma \in \Sigma$, we have $\delta^{\mathrm{R}}(q,\sigma) = \{\, p \mid q \in \delta(p,\sigma)\,\}$ and $\mathcal{L}(A^{\mathrm{R}}) = \mathcal{L}(A)^{\mathrm{R}}$.

The right languages $\mathcal{L}_i$, for $i \in Q = [0, n-1]$, define a system of right equations,

$$\mathcal{L}_i = \bigcup_{j=1}^{k} \sigma_j \left(\bigcup_{m \in I_{ij}} \mathcal{L}_m \right) \cup \varepsilon(\mathcal{L}_i),$$

where $I_{ij} \subseteq [0, n-1]$, $m \in I_{ij} \iff m \in \delta(i, \sigma_j)$, and $\mathcal{L}(A) = \bigcup_{i \in I} \mathcal{L}_i$. In the same manner, the left languages of the states of A define a system of left equations

$$\overleftarrow{\mathcal{L}}_i = \bigcup_{j=1}^{k} \left(\bigcup_{m \in I_{ij}} \overleftarrow{\mathcal{L}}_m \right) \sigma_j \cup \varepsilon(\overleftarrow{\mathcal{L}}_i),$$

where $I_{ij} \subseteq [0, n-1]$, $m \in I_{ij} \iff i \in \delta(m, \sigma_j)$, and $\mathcal{L}(A) = \bigcup_{i \in F} \overleftarrow{\mathcal{L}}_i$.

An equivalence relation $\equiv$ on Q is *right invariant* w.r.t. an NFA A if it satisfies

- $\equiv \subseteq (Q \setminus F)^2 \cup F^2$ and
- $\forall p, q \in Q$, if $p \equiv q$, then $\forall \sigma \in \Sigma, p' \in \delta(p, \sigma)\ \exists q' \in \delta(q, \sigma)$ such that $p' \equiv q'$.

Given a set of states $S \subseteq Q$, we denote $S/\equiv\ = \{\, [q] \mid q \in S \,\}$. Note that $p \equiv q$ implies $\delta(p, \sigma)/\equiv\ = \delta(q, \sigma)/\equiv$, for $p, q \in Q$ and $\sigma \in \Sigma$. If $\equiv$ is a right-invariant relation on Q, the *quotient automaton* $A/\equiv$ is given by $A/\equiv\ = \langle Q/\equiv, \Sigma, \delta/\equiv, I/\equiv, F/\equiv \rangle$, where $\delta/\equiv([p], \sigma) = \{\, [q] \mid q \in \delta(p, \sigma) \,\} = \delta(p, \sigma)/\equiv$. We have $\mathcal{L}\left(A/\equiv\right) = \mathcal{L}(A)$. In the same way, an equivalence relation $\equiv$ on Q is *left invariant* w.r.t. A if

- $\equiv \subseteq (Q \setminus I)^2 \cup I^2$ and
- $\forall p, q \in Q$, if $p \equiv q$, then $\forall \sigma \in \Sigma, p' \in \delta^{\mathrm{R}}(p, \sigma)\ \exists q' \in \delta^{\mathrm{R}}(q, \sigma)$ such that $p' \equiv q'$.

It follows that,

Lemma 1. *A relation $\equiv$ is a left-invariant relation w.r.t. an automaton A if and only if it is a right-invariant relation w.r.t. A^R.*

3. The Prefix Automaton

Yamamoto [21] presented a new algorithm for converting a regular expression into an equivalent ε-free NFA. First, a labelled version of the usual Thompson automaton [20] is obtained, where each state q is labelled with two regular expressions, one that corresponds to its left language, $\mathsf{LP}(q)$, and the other to its right language, $\mathsf{LS}(q)$. States whose in-transitions are labelled with a letter are called *sym-states.* In the set of sym-states two equivalence relations $\equiv_{pre}$ and $\equiv_{suf}$ are defined such that for two states p and q one has $p \equiv_{pre} q$ if and only if $\mathsf{LP}(p) \doteq \mathsf{LP}(q)$; and $p \equiv_{suf} q$ if and only if $\mathsf{LS}(p) \doteq \mathsf{LS}(q)$. Then the author defines the *prefix automaton* $\mathcal{A}_{\mathrm{Pre}}$ and the *suffix automaton* $\mathcal{A}_{\mathrm{Suf}}$ whose sets of states are the equivalence classes of $\equiv_{pre}$ and $\equiv_{suf}$, respectively. The final automaton was a combination of these two. He also shows that $\mathcal{A}_{\mathrm{Suf}}$ coincides with the partial derivatives automaton, $\mathcal{A}_{\mathrm{PD}}$.

In this section we present two different methods for the construction of $\mathcal{A}_{\mathrm{Pre}}$. First, an inductive one, based on a system of expression equations. A second one using an iterative function for computing the states and transitions.

3.1. Inductive Definition of $\mathcal{A}_{\text{Pre}}$

In the following we present Yamamoto's construction for the prefix automaton using a system of left expression equations. The set of states of $\mathcal{A}_{\text{Pre}}(\alpha)$ is $\mathsf{S}(\alpha) \cup \{\alpha_0\}$, where $\alpha_0 \doteq \varepsilon$ and $\mathsf{S}(\alpha)$ is inductively defined by the following equations.

$$\begin{aligned} \mathsf{S}(\emptyset) = \mathsf{S}(\varepsilon) &= \emptyset, & \mathsf{S}(\sigma) &= \{\sigma\}, \\ \mathsf{S}(\alpha+\alpha') &= \mathsf{S}(\alpha) \cup \mathsf{S}(\alpha'), & \mathsf{S}(\alpha\alpha') &= \alpha\,\mathsf{S}(\alpha') \cup \mathsf{S}(\alpha), \\ \mathsf{S}(\alpha^\star) &= \alpha^\star\,\mathsf{S}(\alpha), \end{aligned} \tag{2}$$

where for any $S \subseteq \mathsf{RE}$ we have $S\emptyset = \emptyset S = \emptyset$, $S\varepsilon = \varepsilon S = S$, and $\alpha' S = \{\, \alpha'\alpha \mid \alpha \in S \,\}$ for $\alpha' \in \mathsf{RE} \setminus \{\emptyset, \varepsilon\}$ (and analogously for $S\alpha'$).

Proposition 2. *The set $\mathsf{S}(\alpha) = \{\alpha_1, \dots, \alpha_n\}$ satisfies a system of equations of the form*

$$\alpha_1 = X_1\sigma_{\ell_1}, \dots, \alpha_n = X_n\sigma_{\ell_n}$$

such that X_i are linear combinations of elements of $\{\alpha_0, \alpha_1, \dots, \alpha_n\}$, where $\alpha_0 \doteq \varepsilon$ and $i \in [1,n]$, $\ell_i \in [1,k]$, and $n \geq 0$. Moreover, we have

$$\alpha = \sum_{i \in I \subseteq [0,n]} \alpha_i.$$

Proof. We will prove by structural induction on α that whenever $\mathsf{S}(\alpha) \neq \emptyset$, then the set $\mathsf{S}(\alpha)$ satisfies a system of equations of the form $\alpha_i = X_i\sigma_{\ell_i}$ where $\alpha_0 \doteq \varepsilon$, and such that X_i are linear combinations of elements of $\{\alpha_0, \alpha_1, \dots, \alpha_n\}$, for all $i \in [1,n]$, $\ell_i \in [1,k]$, and $n \geq 0$. Furthermore, $\alpha_0 \doteq \varepsilon$ is a component of at least one X_i. For the cases $\alpha \doteq \emptyset$ or $\alpha \doteq \varepsilon$ we have $S(\alpha) = \emptyset$ and there is nothing to prove. For $\alpha \doteq \sigma$, we have $\mathsf{S}(\alpha) = \{\sigma\} = \{\alpha_1\}$, which satisfies the following set of equations.

$$\begin{aligned} \alpha &= \alpha_1, \\ \alpha_1 &= \alpha_0\sigma, \text{ where } \alpha_0 \doteq \varepsilon. \end{aligned}$$

Now, consider

$$\begin{aligned} \beta &= \sum_{i \in I \subseteq [0,n]} \beta_i, \\ \beta_i &= X_i\sigma_{\ell_i},\ \ell_i \in [1,k], \text{ for all } i \in [1,n] \text{ with } \mathsf{S}(\beta) = \{\beta_1, \dots, \beta_n\} \text{ and } \beta_0 \doteq \varepsilon, \end{aligned}$$

and

$$\begin{aligned} \gamma &= \sum_{i \in I' \subseteq [0,m]} \gamma_i, \\ \gamma_i &= Y_i\sigma_{\ell_i},\ \ell_i \in [1,k], \text{ for all } i \in [1,m] \text{ with } \mathsf{S}(\gamma) = \{\gamma_1, \dots, \gamma_m\} \text{ and } \gamma_0 \doteq \varepsilon. \end{aligned}$$

Let $\alpha \doteq \beta + \gamma$, then

$$\beta + \gamma = \sum_{i \in I \subseteq [1,n]} \beta_i + \sum_{i \in I' \subseteq [1,m]} \gamma_i.$$

Consequently $\mathsf{S}(\alpha) = \{\beta_1, \ldots, \beta_n\} \cup \{\gamma_1, \ldots, \gamma_m\}$ satisfies the system containing the equations for β as well as for γ. The condition on $\alpha_0 \doteq \varepsilon \doteq \beta_0 \doteq \gamma_0$ follows from the induction hypothesis.

Consider $\alpha \doteq \beta\gamma$. Then

$$\begin{aligned} \beta\gamma &= \beta(\sum_{i \in I' \subseteq [0,m]} \gamma_i), \\ &= \begin{cases} \beta(\sum_{i \in I' \subseteq [1,m]} \gamma_i),, & \text{if } 0 \notin I', \\ \beta(\sum_{i \in I' \subseteq [1,m]} \gamma_i) + \sum_{i \in I \subseteq [0,n]} \beta_i & \text{if } 0 \in I', \end{cases} \end{aligned}$$

and $\beta\gamma_i = \beta(Y_i \sigma_{\ell_i})$. We know that ε is a component of at least one of the Y_i for $i \in [0, m]$. Consequently, $\mathsf{S}(\alpha) = \{\beta\gamma_1, \ldots, \beta\gamma_m\} \cup \{\beta_1, \ldots, \beta_n\}$ satisfies the system containing the equations $\beta\gamma_i = \beta Y_i \sigma_{\ell_i}$, as well as the equations for β. By the induction hypothesis, ε is a component of at least one of the X_i for $i \in [1, n]$.

Consider $\alpha \doteq \beta^\star$ with $\mathsf{S}(\alpha) = \{\beta^\star\beta_1, \ldots, \beta^\star\beta_n\}$. Then,

$$\begin{aligned} \beta^\star &= \beta^\star\beta + \varepsilon, \\ &= \beta^\star(\sum_{i \in I \subseteq [1,n]} \beta_i) + \varepsilon. \end{aligned}$$

and

$$\beta^\star\beta_i = \beta^\star(X_i \sigma_{\ell_i}), \quad \text{for all } i \in [1, n].$$

Each $\beta^\star X_i$ can be written as a sum of elements of $\mathsf{S}(\alpha) \cup \{\beta^\star\beta_0\}$. We have that

$$\beta^\star\beta_0 = \beta^\star = \beta^\star(\sum_{i \in I \subseteq [1,n]} \beta_i) + \varepsilon,$$

is also a linear combination of elements of $\mathsf{S}(\alpha) \cup \{\alpha_0\}$, where $\alpha_0 \doteq \varepsilon$. Since β_0 appears in at least one X_i the condition on $\alpha_0 \doteq \varepsilon$ is also satisfied. □

It is easy to see that $\mathsf{S}(\alpha)$ is always finite and contains at most $|\alpha|_\Sigma$ elements. Moreover every element in $\mathsf{S}(\alpha)$ is of the form $\alpha'\sigma$. The system of equations

$$\alpha_1 = X_1 \sigma_{\ell_1}, \ \ldots, \ \alpha_n = X_n \sigma_{\ell_n}$$

satisfied by $\mathsf{S}(\alpha)$ defines the automaton $\mathcal{A}_{\mathrm{Pre}}$, whose set of states is $\mathsf{S}(\alpha) \cup \{\varepsilon\}$. The left language of a state labelled with β is $\mathcal{L}(\beta)$. The initial state has label ε and there is a transition by σ_l from a state α_i to a state α_j if and only if α_i is a component of X_j (which we write as $\alpha_i \in X_j$) and $l = \ell_j$. The set of final states is

$$\mathsf{R}_\varepsilon(\alpha) = \bigcup_{i \in I} \{\alpha_i\},$$

where I is the set of indices such that

$$\alpha = \sum_{i \in I \subseteq [0,n]} \alpha_i.$$

The proof of Proposition 2 provides us with inductive methods for computing the set of final states $\mathsf{R}_\varepsilon(\alpha)$ and the set of transitions of $\mathcal{A}_{\mathrm{Pre}}(\alpha)$. For $\alpha \in \mathsf{RE}$, we have $\mathsf{R}_\varepsilon(\alpha) = \mathsf{R}(\alpha) \cup \varepsilon(\alpha)$, where the set $\mathsf{R}(\alpha)$ is computed by the following rules.

$$\begin{aligned} \mathsf{R}(\emptyset) &= \mathsf{R}(\varepsilon) = \emptyset, & \mathsf{R}(\sigma) &= \{\sigma\}, \\ \mathsf{R}(\alpha + \alpha') &= \mathsf{R}(\alpha) \cup \mathsf{R}(\alpha'), & \mathsf{R}(\alpha\alpha') &= \alpha\,\mathsf{R}(\alpha') \cup \varepsilon(\alpha')\,\mathsf{R}(\alpha), \\ \mathsf{R}(\alpha^\star) &= \alpha^\star\,\mathsf{R}(\alpha). \end{aligned} \tag{3}$$

It is easy to see that $\mathcal{L}(\alpha) = \mathcal{L}(\mathsf{R}_\varepsilon(\alpha))$. The set of outgoing transitions from the initial state ε is $\{\varepsilon\} \times \psi(\alpha)$, where

$$\psi(\alpha) = \{\, (\sigma_{\ell_i}, \alpha_i) \mid \varepsilon \in X_i \wedge i \in [1,n] \wedge \ell_i \in [1,k] \,\}$$

is inductively defined as $\mathsf{R}(\alpha)$ except for the following two cases,

$$\psi(\sigma) = \{(\sigma,\sigma)\} \quad \text{and} \quad \psi(\alpha\alpha') = \psi(\alpha) \cup \varepsilon(\alpha)\alpha\psi(\alpha').$$

In the above definition, the concatenation of an $\alpha \setminus \{\emptyset, \varepsilon\}$ with a tuple (σ,τ) is defined by $(\sigma,\tau)\beta = (\sigma,\tau\beta)$ and $\beta(\sigma,\tau) = (\sigma,\beta\tau)$. These definitions also extend to sets of tuples. The set of remaining transitions

$$\mathsf{T}(\alpha) = \{\, (\alpha_i, \sigma_{\ell_j}, \alpha_j) \mid \alpha_i \in X_j \wedge i,j \in [1,n] \wedge \ell_j \in [1,k] \,\}$$

satisfies the following inductive definition.

$$\begin{aligned} \mathsf{T}(\emptyset) &= \mathsf{T}(\varepsilon) = \mathsf{T}(\sigma) \ = \emptyset, \\ \mathsf{T}(\alpha + \alpha') &= \mathsf{T}(\alpha) \cup \mathsf{T}(\alpha'), \\ \mathsf{T}(\alpha\alpha') &= \mathsf{T}(\alpha) \cup \alpha\,\mathsf{T}(\alpha') \cup (\mathsf{R}(\alpha) \times (\alpha\psi(\alpha'))), \\ \mathsf{T}(\alpha^\star) &= \alpha^\star\,\mathsf{T}(\alpha) \cup \alpha^\star(\mathsf{R}(\alpha) \times \psi(\alpha)). \end{aligned} \tag{4}$$

Note that the result of the $\times$ operation is seen as a set of triples $(\alpha', \sigma, \beta')$. The concatenation of a transition (α,σ,β) with a regular expression $\gamma \in \mathsf{RE} \setminus \{\emptyset, \varepsilon\}$ is defined by $(\alpha,\sigma,\beta)\gamma = (\alpha\gamma,\sigma,\beta\gamma)$ and $\gamma(\alpha,\sigma,\beta) = (\gamma\alpha,\sigma,\gamma\beta)$. Moreover, we define $\emptyset(\alpha,\sigma,\beta) = (\alpha,\sigma,\beta)\emptyset = \emptyset$ and $\varepsilon(\alpha,\sigma,\beta) = (\alpha,\sigma,\beta)\varepsilon = (\alpha,\sigma,\beta)$. These definitions also extend to sets of transitions. Using the above, the *prefix automaton* for α is

$$\mathcal{A}_{\mathrm{Pre}}(\alpha) = \langle \mathsf{S}(\alpha) \cup \{\varepsilon\}, \Sigma, \{\varepsilon\} \times \psi(\alpha) \ \cup \ \mathsf{T}(\alpha), \varepsilon, \mathsf{R}_\varepsilon(\alpha) \rangle. \tag{5}$$

From Proposition 2 we conclude the following.

Proposition 3 [21]. $\mathcal{L}(\mathcal{A}_{\mathrm{Pre}}(\alpha)) = \mathcal{L}(\alpha)$.

Example 4. The prefix automaton $\mathcal{A}_{\mathrm{Pre}}$ for the expression $\alpha = (ab^\star + b)^\star a$ is depicted in Figure 1. We illustrate the computation of the set $S(\alpha)$ following the proof of Proposition 2. Let α_1 be $(ab^\star + b)^\star$. We have $S(a) = \{a\}$, with $a = \varepsilon \cdot a$, and $\mathsf{R}_\varepsilon(a) = \{a\}$. The same for $S(b) = \{b\}$. Next, we have $S(b^\star) = \{b^\star b\}$, with

$$\begin{aligned} b^\star b &= (b^\star b + \varepsilon)b, \\ b^\star &= b^\star b + \varepsilon, \end{aligned}$$

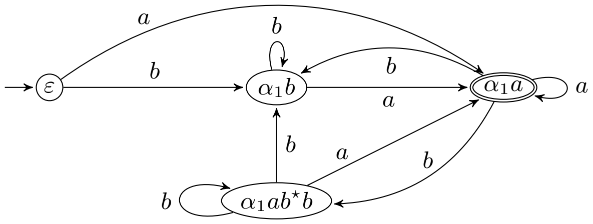


Figure 1: $\mathcal{A}_{\mathrm{Pre}}((ab^\star + b)^\star a)$, where $\alpha_1 = (ab^\star + b)^\star$.

and $\mathsf{R}_\varepsilon(b^\star) = \{b^\star b, \varepsilon\}$. Then $S(ab^\star) = \{ab^\star b, a\}$, with

$$\begin{aligned} ab^\star b &= (ab^\star b + a)b, \\ ab^\star &= ab^\star b + a, \end{aligned}$$

and $\mathsf{R}_\varepsilon(ab^\star) = \{ab^\star b, a\}$. Finally,

$$S(\alpha_1) = \{\alpha_1 ab^\star b, \alpha_1 a, \alpha_1 b\},$$

with

$$\begin{aligned} &\alpha_1 ab^\star b = (\alpha_1 ab^\star b + \alpha_1 a)b, \\ &\alpha_1 b = (\alpha_1 ab^\star b + \alpha_1 a + \alpha_1 b + \varepsilon)b, \\ &\alpha_1 a = (\alpha_1 ab^\star b + \alpha_1 a + \alpha_1 b + \varepsilon)a, \end{aligned}$$

and $\mathsf{R}_\varepsilon(\alpha_1) = \{\alpha_1 ab^\star b, \alpha_1 a, \alpha_1 b, \varepsilon\}$. From these sets $S(\alpha)$ and $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ are easily obtained.

3.2. Iterative Definition of $\mathcal{A}_{\mathrm{Pre}}$

The definition of the prefix automaton above is similar to Mirkin's characterisation of the partial derivative automaton as a solution of a system of right expression equations [18] (see also Section 4.2). In [1] a different iterative construction for $\mathcal{A}_{\mathrm{PD}}$ was independently given by Antimirov. His method starts with expression α as the initial state of $\mathcal{A}_{\mathrm{PD}}$ and the remaining states and transitions are obtained by successively deriving (labels of) states by the symbols of the alphabet. In the following we present a similar iterative approach for the construction of $\mathcal{A}_{\mathrm{Pre}}$. Starting with $\mathsf{R}_\varepsilon(\alpha)$ as the set of final states, the automaton will be successively constructed backwards as follows. For each state of the form $\alpha'\sigma$ the set $\mathsf{R}_\varepsilon(\alpha')$ is computed and a transition by σ is added from each element $\alpha'' \in \mathsf{R}_\varepsilon(\alpha')$ to $\alpha'\sigma$. The state labelled by ε is the initial state. Formally, consider the function $\mathsf{p}_w(\alpha)$ for words $w \in \Sigma^\star$ defined as follows:

$$\mathsf{p}_\varepsilon(\alpha) = \mathsf{R}_\varepsilon(\alpha), \qquad \mathsf{p}_{\sigma w}(\alpha) = \bigcup_{\alpha'\sigma \,\in\, \mathsf{p}_w(\alpha)} \mathsf{R}_\varepsilon(\alpha'). \tag{6}$$

Note that $\mathsf{R}_\varepsilon(\alpha) \subseteq \mathsf{S}(\alpha) \cup \{\varepsilon\}$. It is straightforward to show the following fact.

Fact 1. *Every element in* $\mathsf{p}_w(\alpha)$ *is of the form* $\alpha'\sigma$ *or* ε*.*

Lemma 5. *For any* $w \in \Sigma^\star$ *and* $\alpha \in \mathit{RE}$*, we have*

(I) $\mathcal{L}(\mathsf{p}_w(\alpha)) = \{\, x \mid xw \in \mathcal{L}(\alpha) \,\} = \mathcal{L}(\alpha)w^{-1}$.

(II) $w \in \mathcal{L}(\alpha)$ *if and only if* $\varepsilon \in \mathcal{L}(\mathsf{p}_w(\alpha))$ *if and only if* $\varepsilon \in \mathsf{p}_w(\alpha)$ *if and only if exists* $\alpha'\sigma' \in \mathsf{p}_w(\alpha)$ *such that* $\varepsilon(\alpha') = \varepsilon$*.*

Proof. We prove (I) by induction on the length of w. For $w = \varepsilon$,

$$\mathcal{L}(\mathsf{p}_\varepsilon(\alpha)) = \mathcal{L}(\mathsf{R}_\varepsilon(\alpha)) = \mathcal{L}(\alpha).$$

For $w = \sigma w'$,

$$\mathcal{L}(\mathsf{p}_w(\alpha)) = \mathcal{L}(\bigcup_{\alpha'\sigma \in \mathsf{p}_{w'}(\alpha)} \mathsf{R}_\varepsilon(\alpha')).$$

Then $x \in \mathcal{L}(\mathsf{p}_w(\alpha))$ if and only if exists $\alpha'\sigma \in \mathsf{p}_{w'}(\alpha)$ such that

$$x \in \mathcal{L}(\mathsf{R}_\varepsilon(\alpha')) = \mathcal{L}(\alpha').$$

By the inductive hypothesis one has $x\sigma \in \mathcal{L}(\mathsf{p}_{w'}(\alpha)) = \mathcal{L}(\alpha)w'^{-1}$, and $x \in \mathcal{L}(\alpha)w^{-1}$. By (I), the first equivalence of (II) is immediate. For the second and third, just consider Fact 1 and the definition of $\mathsf{p}_w(\alpha)$. □

Finally, let $\mathsf{Pre}(\alpha) = \bigcup_{w \in \Sigma^\star} \mathsf{p}_w(\alpha)$. Now, consider the automaton

$$\langle \mathsf{Pre}(\alpha), \Sigma, \delta_{\mathrm{Pre}}, \varepsilon, \mathsf{R}_\varepsilon(\alpha) \rangle, \tag{7}$$

where

$$\delta_{\mathrm{Pre}} = \{\, (\tau, \sigma, \beta\sigma) \mid \beta\sigma \in \mathsf{Pre}(\alpha) \wedge \tau \in \mathsf{R}_\varepsilon(\beta) \wedge \sigma \in \Sigma \,\},$$

i.e., we have $\delta^{\mathrm{R}}_{\mathrm{Pre}}(\beta\sigma, \sigma) = \mathsf{R}_\varepsilon(\beta)$, for all $\beta\sigma \in \mathsf{Pre}(\alpha)$, $\sigma \in \Sigma$.

In the following we prove that this automaton coincides with $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ as defined in (5). First note that the initial state and set of final states coincide in both definitions. Thus we need to prove that $\mathsf{S}(\alpha) \cup \{\varepsilon\} = \mathsf{Pre}(\alpha)$ and $\delta_{\mathrm{Pre}} = \{\varepsilon\} \times \psi(\alpha) \,\cup\, \mathsf{T}(\alpha)$.

Lemma 6. *For all* $\alpha \in \mathit{RE}$ *we have,*

(I) *If* $\alpha'\sigma \in \mathsf{S}(\alpha)$*, then* $\mathsf{S}(\alpha') \subseteq \mathsf{S}(\alpha)$*.*

(II) *For* $w \in \Sigma^\star$*, one has* $\mathsf{p}_w(\alpha) \subseteq \mathsf{S}(\alpha) \cup \{\varepsilon\}$*.*

Proof.

(I) We proceed by structural induction on α. For $\emptyset$, ε, and σ, the result is vacuously true. Let $\alpha = \alpha_1 + \alpha_2$ and consider $\alpha'\sigma \in \mathsf{S}(\alpha_1 + \alpha_2) = \mathsf{S}(\alpha_1) \cup \mathsf{S}(\alpha_2)$. Without loss of generality we suppose that $\alpha'\sigma \in \mathsf{S}(\alpha_1)$. Then, $\mathsf{S}(\alpha') \subseteq \mathsf{S}(\alpha_1) \subseteq \mathsf{S}(\alpha)$ follows by the induction hypothesis.

Let $\alpha = \alpha_1\alpha_2$, and let $\alpha'\sigma \in \alpha_1 \mathsf{S}(\alpha_2) \cup \mathsf{S}(\alpha_1)$. If $\alpha'\sigma \in \mathsf{S}(\alpha_1)$, then as in the previous case, the result follows from the induction hypothesis and by definition of S. If $\alpha'\sigma \in \alpha_1 \mathsf{S}(\alpha_2)$, there are two cases to consider. Either, we have $\alpha'\sigma = \alpha_1\alpha''\sigma$ with $\alpha''\sigma \in \mathsf{S}(\alpha_2)$, or $\alpha' = \alpha_1$ and $\sigma \in \mathsf{S}(\alpha_2)$. In the first case, $\mathsf{S}(\alpha'') \subseteq \mathsf{S}(\alpha_2)$, and consequently

$$\mathsf{S}(\alpha_1\alpha'') = \alpha_1\,\mathsf{S}(\alpha'') \cup \mathsf{S}(\alpha_1) \subseteq \alpha_1\,\mathsf{S}(\alpha_2) \cup \mathsf{S}(\alpha_1) = \mathsf{S}(\alpha).$$

In the second case, $\mathsf{S}(\alpha') = \mathsf{S}(\alpha_1) \subseteq \alpha_1\,\mathsf{S}(\alpha_2) \cup \mathsf{S}(\alpha_1) = \mathsf{S}(\alpha)$.

Let $\alpha = \alpha_1^\star$, and let $\alpha'\sigma \in \mathsf{S}(\alpha_1^\star) = \alpha_1^\star\,\mathsf{S}(\alpha_1)$. If $\alpha'\sigma = \alpha_1^\star\alpha''\sigma$ for some $\alpha''\sigma \in \mathsf{S}(\alpha_1)$, then

$$\mathsf{S}(\alpha') = \mathsf{S}(\alpha_1^\star\alpha'') = \alpha_1^\star\,\mathsf{S}(\alpha'') \cup \mathsf{S}(\alpha_1^\star) \subseteq \alpha_1^\star\,\mathsf{S}(\alpha_1) = \mathsf{S}(\alpha),$$

because $\mathsf{S}(\alpha'') \subseteq \mathsf{S}(\alpha_1)$ by the induction hypothesis. Otherwise, $\alpha' = \alpha_1^\star = \alpha$ and $\sigma \in \mathsf{S}(\alpha_1)$ and the result is trivially true.

(ii) We prove the statement by induction on length of $w \in \Sigma^\star$. If $w = \varepsilon$, then

$$\mathsf{p}_\varepsilon(\alpha) = \mathsf{R}_\varepsilon(\alpha) \subseteq \mathsf{S}(\alpha) \cup \{\varepsilon\}.$$

Now, let $w = \sigma w'$ and consider $\alpha'' \in \mathsf{p}_{\sigma w'}(\alpha)$, i.e. $\alpha'' \in \mathsf{R}_\varepsilon(\alpha')$, such that $\alpha'\sigma \in \mathsf{p}_{w'}(\alpha)$ for some expression α'. Furthermore, suppose that $\alpha'' \neq \varepsilon$. It follows from the induction hypothesis that $\alpha'\sigma \in \mathsf{S}(\alpha)$. By (i), $\mathsf{S}(\alpha') \subseteq \mathsf{S}(\alpha)$. Thus $\alpha'' \in \mathsf{R}_\varepsilon(\alpha') \subseteq \mathsf{S}(\alpha') \cup \{\varepsilon\} \subseteq \mathsf{S}(\alpha) \cup \{\varepsilon\}$.

□

As a consequence of Lemma 6 (ii) we obtain the following result.

Proposition 7. $\mathsf{Pre}(\alpha) \subseteq \mathsf{S}(\alpha) \cup \{\varepsilon\}$.

To prove the inclusion $\mathsf{S}(\alpha) \cup \{\varepsilon\} \subseteq \mathsf{Pre}(\alpha)$, we define

$$\mathsf{Pre}^+(\alpha) = \mathsf{Pre}(\alpha) \setminus \{\varepsilon\}$$

and

$$\mathsf{p^+}_w(\alpha) = \mathsf{p}_w(\alpha) \setminus \{\varepsilon\}.$$

From the Proposition 7 we have that $\mathsf{Pre}^+(\alpha) \subseteq \mathsf{S}(\alpha)$. For the other inclusion we need the two lemmas below. In the following, to show that an inclusion $\mathsf{Pre}^+(\alpha) \subseteq E$ (resp. $\alpha'\mathsf{Pre}^+(\alpha) \subseteq E$) holds for some set E, we show by induction on the length of w that for every $w \in \Sigma^\star$ one has $\mathsf{p^+}_w(\alpha) \subseteq E$ (resp. $\alpha'\mathsf{p^+}_w(\alpha) \subseteq E$). First, we show that the set Pre^+ also satisfies the statement of Lemma 6 (i).

Lemma 8. *For all $\alpha \in \mathsf{RE}$ we have,*

(i) *If $\alpha'\sigma \in \mathsf{Pre}^+(\alpha)$, then $\mathsf{Pre}^+(\alpha') \subseteq \mathsf{Pre}^+(\alpha)$.*

(ii) *If $\alpha'\sigma \in \mathsf{Pre}^+(\alpha)$, then $\alpha''\alpha'\sigma \in \mathsf{Pre}^+(\alpha''\alpha)$, for all $\alpha'' \in \mathsf{RE}$.*

Proof.

(I) If $\alpha'\sigma \in \mathsf{Pre}^+(\alpha)$ there exists $w \in \Sigma^\star$ such that $\alpha'\sigma \in \mathsf{p}_w(\alpha)$ and, by definition $\mathsf{R}(\alpha') \subseteq \mathsf{p^+}_{\sigma w}(\alpha) \subseteq \mathsf{Pre}^+(\alpha)$. Suppose that $\mathsf{p^+}_x(\alpha') \subseteq \mathsf{Pre}^+(\alpha)$ for $x \in \Sigma^\star$. Then, for $\sigma' \in \Sigma$, we have

$$\mathsf{p^+}_{\sigma' x}(\alpha') = \bigcup_{\alpha''\sigma' \in \mathsf{p^+}_x(\alpha')} \mathsf{R}(\alpha'') \subseteq \bigcup_{\alpha''\sigma' \in \mathsf{Pre}^+(\alpha)} \mathsf{R}(\alpha'') \subseteq \mathsf{Pre}^+(\alpha).$$

(II) We prove the result by induction on the length of an word $w' \in \Sigma^\star$ such that $\alpha'\sigma \in \mathsf{p^+}_{w'}(\alpha)$. If $\alpha'\sigma \in \mathsf{R}(\alpha)$, then

$$\alpha''\alpha'\sigma \in \mathsf{R}(\alpha''\alpha) \subseteq \mathsf{Pre}^+(\alpha''\alpha).$$

In the case $\alpha'\sigma \in \mathsf{p^+}_{\sigma' w}(\alpha)$, there exists $\alpha'''\sigma' \in \mathsf{p^+}_w(\alpha)$ such that $\alpha'\sigma \in \mathsf{R}(\alpha''')$. Then

$$\alpha''\alpha'\sigma \in \mathsf{R}(\alpha''\alpha''') \subseteq \mathsf{Pre}^+(\alpha''\alpha''')$$

and, by the inductive hypothesis, we conclude that $\alpha''\alpha'''\sigma' \in \mathsf{Pre}^+(\alpha''\alpha)$. By (I), $\mathsf{Pre}^+(\alpha''\alpha''') \subseteq \mathsf{Pre}^+(\alpha''\alpha)$, and thus, finally, $\alpha''\alpha'\sigma \in \mathsf{Pre}^+(\alpha''\alpha)$.

□

Lemma 9. Pre^+ *satisfies the following*

$$\begin{array}{ll} \mathsf{Pre}^+(\emptyset) = \mathsf{Pre}^+(\varepsilon) = \emptyset, & \mathsf{Pre}^+(\sigma) = \{\sigma\}, \\ \mathsf{Pre}^+(\alpha+\alpha') \supseteq \mathsf{Pre}^+(\alpha) \cup \mathsf{Pre}^+(\alpha'), & \mathsf{Pre}^+(\alpha\alpha') \supseteq \alpha\mathsf{Pre}^+(\alpha') \cup \mathsf{Pre}^+(\alpha), \\ \mathsf{Pre}^+(\alpha^\star) \supseteq \alpha^\star\mathsf{Pre}^+(\alpha). & \end{array}$$

Proof. The proof proceeds by induction on the structure of α. For $\emptyset$, ε, and σ the result is obvious.

- For $\alpha + \alpha'$, we have $\mathsf{p^+}_\varepsilon(\alpha) = \mathsf{R}(\alpha) \subseteq \mathsf{R}(\alpha+\alpha') \subseteq \mathsf{Pre}^+(\alpha+\alpha')$ and

$$\mathsf{p^+}_{\sigma w}(\alpha) = \bigcup_{\alpha''\sigma \in \mathsf{p^+}_w(\alpha)} \mathsf{R}(\alpha'') \subseteq \bigcup_{\alpha''\sigma \in \mathsf{Pre}^+(\alpha+\alpha')} \mathsf{R}(\alpha'') \subseteq \mathsf{Pre}^+(\alpha+\alpha').$$

  The same applies for α', and thus $\mathsf{Pre}^+(\alpha) \cup \mathsf{Pre}^+(\alpha') \subseteq \mathsf{Pre}^+(\alpha+\alpha')$.
- For $\alpha\alpha'$, we first note that $\alpha\mathsf{Pre}^+(\alpha') \subseteq \mathsf{Pre}^+(\alpha\alpha')$ is a direct consequence of Lemma 8 (II). Now, we prove that $\mathsf{Pre}^+(\alpha) \subseteq \mathsf{Pre}^+(\alpha\alpha')$. Suppose that $\varepsilon(\alpha') = \varepsilon$, then $\mathsf{p^+}_\varepsilon(\alpha) = \mathsf{R}(\alpha) \subseteq \mathsf{R}(\alpha\alpha') \subseteq \mathsf{Pre}^+(\alpha\alpha')$, as well as

$$\mathsf{p^+}_{\sigma w}(\alpha) = \bigcup_{\alpha''\sigma \in \mathsf{p^+}_w(\alpha)} \mathsf{R}(\alpha'') \subseteq \bigcup_{\alpha''\sigma \in \mathsf{Pre}^+(\alpha\alpha')} \mathsf{R}(\alpha'') \subseteq \mathsf{Pre}^+(\alpha\alpha').$$

  If $\varepsilon(\alpha') = \emptyset$ then applying Lemma 5 (II) there exists $\alpha''\sigma \in \mathsf{Pre}^+(\alpha')$ such that $\varepsilon(\alpha'') = \varepsilon$. By the previous case, one has $\mathsf{Pre}^+(\alpha) \subseteq \mathsf{Pre}^+(\alpha\alpha'')$. On the other hand, by Lemma 8 (II), $\alpha\alpha''\sigma \in \mathsf{Pre}^+(\alpha\alpha')$. Applying once more Lemma 8 (I), we conclude that $\mathsf{Pre}^+(\alpha\alpha'') \subseteq \mathsf{Pre}^+(\alpha\alpha')$. By transitivity, it follows that $\mathsf{Pre}^+(\alpha) \subseteq \mathsf{Pre}^+(\alpha\alpha')$.

- For $\alpha^\star$, we have $\alpha^\star {\sf p^+}_\varepsilon(\alpha) = \alpha^\star\, {\sf R}(\alpha) = {\sf R}(\alpha^\star) \subseteq {\sf Pre}^+(\alpha^\star)$ and

  $$\alpha^\star {\sf p^+}_{\sigma w}(\alpha) = \bigcup_{\alpha'\sigma \,\in\, {\sf p^+}_w(\alpha)} \alpha^\star\, {\sf R}(\alpha') \subseteq \bigcup_{\alpha'\sigma \,\in\, {\sf p^+}_w(\alpha)} {\sf Pre}^+(\alpha^\star\alpha') \subseteq {\sf Pre}^+(\alpha^\star),$$

  where the first inclusion follows from Lemma 8 (ii) and the second inclusion follows from Lemma 8 (i) (as if $\alpha^\star\alpha'\sigma \in \alpha^\star {\sf p^+}_w(\alpha) \subseteq {\sf Pre}^+(\alpha^\star)$ then ${\sf Pre}^+(\alpha^\star\alpha') \subseteq {\sf Pre}^+(\alpha^\star)$).

□

From Lemma 9 and the definition of ${\sf S}$ it is immediate that ${\sf S}(\alpha) \subseteq {\sf Pre}^+(\alpha)$, and so we proved the following proposition.

Proposition 10. ${\sf S}(\alpha) \cup \{\varepsilon\} \subseteq {\sf Pre}(\alpha)$.

The following theorem ensures that $\mathcal{A}_{\rm Pre}(\alpha)$ is the automaton defined in (7), i.e.,

$$\langle {\sf Pre}(\alpha), \Sigma, \delta_{\rm Pre}, \varepsilon, {\sf R}_\varepsilon(\alpha)\rangle.$$

Theorem 11. ${\sf S}(\alpha) \cup \{\varepsilon\} = {\sf Pre}(\alpha)$ *and* $\delta_{\rm Pre} = \{\varepsilon\} \times \psi(\alpha) \;\cup\; {\sf T}(\alpha)$.

Proof. The equality ${\sf S}(\alpha) \cup \{\varepsilon\} = {\sf Pre}(\alpha)$ follows from Proposition 10 and Proposition 7. Moreover, we have ${\sf Pre}^+(\alpha) = {\sf S}(\alpha)$. The set of transitions $\delta_{\rm Pre}$ can be seen as the union of following two sets, where we assume that $\sigma \in \Sigma$.

$$\{\, (\varepsilon, \sigma, \beta\sigma) \mid \beta\sigma \in {\sf Pre}(\alpha) \wedge \varepsilon(\beta) = \varepsilon \,\} \cup \{\, (\tau, \sigma, \beta\sigma) \mid \beta\sigma \in {\sf Pre}^+(\alpha) \wedge \tau \in {\sf R}(\beta) \,\}.$$

The first set is exactly $\{\varepsilon\} \times \psi(\alpha)$. Using induction on the structure of α we show that the second set is equal to ${\sf T}(\alpha)$, i.e.,

$${\sf T}(\alpha) = \{\, (\tau, \sigma, \beta\sigma) \mid \beta\sigma \in {\sf Pre}^+(\alpha) \wedge \tau \in {\sf R}(\beta) \,\}.$$

For the base cases the equality holds. Suppose that the equality holds for α_1 and α_2. Let $\alpha = \alpha_1 + \alpha_2$. Then

$$\begin{aligned}
&\{\, (\tau, \sigma, \beta\sigma) \mid \beta\sigma \in {\sf Pre}^+(\alpha_1 + \alpha_2) \wedge \tau \in {\sf R}(\beta) \,\}\\
&\qquad = \{\, (\tau, \sigma, \beta\sigma) \mid \beta\sigma \in {\sf Pre}^+(\alpha_1) \wedge \tau \in {\sf R}(\beta) \,\}\\
&\qquad \cup \{\, (\tau, \sigma, \beta\sigma) \mid \beta\sigma \in {\sf Pre}^+(\alpha_2) \wedge \tau \in {\sf R}(\beta) \,\}\\
&\qquad = {\sf T}(\alpha_1) + {\sf T}(\alpha_2) = {\sf T}(\alpha_1 + \alpha_2).
\end{aligned}$$

Let $\alpha = \alpha_1\alpha_2$. Then

$$\begin{aligned}
&\{\, (\tau, \sigma, \beta\sigma) \mid \beta\sigma \in {\sf Pre}^+(\alpha_1\alpha_2) \wedge \tau \in {\sf R}(\beta) \,\}\\
&\qquad = \{\, (\tau, \sigma, \beta\sigma) \mid \beta\sigma \in \alpha_1{\sf Pre}^+(\alpha_2) \wedge \tau \in {\sf R}(\beta) \,\}\\
&\qquad \cup \{\, (\tau, \sigma, \beta\sigma) \mid \beta\sigma \in {\sf Pre}^+(\alpha_2) \wedge \tau \in {\sf R}(\beta) \,\}
\end{aligned}$$

By the induction hypothesis, the second set is $\mathsf{T}(\alpha_2)$. On the other hand, we have

$$\begin{aligned}
&\{\, (\tau,\sigma,\beta\sigma) \mid \beta\sigma \in \alpha_1\mathsf{Pre}^{+}(\alpha_2) \wedge \tau \in \mathsf{R}(\beta) \,\}\\
&\qquad = \{\, (\tau,\sigma,\beta\sigma) \mid \beta \doteq \alpha_1\gamma \wedge \gamma\sigma \in \mathsf{Pre}^{+}(\alpha_2) \wedge \tau \in \mathsf{R}(\alpha_1\gamma) \,\}\\
&\qquad = \{\, (\tau,\sigma,\beta\sigma) \mid \beta \doteq \alpha_1\gamma \wedge \gamma\sigma \in \mathsf{Pre}^{+}(\alpha_2) \wedge \tau \doteq \alpha_1\gamma' \wedge \gamma' \in \mathsf{R}(\gamma) \,\}\\
&\qquad \cup \{\, (\tau,\sigma,\beta\sigma) \mid \beta \doteq \alpha_1\gamma \wedge \gamma\sigma \in \mathsf{Pre}^{+}(\alpha_2) \wedge \varepsilon(\gamma) = \varepsilon \wedge \tau \in \mathsf{R}(\alpha_1) \,\}\\
&\qquad = \alpha_1\, \mathsf{T}(\alpha_2) \cup \mathsf{R}(\alpha_1) \times \alpha_1\psi(\alpha_2).
\end{aligned}$$

We conclude that

$$\{\, (\tau,\sigma,\beta\sigma) \mid \beta\sigma \in \mathsf{Pre}^{+}(\alpha_1\alpha_2) \wedge \tau \in \mathsf{R}(\beta) \,\} = \mathsf{T}(\alpha_2) \cup \alpha_1\, \mathsf{T}(\alpha_2) \cup \mathsf{R}(\alpha_1) \times \alpha_1\psi(\alpha_2).$$

Let $\alpha = \alpha_1^{\star}$. Then

$$\begin{aligned}
&\{\, (\tau,\sigma,\beta\sigma) \mid \beta\sigma \in \mathsf{Pre}^{+}(\alpha_1^{\star}) \wedge \tau \in \mathsf{R}(\beta) \,\}\\
&\qquad = \{\, (\tau,\sigma,\beta\sigma) \mid \beta\sigma \in \alpha_1^{\star}\mathsf{Pre}^{+}(\alpha_1) \wedge \tau \in \mathsf{R}(\beta) \,\}\\
&\qquad = \{\, (\tau,\sigma,\beta\sigma) \mid \beta \doteq \alpha_1^{\star}\gamma \wedge \gamma\sigma \in \mathsf{Pre}^{+}(\alpha_1) \wedge \tau \in \mathsf{R}(\alpha_1^{\star}\gamma) \,\}\\
&\qquad = \{\, (\tau,\sigma,\beta\sigma) \mid \beta \doteq \alpha_1^{\star}\gamma \wedge \gamma\sigma \in \mathsf{Pre}^{+}(\alpha_1) \wedge \tau \doteq \alpha_1^{\star}\gamma' \wedge \gamma' \in \mathsf{R}(\gamma) \,\}\\
&\qquad \cup \{\, (\tau,\sigma,\beta\sigma) \mid \beta \doteq \alpha_1^{\star}\gamma \wedge \gamma\sigma \in \mathsf{Pre}^{+}(\alpha_1) \wedge \varepsilon(\gamma) = \varepsilon \wedge \tau \in \alpha_1^{\star}\, \mathsf{R}(\alpha_1) \,\}\\
&\qquad = \alpha_1^{\star}\, \mathsf{T}(\alpha_1) \cup \alpha_1^{\star}(\mathsf{R}(\alpha_1) \times \psi(\alpha_1)).
\end{aligned}$$

□

4. Relation with Other Constructions

In this section we relate the prefix automaton with several other constructions from regular expressions to NFAs. From Section 4.1 to Section 4.3 we recall the definitions and some properties of the position automaton and the partial derivative automaton as well as their duals, i.e., the dual position automaton ($\mathcal{A}_{\overleftarrow{\mathrm{POS}}}$) and the right-partial derivative automaton ($\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$). For the $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ we also show that it is (isomorphic to) a quotient $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}$. The subsequent sections relate $\mathcal{A}_{\mathrm{Pre}}$ with these automata. While $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ is a quotient of $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}$, $\mathcal{A}_{\mathrm{Pre}}$ is a quotient of $\mathcal{A}_{\mathrm{POS}}$. For completeness we also consider the dual prefix automaton $\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}$ and relate it with $\mathcal{A}_{\mathrm{PD}}$.

4.1. The Position Automaton and its the Dual

The position automaton, introduced by Glushkov [14], permits us to convert a regular expression α into an equivalent NFA without ε-transitions. The states in the position automaton correspond to the positions of letters in α plus an additional initial state. Formally, given $\alpha \in \mathsf{RE}$, one can mark each occurrence of a letter σ with its position in α, considering reading it from left to right. The resulting regular expression is a *marked* regular expression $\overline{\alpha}$ with all symbols distinct and over the alphabet $\Sigma_{\overline{\alpha}}$. Then, a position $i \in [1, |\alpha|_{\Sigma}]$ corresponds to the symbol σ_i in $\overline{\alpha}$, and consequently to exactly one occurrence of σ in α. The same notation is used to remove the markings, i.e., $\overline{\overline{\alpha}} = \alpha$. Marking and unmarking can also be applied to a set of expressions.

Example 12. The marked version of $\tau = (ab^\star + b)^\star a$ is $\overline{\tau} = (a_1 b_2^\star + b_3)^\star a_4$.

Let $\mathsf{Pos}(\alpha) = \{1, 2, \dots, |\alpha|_\Sigma\}$, and let $\mathsf{Pos}_0(\alpha) = \mathsf{Pos}(\alpha) \cup \{0\}$. To define the $\mathcal{A}_{\mathrm{POS}}(\alpha)$ we consider the following sets:

$$\begin{aligned}
\mathsf{First}(\alpha) &= \{\, i \mid \sigma_i w \in \mathcal{L}(\overline{\alpha}) \,\},\\
\mathsf{Last}(\alpha) &= \{\, i \mid w\sigma_i \in \mathcal{L}(\overline{\alpha}) \,\},\\
\mathsf{Follow}(\alpha, i) &= \{\, j \mid u\sigma_i\sigma_j v \in \mathcal{L}(\overline{\alpha}) \,\}.
\end{aligned}$$

It is convenient to extend $\mathsf{Follow}(\alpha, 0) = \mathsf{First}(\alpha)$ and define that $\mathsf{Last}_0(\alpha)$ is $\mathsf{Last}(\alpha)$ if $\varepsilon(\alpha) = \emptyset$, or is $\mathsf{Last}(\alpha) \cup \{0\}$, otherwise. We define the position automaton using the approach by Broda et al. [3], where the transition function is expressed as the composition of functions Select and Follow. Given a letter σ and a set of positions S, the function Select selects the subset of positions in S that correspond to letter σ. Formally, given a subset S of $\mathsf{Pos}(\alpha)$ and $\sigma \in \Sigma$, let

$$\mathsf{Select}(S, \sigma) = \{\, i \mid i \in S \wedge \overline{\sigma_i} = \sigma \,\}.$$

Then, the *position automaton* for α is

$$\mathcal{A}_{\mathrm{POS}}(\alpha) = \langle \mathsf{Pos}_0(\alpha), \Sigma, \delta_{\mathrm{POS}}, 0, \mathsf{Last}_0(\alpha) \rangle,$$

where $\delta_{\mathrm{POS}}(i, \sigma) = \mathsf{Select}(\mathsf{Follow}(\alpha, i), \sigma)$. Broda et al. [3] defined a dual of the position automaton, $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}$, which has only one final state $n+1$, where $n = |\alpha|_\Sigma$ and the initial states are $\mathsf{Follow}(\alpha, 0) \cup \varepsilon(\alpha)\{n+1\}$. Formally

$$\mathcal{A}_{\overleftarrow{\mathrm{POS}}}(\alpha) = \langle \mathsf{Pos}(\alpha) \cup \{n+1\}, \Sigma, \delta_{\overleftarrow{\mathrm{POS}}}, \mathsf{Follow}(\alpha, 0) \cup \varepsilon(\alpha)\{n+1\}, \{n+1\} \rangle,$$

where for $i \in \mathsf{Pos}(\alpha) \cup \{n+1\}$ one has $\delta_{\overleftarrow{\mathrm{POS}}}(i, \sigma) = \mathsf{Follow}(\alpha, i) \cup \varepsilon(i)\{n+1\}$, if $i \neq n+1$ and $\overline{\sigma_i} = \sigma$, being the empty set otherwise. In particular, it was shown that

$$\mathcal{A}_{\mathrm{POS}}(\alpha^{\mathrm{R}})^{\mathrm{R}} \simeq \mathcal{A}_{\overleftarrow{\mathrm{POS}}}(\alpha).$$

Example 13. For $\alpha = (ab^\star + b)^\star a$ with $\overline{\alpha} = (a_1 b_2^\star + b_3)^\star a_4$ we can compute the sets:

$$\begin{aligned}
\mathsf{First}(\alpha) &= \{1, 3, 4\}, & \mathsf{Last}(\alpha) &= \{4\},\\
\mathsf{Follow}(\alpha, 1) &= \{1, 2, 3, 4\}, & \mathsf{Follow}(\alpha, 2) &= \{1, 2, 3, 4\},\\
\mathsf{Follow}(\alpha, 3) &= \{1, 3, 4\}, & \mathsf{Follow}(\alpha, 4) &= \emptyset.
\end{aligned}$$

The corresponding position automaton and its dual are depicted in Figure 2 and in Figure 3, respectively.

4.2. The Partial Derivative Automaton

The partial derivative automaton of a regular expression was introduced independently by Mirkin [18] and Antimirov [1]. Champarnaud and Ziadi [9] proved that

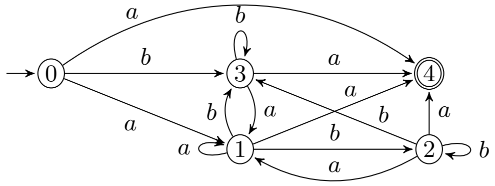


Figure 2: $\mathcal{A}_{\text{POS}}((ab^\star + b)^\star a)$.

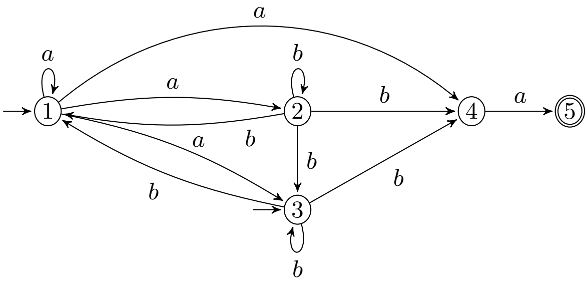


Figure 3: $\mathcal{A}_{\overleftarrow{\text{POS}}}((ab^\star + b)^\star a)$.

the two formulations are equivalent. For a regular expression $\alpha \in \mathsf{RE}$ and a symbol $\sigma \in \Sigma$, the set of *partial derivatives* of α w.r.t. σ is defined inductively as follows:

$$\begin{array}{ll} \partial_\sigma(\emptyset) = \partial_\sigma(\varepsilon) = \emptyset, & \partial_\sigma(\alpha + \alpha') = \partial_\sigma(\alpha) \cup \partial_\sigma(\alpha'), \\ \partial_\sigma(\sigma') = \begin{cases} \{\varepsilon\} & \text{if } \sigma' = \sigma, \\ \emptyset & \text{otherwise}, \end{cases} & \begin{array}{r} \partial_\sigma(\alpha\alpha') = \partial_\sigma(\alpha)\alpha' \cup \varepsilon(\alpha)\partial_\sigma(\alpha'), \\ \partial_\sigma(\alpha^\star) = \partial_\sigma(\alpha)\alpha^\star. \end{array} \end{array} \tag{8}$$

The definition of partial derivatives can be extended in a natural way to sets of regular expressions, words, and languages. For $w \in \Sigma^*$, we have

$$w^{-1}\mathcal{L}(\alpha) = \mathcal{L}(\partial_w(\alpha)) = \bigcup_{\tau \in \partial_w(\alpha)} \mathcal{L}(\tau).$$

The set of all partial derivatives of α w.r.t. words is denoted by $\mathsf{PD}(\alpha) = \partial_{\Sigma^*}(\alpha)$. The *partial derivative automaton* of α is

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

where $F_{\text{PD}} = \{\, \tau \in \mathsf{PD}(\alpha) \mid \varepsilon(\tau) = \varepsilon \,\}$, and $\delta_{\text{PD}}(\tau, \sigma) = \partial_\sigma(\tau)$, for all $\tau \in \mathsf{PD}(\alpha)$ and $\sigma \in \Sigma$. Mirkin's construction of the $\mathcal{A}_{\text{PD}}(\alpha)$ is based on the existence of a set of expressions $\pi(\alpha) = \{\alpha_1, \ldots, \alpha_n\}$ that satisfies a system of equations

$$\alpha_i = \sigma_1\alpha_{i1} + \cdots + \sigma_k\alpha_{ik} + \varepsilon(\alpha_i),$$

with $\alpha_0 \doteq \alpha$ and such that α_{ij} are linear combinations of elements of $\pi(\alpha)$, for all $i \in [1, n]$ and $j \in [1, k]$. It follows that $\mathsf{PD}(\alpha) = \pi(\alpha) \cup \{\alpha\}$ and that $\mathcal{A}_{\text{PD}}$ can be

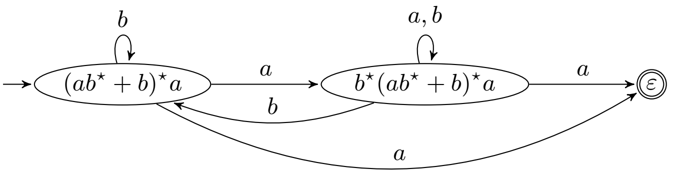


Figure 4: $\mathcal{A}_{\mathrm{PD}}((ab^\star + b)^\star a)$.

defined in an inductive manner [4, 17, 16]. For completeness we give that inductive construction in the appendix.

Champarnaud and Ziadi [10] proved that $\mathcal{A}_{\mathrm{PD}}$ is a quotient of $\mathcal{A}_{\mathrm{POS}}$ by the relation $\equiv_c$, which is a right-invariant relation w.r.t. $\mathcal{A}_{\mathrm{POS}}$. Given a position i there is some expression $c_i(\alpha)$ such that for all $w \in \Sigma^\star_{\overline{\alpha}}$, either $\partial_{w\sigma_i}(\overline{\alpha}) = \emptyset$ or $\partial_{w\sigma_i}(\overline{\alpha}) = \{c_i(\alpha)\}$. For $i, j \in \mathsf{Pos}_0(\alpha)$ and considering $c_0(\alpha) = \alpha$, one has

$$i \equiv_c j \iff \overline{c_i(\alpha)} \doteq \overline{c_j(\alpha)}.$$

Proposition 14 [10]. $\mathcal{A}_{\mathrm{PD}}(\alpha) \simeq \mathcal{A}_{\mathrm{POS}}(\alpha)/{\equiv_c}$.

Example 15. Figure 4 presents the automaton $\mathcal{A}_{\mathrm{PD}}((ab^\star + b)^\star a)$. Considering the set $\mathsf{Pos}_0((a_1 b_2^\star + b_3)^\star a_4)) = \{0, 1, 2, 3, 4\}$ we have $0 \equiv_c 3$ and $1 \equiv_c 2$.

4.3. The Right-Partial Derivative Automaton

Partial derivatives correspond to left-quotients of the language of a regular expression. In the same way one can consider the right-quotients and define the *right-partial derivatives* of an expression in a dual manner. For a regular expression $\alpha \in \mathsf{RE}$ and a symbol $\sigma \in \Sigma$, the set of right-partial derivatives of α w.r.t. σ, $\overleftarrow{\partial}_\sigma(\alpha)$, is defined inductively as in (8) except for the following cases:

$$\overleftarrow{\partial}_\sigma(\alpha\beta) = \alpha\overleftarrow{\partial}_\sigma(\beta) \cup \varepsilon(\beta)\overleftarrow{\partial}_\sigma(\alpha) \quad \text{and} \quad \overleftarrow{\partial}_\sigma(\alpha^\star) = \alpha^\star\overleftarrow{\partial}_\sigma(\alpha). \tag{9}$$

The definition of right-partial derivative can be extended in a natural way to sets of regular expressions, words, and languages. The set of all right-partial derivatives of α w.r.t. words is denoted by $\overleftarrow{\mathsf{PD}}(\alpha) = \bigcup_{w\in\Sigma^\star} \overleftarrow{\partial}_w(\alpha)$. The next results relate the left and the right partial derivatives, where the reversal of set a of regular expressions is the set of the reversals of its elements.

Lemma 16. *For any $\alpha \in \mathsf{RE}$, $\sigma \in \Sigma$, and $w \in \Sigma^\star$ we have*

(I) $(\partial_\sigma(\alpha^R))^R = \overleftarrow{\partial}_\sigma(\alpha)$.

(II) $(\partial_{w^R}(\alpha^R))^R = \overleftarrow{\partial}_w(\alpha)$.

(III) $\mathcal{L}(\overleftarrow{\partial}_w(\alpha)) = \mathcal{L}(\alpha)w^{-1}$.

(IV) $\overleftarrow{\mathsf{PD}}(\alpha) = (\mathsf{PD}(\alpha^R))^R$.

Proof. The proof of (I) proceeds by induction on the structure of α using the definitions. The proof of (II) proceeds by induction on the size of w using (I). Finally, (III) and (IV) are consequences of the first two items. □

The *right-partial derivative automaton* of α is

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

where

$$\begin{aligned}\overleftarrow{\delta_{\mathrm{PD}}} &= \{\, (\tau, \sigma, \beta) \mid \tau \in \overleftarrow{\partial}_{\sigma}(\beta) \wedge \tau \in \overleftarrow{\mathsf{PD}}(\alpha) \wedge \sigma \in \Sigma \,\},\\ \overleftarrow{F_{\mathrm{PD}}} &= \{\, \tau \in \overleftarrow{\mathsf{PD}}(\alpha) \mid \varepsilon(\tau) = \varepsilon \,\}.\end{aligned}$$

Proposition 17 [16]. $\mathcal{L}(\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)) = \mathcal{L}(\alpha)$.

Note that $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$ has always one final state although it can have more than one initial state. Similarly to $\mathcal{A}_{\mathrm{PD}}$, the $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$ can also be defined based on the existence of a set of expressions $\overleftarrow{\pi}(\alpha) = \{\alpha_1, \ldots, \alpha_n\}$ that satisfies a system of equations of the form

$$\alpha_i = \alpha_{i1}\sigma_1 + \cdots + \alpha_{ik}\sigma_k + \varepsilon(\alpha_i),$$

with $\alpha_0 \doteq \alpha$ and such that α_{ij} are linear combinations of $\overleftarrow{\pi}(\alpha)$ for all $i \in [1, n]$ and $j \in [1, k]$ [16, 17]. Using the results above we can relate $\mathcal{A}_{\mathrm{PD}}$ with $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$.

Proposition 18. $(\mathcal{A}_{\mathrm{PD}}(\alpha^R))^R \simeq \mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$.

From Proposition 18 and Proposition 14 we have the following.

Corollary 19. $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha) \simeq (\mathcal{A}_{\mathrm{POS}}(\alpha^R))^R/{\equiv_c}$.

Proof. The following hold

$$\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha) \simeq (\mathcal{A}_{\mathrm{PD}}(\alpha^R))^R \simeq \left(\mathcal{A}_{\mathrm{POS}}(\alpha^R) \big/ {\equiv_c} \right)^R \simeq (\mathcal{A}_{\mathrm{POS}}(\alpha^R))^R \big/ {\equiv_c}.$$

□

Note that $\equiv_c$ is a left-invariant relation on the set of states of $(\mathcal{A}_{\mathrm{POS}}(\alpha^R))^R$. As mentioned in Section 4.1 this last automaton is isomorphic to the $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}$ [3].

Example 20. For $\alpha = (ab^\star + b)^\star a$, $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$ is represented in Figure 5 and can be obtained from $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}(\alpha)$ in Figure 3, by merging states 1, 3 and 4.

4.4. $\mathcal{A}_{\mathrm{Pre}}$ *as a quotient of* $\mathcal{A}_{\mathrm{POS}}$

As mentioned before Yamamoto showed that $\mathcal{A}_{\mathrm{Suf}}$ coincides with $\mathcal{A}_{\mathrm{PD}}$. This fact could lead us to think that $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ coincides with $\mathcal{A}_{\mathrm{Pre}}$, which is not true. For instance, considering $\alpha = a + b$, the $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$ has two states and the $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ has three states

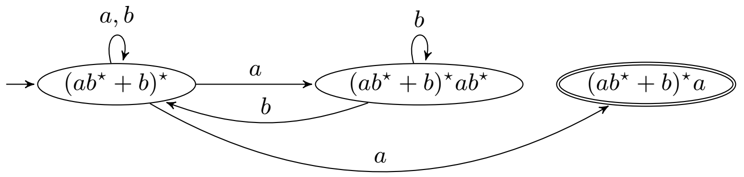
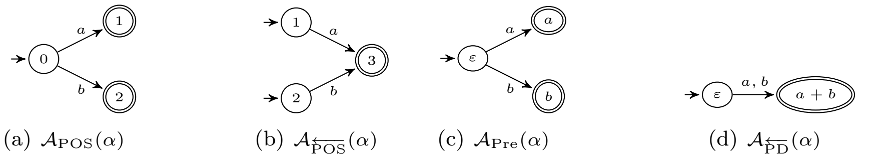


Figure 5: $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}((ab^\star + b)^\star a)$

(see Figure 6). While $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$ is a left-quotient of $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}$, we will see that $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ is a left-quotient of $\mathcal{A}_{\mathrm{POS}}(\alpha)$.



(a) $\mathcal{A}_{\mathrm{POS}}(\alpha)$ (b) $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}(\alpha)$ (c) $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ (d) $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$

Figure 6: Automata for $\alpha = a + b$.

Considering a marked regular expression $\overline{\alpha}$, $\mathcal{A}_{\mathrm{POS}}(\overline{\alpha})$ is deterministic and thus all its states, labelled with indexes $i \in \mathsf{Pos}(\alpha) \cup \{0\}$, have distinct left-languages. On the other hand, for each index $i \in \mathsf{Pos}(\alpha)$ there is exactly one state $\beta\sigma_i$ in $\mathcal{A}_{\mathrm{Pre}}(\overline{\alpha})$. The initial states of the two automata are respectively labelled with 0 and ε. We conclude that $|\mathsf{Pre}(\overline{\alpha})| = |\overline{\alpha}|_\Sigma$. The fact that $\mathcal{A}_{\mathrm{Pre}}(\overline{\alpha}) \simeq \mathcal{A}_{\mathrm{POS}}(\overline{\alpha})$ follows from the following lemma which can be proved by induction on the structure of α [16].

Lemma 21. *For any marked regular expression α we have*

(i) $\mathsf{First}(\alpha) = \{\, i \mid \beta\sigma_i \in \mathsf{Pre}^+(\alpha) \wedge \varepsilon(\beta) = \varepsilon \,\}$.

(ii) $\mathsf{Last}(\alpha) = \{\, i \mid \beta\sigma_i \in \mathsf{R}(\alpha) \,\}$.

(iii) *For all $\beta\sigma_i, \tau\sigma_j \in \mathsf{Pre}^+(\alpha)$, one has $(\beta\sigma_i, \sigma_j, \tau\sigma_j) \in \delta_{\mathrm{Pre}}$ iff $j \in \mathsf{Follow}(\alpha, i)$.*

Proposition 22. $\mathcal{A}_{\mathrm{Pre}}(\overline{\alpha}) \simeq \mathcal{A}_{\mathrm{POS}}(\overline{\alpha})$.

Proof. To prove that these automata are isomorphic we just consider the bijection $\varphi_{\mathrm{p}} : \mathsf{Pre} \to \mathsf{Pos}_0$ such that $\varphi_{\mathrm{p}}(\varepsilon) = 0$, and $\varphi_{\mathrm{p}}(\gamma) = \mathsf{Last}(\gamma)$, if $\gamma \in \mathsf{Pre}(\overline{\alpha}) \setminus \{\varepsilon\}$. For the initial and final states the isomorphism is obvious. For the transition functions the isomorphism follows from Lemma 21. □

Moreover, considering the automaton $\overline{\mathcal{A}_{\mathrm{Pre}}(\overline{\alpha})}$ that is obtained from $\mathcal{A}_{\mathrm{Pre}}(\overline{\alpha})$ by unmarking the letters labelling the transitions, we have the following result.

Proposition 23. $\overline{\mathcal{A}_{\mathrm{Pre}}(\overline{\alpha})} \simeq \mathcal{A}_{\mathrm{POS}}(\alpha)$.

To show that $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ is a quotient of $\mathcal{A}_{\mathrm{POS}}(\alpha)$, we first consider an equivalence relation on the set $\mathsf{Pre}(\overline{\alpha})$. For $\beta, \tau \in \mathsf{Pre}(\overline{\alpha})$, let $\equiv_p$ be defined by

$$\beta \equiv_p \tau \iff \overline{\beta} \doteq \overline{\tau}.$$

The following lemma is immediate by the definition of R.

Lemma 24. *For all $\beta\sigma_i, \tau\sigma_j \in \mathsf{Pre}(\overline{\alpha})$, if $\overline{\beta\sigma_i} \doteq \overline{\tau\sigma_j}$, then $\overline{\mathsf{R}_\varepsilon(\beta)} = \overline{\mathsf{R}_\varepsilon(\tau)}$.*

Lemma 25. *The relation $\equiv_p$ is left-invariant w.r.t. $\overline{\mathcal{A}_{\mathrm{Pre}}(\overline{\alpha})}$.*

Proof. We have $\equiv_p \subseteq (\mathsf{Pre}^+(\overline{\alpha}))^2 \cup \{(\varepsilon, \varepsilon)\}$. Let $\beta\sigma_i, \tau\sigma_j \in \mathsf{Pre}(\overline{\alpha})$ with $\beta\sigma_i \equiv_p \tau\sigma_j$. Let $\tau' \in \delta^R(\tau\sigma_j, \sigma) = \mathsf{R}_\varepsilon(\tau)$ with $\overline{\sigma_j} \doteq \sigma$. Then $\overline{\tau'} \in \overline{\mathsf{R}_\varepsilon(\tau)}$ and by Lemma 24 there exists $\beta' \in \mathsf{R}_\varepsilon(\beta) = \delta^R(\beta\sigma_i, \sigma)$ such that $\overline{\tau'} \doteq \overline{\beta'}$ (i.e., $\overline{\beta'} \in \overline{\mathsf{R}_\varepsilon(\beta)}$). Thus $\beta' \equiv_p \tau'$. □

Corollary 26. $\mathcal{A}_{\mathrm{Pre}}(\alpha) \simeq \overline{\mathcal{A}_{\mathrm{Pre}}(\overline{\alpha})}/\equiv_p$.

Proof. Consider $\varphi_{\mathrm{u}} : \mathsf{Pre}(\overline{\alpha})/\equiv_p \to \mathsf{Pre}(\alpha)$ defined by $\varphi_{\mathrm{u}}([\varepsilon]) = \varepsilon$ and $\varphi_{\mathrm{u}}([\beta]) = \overline{\beta}$. It is obvious that φ_{u} is a bijection and defines an isomorphism between the automaton $\overline{\mathcal{A}_{\mathrm{Pre}}(\overline{\alpha})}/\equiv_p$ and the automaton $\mathcal{A}_{\mathrm{Pre}}(\alpha)$. □

As seen before, all the expressions of $\mathsf{Pre}(\overline{\alpha})$ are of the form that $\alpha'\sigma_i$ or ε and for each position $i \in \mathsf{Pos}(\alpha)$ there exists a unique $\alpha'\sigma_i \in \mathsf{Pre}(\overline{\alpha})$. Let $p_i(\alpha)$ be that expression. Considering the isomorphism φ_{p} defined between Pre and Pos_0 (cf. Lemma 23) we can define $\equiv_\ell \; = \; \equiv_p \circ \varphi_{\mathrm{p}}$ which is left-invariant w.r.t. $\mathcal{A}_{\mathrm{POS}}$ and for $i, j \in \mathsf{Pos}_0(\alpha)$ verifies the following

$$i \equiv_\ell j \iff \overline{p_i(\alpha)} \doteq \overline{p_j(\alpha)}.$$

Theorem 27. $\mathcal{A}_{\mathrm{Pre}}(\alpha) \simeq \mathcal{A}_{\mathrm{POS}}(\alpha)/\equiv_\ell$.

Proof. Let $\varphi_\ell : \mathsf{Pos}(\alpha)/\equiv_\ell \to \mathsf{Pre}(\alpha)$ be defined by $\varphi_\ell([0]) = \varepsilon$ and $\varphi_\ell([i]) = \overline{p_i(\alpha)}$. From the above, it is obvious that φ_ℓ is a bijection and defines an isomorphism between the automaton $\mathcal{A}_{\mathrm{POS}}(\alpha)/\equiv_\ell$ and the automaton $\mathcal{A}_{\mathrm{Pre}}(\alpha)$. □

Example 28. Consider $\alpha = (ab^\star + b)^\star a$, with $\overline{\alpha} = (a_1 b_2^\star + b_3)^\star a_4$, and the automata $\mathcal{A}_{\mathrm{POS}}$ and $\mathcal{A}_{\mathrm{Pre}}$ depicted in Fig 2 and 1, respectively. We have $1 \equiv_\ell 4$.

4.5. $\mathcal{A}_{\mathrm{Pre}}$ versus $\mathcal{A}_{\mathrm{PD}}$, and their Duals

Broda et al. [3] defined a dual version of the prefix automaton, denoted by $\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}$, such that

$$\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}(\alpha) \simeq (\mathcal{A}_{\mathrm{Pre}}(\alpha^{\mathrm{R}}))^{\mathrm{R}}$$

and

$$\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}(\alpha) \simeq \mathcal{A}_{\overleftarrow{\mathrm{POS}}}(\alpha)/_{\equiv_\ell}.$$

To define that automaton one uses a set $\mathsf{L}_\varepsilon(\alpha) = \mathsf{L}(\alpha) \cup \{\varepsilon(\alpha)\}$, where $\mathsf{L}(\alpha)$ is defined as $\mathsf{R}(\alpha)$ except for concatenation and Kleene star, where

$$\mathsf{L}(\alpha\alpha') = \mathsf{L}(\alpha)\alpha' \cup \varepsilon(\alpha)\,\mathsf{L}(\alpha'), \qquad \mathsf{L}(\alpha^\star) = \mathsf{L}(\alpha)\alpha^\star.$$

These sets relate to partial (left or right) derivatives as follows.

Table 1: Experimental results for the number of states of some automata constructions.

k	$\lvert\alpha\rvert$	$\lvert \mathsf{Pos}_0 \rvert$	$\lvert \mathrm{PD} \rvert$	$\frac{\lvert \mathrm{PD} \rvert}{\lvert \mathrm{POS} \rvert}$	$\lvert \overleftarrow{\mathrm{PD}} \rvert$	$\frac{\lvert \overleftarrow{\mathrm{PD}} \rvert}{\lvert \mathrm{POS} \rvert}$	$\lvert \mathrm{Pre} \rvert$	$\frac{\lvert \mathrm{Pre} \rvert}{\lvert \mathrm{POS} \rvert}$	$1 - \eta_k$
2	100	28.9	15.7	0.55	15.9	0.55	20.1	0.71	0.90
	500	139.9	71.6	0.51	71.5	0.51	91.9	0.66	
10	100	42.5	23.8	0.56	23.8	0.56	38.5	0.91	0.99
	500	207.1	113.2	0.55	112.4	0.54	186	0.90	
	1000	412.1	223.7	0.54	223.1	0.54	369.5	0.90	

Lemma 29 **[3].** *For any $\alpha \in \mathsf{RE}$,*

(I) $\mathsf{R}_\varepsilon(\alpha) = \bigcup_{\sigma\in\Sigma} \overleftarrow{\partial}_\sigma(\alpha)\sigma \cup \varepsilon(\alpha)$.

(II) $\mathsf{L}_\varepsilon(\alpha) = \bigcup_{\sigma\in\Sigma} \sigma\partial_\sigma(\alpha) \cup \varepsilon(\alpha)$.

In particular, the established relation implies that the number of states of $\mathcal{A}_{\mathrm{PD}}(\alpha)$ is always less than or equal to the number of states of $\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}(\alpha)$. The same holds for $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$ and $\mathcal{A}_{\mathrm{Pre}}(\alpha)$. We note that $\mathcal{A}_{\mathrm{PD}}$ and $\mathcal{A}_{\mathrm{Pre}}$ are generally not comparable. For instance, for $\alpha = (ab^\star + b)^\star a$ from Examples 4 and 15 the number of states in $\mathcal{A}_{\mathrm{PD}}(\alpha)$ is three, while $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ has four states. On the other hand, for $\alpha = a^\star ab + (ab)^\star + a^\star$ there are seven states in $\mathcal{A}_{\mathrm{PD}}(\alpha)$ and five in $\mathcal{A}_{\mathrm{Pre}}(\alpha)$.

5. Average Size Complexity

In this section we analyse the average size of the automata obtained from regular expressions by the different constructions considered in the previous sections. We use both experimental as well as theoretical asymptotical methods considering regular expressions of a given size following a uniform distribution. Note that although this distribution on expressions is an adequate choice, it does not relate directly with any distribution in the realm of regular languages.

First we consider results of experimental tests carried out in order to compare the sizes of $\mathcal{A}_{\mathrm{POS}}$, $\mathcal{A}_{\mathrm{PD}}$, $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ and $\mathcal{A}_{\mathrm{Pre}}$ automata. We used the FAdo library [1] that includes implementations of those NFA conversions, and several tools for uniformly random generate regular expressions. In order to obtain regular expressions uniformly generated in the size of the syntactic tree, we used a prefix notation version of the grammar (1). For each alphabet size, k, and expression size, $|\alpha|$, samples of 10 000 regular expressions were generated, which is sufficient to ensure a 95% confidence level within a 1% error margin [11]. Tables 1 and 2 present the average values obtained for $|\alpha| \in \{100, 500, 1000\}$ and $k \in \{2, 10\}$.

These experiments suggest that, on average, the $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ and the $\mathcal{A}_{\mathrm{PD}}$ have the same size and the $\mathcal{A}_{\mathrm{Pre}}$ is not significantly smaller than the $\mathcal{A}_{\mathrm{POS}}$.

Nicaud [19] showed that on average and asymptotically the number of transitions of $\mathcal{A}_{\mathrm{POS}}$ is linear on the size of the expression. Broda et al. [4, 5] studied the average size of $\mathcal{A}_{\mathrm{PD}}$ and concluded that, on average and asymptotically, the $\mathcal{A}_{\mathrm{PD}}$ has at most half the number of states and transitions of the $\mathcal{A}_{\mathrm{POS}}$. By Proposition 18, $|\alpha^R|_\Sigma = |\alpha|_\Sigma$

[1] `fado.dcc.fc.up.pt`

Table 2: Experimental results for the number of transitions of some automata constructions.

k	$\lvert\alpha\rvert$	$\lvert\delta_{\mathrm{POS}}\rvert$	$\lvert\delta_{\mathrm{PD}}\rvert$	$\frac{\lvert\delta_{\mathrm{PD}}\rvert}{\lvert\delta_{\mathrm{POS}}\rvert}$	$\lvert\delta_{\overleftarrow{\mathrm{PD}}}\rvert$	$\frac{\lvert\delta_{\overleftarrow{\mathrm{PD}}}\rvert}{\lvert\delta_{\mathrm{POS}}\rvert}$	$\lvert\delta_{\mathrm{Pre}}\rvert$	$\frac{\lvert\delta_{\overleftarrow{\mathrm{PD}}}\rvert}{\lvert\delta_{\mathrm{POS}}\rvert}$
2	100	167.5	56.0	0.33	56.4	0.34	73.7	0.44
	500	1486.5	389.8	0.26	393.1	0.26	530.8	0.36
10	100	159.4	73.7	0.46	72.9	0.46	130.4	0.82
	500	1019.1	423.8	0.42	425.6	0.42	807.1	0.79
	1000	2182.1	884.1	0.41	884.5	0.41	1717.6	0.79

and by the fact that $\varepsilon \in \pi(\alpha)$ if and only if $\varepsilon \in \overleftarrow{\pi}(\alpha)$, this analysis of the average size of $\mathcal{A}_{\mathrm{PD}}(\alpha)$ still holds for $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$. Thus the average sizes of $\mathcal{A}_{\mathrm{PD}}$ and $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ are asymptotically the same. However, $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha)$ has only one final state and its number of initial states is the number of final states of $\mathcal{A}_{\mathrm{PD}}(\alpha^R)$.

Again following the ideas in Broda et al., we estimate the number of mergings of states that arise when computing $\mathcal{A}_{\mathrm{Pre}}$ from $\mathcal{A}_{\mathrm{POS}}$. The $\mathcal{A}_{\mathrm{Pre}}$ has at most $|\alpha|_\Sigma + 1$ states and this only occurs when all unions in $\mathsf{Pre}^+(\alpha)$ are disjoint. However for some cases this does not happen. For instance, when $\sigma \in \mathsf{Pre}^+(\beta) \cap \mathsf{Pre}^+(\gamma)$, then

$$\begin{aligned} |\mathsf{Pre}^+(\beta+\gamma)| &= |\mathsf{Pre}^+(\beta) \cup \mathsf{Pre}^+(\gamma)| \leq |\mathsf{Pre}^+(\beta)| + |\mathsf{Pre}^+(\gamma)| - 1, \\ |\mathsf{Pre}^+(\beta^\star\gamma)| &= |\beta^\star\mathsf{Pre}^+(\gamma) \cup \beta^\star\mathsf{Pre}^+(\beta)| \leq |\mathsf{Pre}^+(\beta)| + |\mathsf{Pre}^+(\gamma)| - 1. \end{aligned} \tag{10}$$

In what follows, we estimate the number of these non-disjoint unions, which corresponds to a lower bound for the number of states merged in the $\mathcal{A}_{\mathrm{POS}}$ automaton. This is done in the framework of analytic combinatorics as expounded by Flajolet and Sedgewick [12] (see also [5]). The method applies to generating functions

$$A(z) = \sum_n a_n z^n$$

for a combinatorial class $\mathcal{A}$ with a_n objects of size n, denoted by $[z^n]A(z)$, and also bivariate functions

$$C(u,z) = \sum_\alpha u^{c(\alpha)} z^{|\alpha|},$$

where $c(\alpha)$ is some measure of the object $\alpha \in \mathcal{A}$. The symbolic method [12] is a framework that allows to express a combinatorial class C in terms of simpler ones, $\mathsf{B}_1,\ldots,\mathsf{B}_n$, by means of specific operations, yielding the generating function $C(z)$ as a function of the generating functions $B_i(z)$ of B_i, for $1 \leq i \leq n$.

Generating functions can be seen as complex analytic functions, and the study of their behaviour around their dominant singularities gives us access to an asymptotic estimate for their coefficients. We refer the reader to Flajolet and Sedgewick for an extensive study on this topic. Here we only state the results relevant for this paper. For $\rho \in \mathbb{C}$, $R > 1$ and $0 < \phi < \pi/2$, consider the domain

$$\Delta(\rho,\phi,R) = \{\, z \in \mathbb{C} \mid |z| < R,\ z \neq \rho,\ \text{and } |Arg(z-\rho)| > \phi \,\},$$

where $Arg(z)$ denotes the argument of $z \in \mathbb{C}$. A region is a Δ-*domain* at ρ if it is a $\Delta(\rho, \phi, R)$, for some R and ϕ. The generating functions we consider have always a unique dominant singularity, and satisfy one of the two conditions of the following proposition, used by Nicaud [19].

Proposition 30. *Let $f(z)$ be a function that is analytic in some Δ-domain at $\rho \in \mathbb{R}^+$. If at the intersection of a neighborhood of ρ and its Δ-domain,*

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

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

(II) $f(z) = \frac{a}{\sqrt{1-z/\rho}} + o\left(\frac{1}{\sqrt{1-z/\rho}}\right)$, *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. The Average State Complexity of $\mathcal{A}_{\text{Pre}}$

The regular expressions α_σ for which $\sigma \in \mathsf{Pre}(\alpha_\sigma)$, $\sigma \in \Sigma$ are generated by following grammar

$$\alpha_\sigma := \sigma \mid (\alpha_\sigma + \alpha) \mid (\alpha_{\overline{\sigma}} + \alpha_\sigma) \mid (\alpha_\sigma \cdot \alpha) \mid (\varepsilon \cdot \alpha_\sigma).$$

The regular expressions that are not generated by α_σ are denoted by $\alpha_{\overline{\sigma}}$ and α are regular expressions given by grammar (1) (omitting the $\emptyset$). The generating function for regular expressions is

$$R_k(z) = \frac{1 - z - \sqrt{\Delta_k(z)}}{4z}, \tag{11}$$

where $\Delta_k(z) = 1 - 2z - (7+8k)z^2$ and the zeros of $\Delta_k(z)$ are

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

Moreover [19, 4],

$$[z^n]R_k(z) \sim \frac{\sqrt{2(1-\rho_k)}}{8\rho_k\sqrt{\pi}}\rho_k^{-n}n^{-3/2}. \tag{12}$$

The generating function for α_σ, $R_{\sigma,k}(z)$, satisfies

$$R_{\sigma,k}(z) = z + zR_{\sigma,k}(z)R_k(z) + z(R_k(z) - R_{\sigma,k}(z))R_{\sigma,k}(z) + zR_{\sigma,k}(z)R_k(z) + z^2R_{\sigma,k}(z)$$

which is equivalent to

$$zR_{\sigma,k}(z)^2 - (3zR_k(z) + z^2 - 1)R_{\sigma,k}(z) - z = 0.$$

From this one gets

$$R_{\sigma,k}(z) = \frac{(z^2 + 3zR_k(z) - 1) + \sqrt{(z^2 + 3zR_k(z) - 1)^2 + 4z^2}}{2z}.$$

Using Equation (11) for $R_k(z)$ one has

$$8zR_{\sigma,k}(z) = -b(z) - 3\sqrt{\Delta_k(z)} + \sqrt{a(z) + 6b(z)\sqrt{\Delta_k(z)} + 9\Delta_k(z)},$$

where $a(z) = 16z^4 - 24z^3 + 65z^2 + 6z + 1$ and $b(z) = -4z^2 + 3z + 1$. Using the binomial theorem, we know that

$$\sqrt{a(z) + 6b(z)\sqrt{\Delta_k(z)} + 9\Delta_k(z)} = \sqrt{a(z)} + 3\frac{b(z)}{\sqrt{a(z)}}\sqrt{\Delta_k(z)} + o(\Delta_k(z)^{\frac{1}{2}}).$$

Thus,

$$8zR_{\sigma,k}(z) = -b(z) + \sqrt{a(z)} + 3\left(\frac{b(z)}{\sqrt{a(z)}} - 1\right)\sqrt{\Delta_k(z)} + o(\Delta_k(z)^{\frac{1}{2}}).$$

As we know that the following equalities are true:

$$\sqrt{\Delta_k(z)} = \sqrt{(7 + 8k)\rho_k(z - \overline{\rho}_k)}\sqrt{1 - z/\rho_k},$$

$$\sqrt{(7 + 8k)\rho_k(\rho_k - \overline{\rho}_k)} = \sqrt{2 - 2\rho_k},$$

and using the Proposition 30 one has

$$[z^n]R_{\sigma,k}(z) \sim \frac{3}{16\sqrt{\pi}}\left(1 - \frac{b(\rho_k)}{\sqrt{a(\rho_k)}}\right)\sqrt{2(1 - \rho_k)}\rho_k^{-(n+1)}n^{-\frac{3}{2}}.$$

Thus the asymptotic ratio of regular expressions with $\sigma \in \mathsf{Pre}(\alpha)$ is:

$$\frac{[z^n]R_{\sigma,k}(z)}{[z^n]R_k(z)} \sim \frac{3}{2}\left(1 - \frac{b(\rho_k)}{\sqrt{a(\rho_k)}}\right).$$

As $\lim\limits_{k\to\infty} \rho_k = 0$, $\lim\limits_{k\to\infty} a(\rho_k) = 1$, and $\lim\limits_{k\to\infty} b(\rho_k) = 1$, the asymptotic ratio of regular expressions with $\sigma \in \mathsf{Pre}$ approaches 0 when $k \to \infty$.

Let $i(\alpha)$ be the number of non-disjoint unions appearing during the computation of $\mathsf{Pre}(\alpha)$, $\alpha \in \mathsf{RE}$ originated by the two cases described in (10). Then $i(\alpha)$ verifies the following equations

$$\begin{aligned} i(\varepsilon) = i(\sigma) &= 0, & i(\alpha_\sigma^\star\alpha_\sigma) &= i(\alpha_\sigma^\star) + i(\alpha_\sigma) + 1,\\ i(\alpha_\sigma + \alpha_\sigma) &= i(\alpha_\sigma) + i(\alpha_\sigma) + 1, & i(\overline{\alpha_\sigma^\star}\alpha_\sigma) &= i(\overline{\alpha_\sigma^\star}) + i(\alpha_\sigma),\\ i(\alpha_\sigma + \alpha_{\overline{\sigma}}) &= i(\alpha_\sigma) + i(\alpha_{\overline{\sigma}}), & i(\alpha\alpha_{\overline{\sigma}}) &= i(\alpha) + i(\alpha_{\overline{\sigma}}),\\ i(\alpha_{\overline{\sigma}} + \alpha) &= i(\alpha_{\overline{\sigma}}) + i(\alpha), & i(\alpha^\star) &= i(\alpha). \end{aligned}$$

From these equations we can obtain the cost generating function of the mergings, $I_\sigma(z)$, by adding the contributions of each one of them. For example, the contribution of the regular expressions of the form $\alpha_\sigma + \alpha_\sigma$ can be computed as follows:

$$\begin{aligned}\sum_{\alpha_\sigma+\alpha_\sigma} i(\alpha_\sigma + \alpha_\sigma) z^{|(\alpha_\sigma+\alpha_\sigma)|} &= z \sum_{\alpha_\sigma}\sum_{\alpha_\sigma} (i(\alpha_\sigma) + i(\alpha_\sigma) + 1) z^{|\alpha_\sigma|} z^{|\alpha_\sigma|} \\ &= z \sum_{\alpha_\sigma}\sum_{\alpha_\sigma} (i(\alpha_\sigma) + i(\alpha_\sigma)) z^{|\alpha_\sigma|} z^{|\alpha_\sigma|} + z \sum_{\alpha_\sigma}\sum_{\alpha_\sigma} z^{|\alpha_\sigma|} z^{|\alpha_\sigma|} \\ &= 2z I_{\alpha_\sigma,k}(z) R_{\sigma,k}(z) + z R_{\sigma,k}(z)^2,\end{aligned}$$

where $I_{\alpha_\sigma,k}(z)$ is the generating function for the mergings coming from α_σ. Applying this technique to the remaining cases, we obtain

$$I_\sigma(z) = \frac{(z+z^2)R_{\sigma,k}(z)^2}{\sqrt{\Delta_k(z)}}.$$

Using again the same Proposition 30, we can conclude that:

$$[z^n]I_\sigma(z) \sim \frac{1+\rho_k}{64} \frac{\left(a(\rho_k) + b(\rho_k)^2 - 2b(\rho_k)\sqrt{a(\rho_k)}\right)}{\sqrt{\pi}\sqrt{2-2\rho_k}} \rho_k^{-(n+1)} n^{-\frac{1}{2}}.$$

Recall that the number of states of $\mathcal{A}_{\mathrm{POS}}(\alpha)$ is equal to the number of letters in α. Thus in order to obtain a lower bound for the reduction in the number of states of the $\mathcal{A}_{\mathrm{Pre}}$ automaton, as compared to the ones of the $\mathcal{A}_{\mathrm{POS}}$ automaton, it is enough to compare the number of mergings for an expression α, with the number of letters in α. From Nicaud [19] one knows that the generating function for the number of letters $L_k(z)$ satisfies the following

$$[z^n]L_k(z) \sim \frac{k\rho_k}{\sqrt{\pi(2-2\rho_k)}} \rho_k^{-n} n^{-1/2}.$$

Therefore, the asymptotic estimate for the average number of mergings is given by:

$$\frac{[z^n]I_\sigma(z)}{[z^n]L_k(z)} \sim \frac{1-\rho_k}{4k\rho_k^2}\lambda_k = \eta_k, \text{ where}$$
$$\lambda_k = \frac{(1+\rho_k)}{16(1-\rho_k)}\left(a(\rho_k) + b(\rho_k)^2 - 2b(\rho_k)\sqrt{a(\rho_k)}\right).$$

It is not difficult to conclude that $\lim_{k\to\infty} \lambda_k = 0$, therefore $\lim_{k\to\infty} \eta_k = 0$.

As it is evident from the last two columns of Table 1, for small values of k, the lower bound η_k does not capture all the mergings that occur in $\mathcal{A}_{\mathrm{Pre}}$. However, it seems that for larger values of k, the average number of states of the $\mathcal{A}_{\mathrm{Pre}}$ automaton approaches the number of states of the $\mathcal{A}_{\mathrm{POS}}$ automaton.

6. The Prefix Automaton for Regular Expressions with Intersection

In this section we extend the prefix automaton to regular expressions with the intersection operator. The set $\mathsf{RE}_\cap$ of *regular expressions with intersection* over Σ is obtained by adding the rule $\alpha := (\alpha \cap \alpha)$ in grammar (1). We have

$$\mathcal{L}(\alpha \cap \beta) = \mathcal{L}(\alpha) \cap \mathcal{L}(\beta).$$

Recently, the partial derivative automaton and the position automaton were extended to $\mathsf{RE}_\cap$ [2, 6, 8]. In the case of the position automaton, the states were labelled by sets of indexes. If reading a sequence of letters leads to a state with a label $I = \{i_1, \ldots, i_n\}$, then in the corresponding marked regular expression, one just reads simultaneously letters $\sigma_{i_1}, \ldots, \sigma_{i_n}$ for some (unmarked) letter σ.

Example 31. Let $\alpha = (ab^\star a + a)^\star \cap (aa + b)^\star$ with $\overline{\alpha} = (a_1 b_2^\star a_3 + a_4)^\star \cap (a_5 a_6 + b_7)^\star$. After reading the sequence of letters aa, the letters read in the marked expressions are either a_3 and a_6, or a_4 and a_6. Thus there will be a path in the position automaton from the initial state to a state with label $\{3, 6\}$, as well as to a state with label $\{4, 6\}$.

Furthermore, it was shown that the partial derivative automaton is a quotient of this position automaton construction by an extension of relation $\equiv_c$. In the case of expressions containing intersection, and due to the fact that some subexpressions may describe the empty language, the inductive constructions of these automata may include useless states, i.e., states with an empty right language.

Example 32. For α from Example 31 and $\beta = (c \cap d)$ we have $\overline{\alpha\beta} = \overline{\alpha}(c_8 \cap d_9)$. In this case states $\{3, 6\}$ and $\{4, 6\}$ in the position automaton of $\alpha\beta$ are now useless states.

We now define the position automaton for expressions in $\mathsf{RE}_\cap$, using the Select function defined in Section 4.1. This construction leads to an initially connected automaton, which improves the inductive definition presented in [8]. That is obtained by defining the Follow set only for the necessary sets of indexes, i.e., labels of reachable states. Given $\alpha \in \mathsf{RE}_\cap$, both $\overline{\alpha}$ and $\mathsf{Pos}(\alpha)$ are defined as before. For the labels of states of $\mathcal{A}_{\mathrm{POS}}(\alpha)$ one has to consider non-empty subsets of $\mathsf{Pos}(\alpha)$ where all indexes correspond to the same letter. For this, we define $\ell(i) = \sigma$ for $\overline{\sigma_i} = \sigma$, as well as $\ell(I) = \sigma$ if for all $i \in I \subseteq \mathsf{Pos}(\alpha)$ one has $\ell(i) = \sigma$. The set of all non-empty subsets I of $\mathsf{Pos}(\alpha)$, such that $\ell(I) = \sigma$ for some $\sigma \in \Sigma$, is denoted by $\mathsf{Ind}(\alpha)$. For $S_1, S_2 \subseteq \mathsf{Ind}(\alpha)$, we define

$$S_1 \otimes S_2 = \{\, I_1 \cup I_2 \mid \ell(I_1) = \ell(I_2) \wedge I_1 \in S_1, I_2 \in S_2 \,\}.$$

Given a marked expression α, a subexpression β of α, and a set of indexes $I \in \mathsf{Ind}(\alpha)$, let $I\big|_\beta$ denote the set of indexes in I that occur in β. This definition is naturally extended to words $x = I_1 \cdots I_n \in \mathsf{Ind}(\alpha)^\star$ by $x\big|_\beta = I_1\big|_\beta \cdots I_n\big|_\beta$, for $n \geq 0$.

Now, we consider the sets $\mathsf{First}(\alpha), \mathsf{Last}(\alpha)$ and $\mathsf{Follow}(\alpha, I) \subseteq \mathsf{Ind}(\alpha)$, for $I \in \mathsf{Ind}(\alpha)$. These sets are defined on a marked expression and as usual [3, 15], except for the base cases and for intersection. First is defined as follows.

$$\begin{aligned}
&\mathsf{First}(\emptyset) = \mathsf{First}(\varepsilon) = \emptyset, && \mathsf{First}(\alpha_1 + \alpha_2) = \mathsf{First}(\alpha_1) \cup \mathsf{First}(\alpha_2),\\
&\mathsf{First}(\sigma_i) = \{\{i\}\}, && \mathsf{First}(\alpha_1\alpha_2) = \begin{cases} \mathsf{First}(\alpha_1) \cup \mathsf{First}(\alpha_2), & \text{if } \varepsilon(\alpha_1) = \varepsilon;\\ \mathsf{First}(\alpha_1), & \text{otherwise,}\end{cases}\\
&\mathsf{First}(\alpha^\star) = \mathsf{First}(\alpha), && \mathsf{First}(\alpha_1 \cap \alpha_2) = \mathsf{First}(\alpha_1) \otimes \mathsf{First}(\alpha_2).
\end{aligned}$$

As usual, Last is defined as First except for the concatenation operator.

$$\mathsf{Last}(\alpha_1\alpha_2) = \begin{cases} \mathsf{Last}(\alpha_1) \cup \mathsf{Last}(\alpha_2), & \text{if } \varepsilon(\alpha_2) = \varepsilon;\\ \mathsf{Last}(\alpha_2), & \text{otherwise.}\end{cases}$$

Now, consider the set Follow with $I \in \mathsf{Ind}(\alpha)$, given by the following rules.

$$\mathsf{Follow}(\sigma_i, I) = \emptyset,$$

$$\mathsf{Follow}(\alpha_1 + \alpha_2, I) = \begin{cases} \mathsf{Follow}(\alpha_1, I) & \text{if } I \in \mathsf{Ind}(\alpha_1),\\ \mathsf{Follow}(\alpha_2, I) & \text{if } I \in \mathsf{Ind}(\alpha_2),\\ \emptyset & \text{otherwise.}\end{cases}$$

$$\mathsf{Follow}(\alpha_1\alpha_2, I) = \begin{cases} \mathsf{Follow}(\alpha_1, I) & \text{if } I \in \mathsf{Ind}(\alpha_1) \wedge I \notin \mathsf{Last}(\alpha_1),\\ \mathsf{Follow}(\alpha_1, I) \cup \mathsf{First}(\alpha_2) & \text{if } I \in \mathsf{Last}(\alpha_1),\\ \mathsf{Follow}(\alpha_2, I) & \text{if } I \in \mathsf{Ind}(\alpha_2),\\ \emptyset & \text{otherwise.}\end{cases}$$

$$\mathsf{Follow}(\alpha^\star, I) = \begin{cases} \mathsf{Follow}(\alpha, I) & \text{if } I \notin \mathsf{Last}(\alpha),\\ \mathsf{Follow}(\alpha, I) \cup \mathsf{First}(\alpha) & \text{if } I \in \mathsf{Last}(\alpha),\\ \emptyset & \text{otherwise.}\end{cases}$$

$$\mathsf{Follow}(\alpha_1 \cap \alpha_2, I) = \begin{cases} \mathsf{Follow}(\alpha_1, I_1) \otimes \mathsf{Follow}(\alpha_2, I_2) & \text{if } I = I_1 \cup I_2 \wedge \\ & \quad I_1 \in \mathsf{Ind}(\alpha_1) \wedge I_2 \in \mathsf{Ind}(\alpha_2),\\ \emptyset & \text{otherwise.}\end{cases}$$

Finally, for $S \subseteq \mathsf{Ind}(\alpha)$ and $\sigma \in \Sigma$ one has $\mathsf{Select}(S, \sigma) = \{\, I \in S \mid \ell(I) = \sigma \,\}$. Then, the position automaton $\mathcal{A}_{\mathrm{POS}}(\alpha)$ is

$$\mathcal{A}_{\mathrm{POS}}(\alpha) = \langle \mathsf{Ind}(\alpha) \cup \{\{0\}\}, \Sigma, \delta_{\mathsf{Pos}}, 0, \mathsf{Last}_0(\alpha)\rangle,$$

where $\mathsf{Last}_0(\alpha)$ is defined as before,

$$\delta_{\mathsf{Pos}}(I, \sigma) = \mathsf{Select}(\mathsf{Follow}(\alpha, I), \sigma), \text{ and}$$

$$\mathsf{Follow}(\alpha, \{0\}) = \mathsf{First}(\alpha).$$

With these definitions we have that, given an expression $\alpha \in \mathsf{RE}_\cap$, $\mathcal{A}_{\mathrm{POS}}(\alpha)$ is equivalent to α [8].

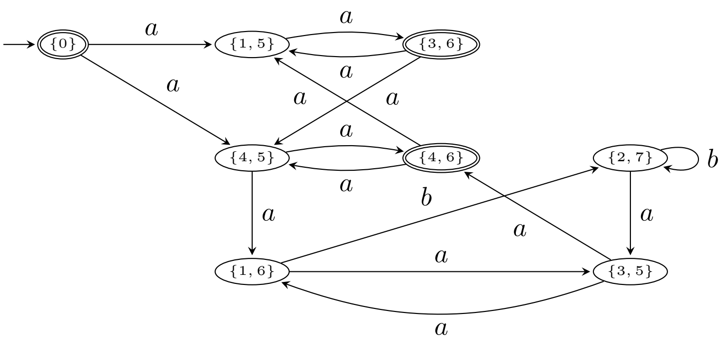


Figure 7: $\mathcal{A}_{\mathrm{POS}}((ab^\star a + a)^\star \cap (aa + b)^\star)$

Example 33. Let $\alpha = (ab^\star a + a)^\star \cap (aa+b)^\star$ with $\overline{\alpha} = (a_1 b_2^\star a_3 + a_4)^\star \cap (a_5 a_6 + b_7)^\star$. In Figure 7 we depict the position automaton for α. We have

$$\begin{aligned}
&\mathsf{First}(\alpha) = \{\{1,5\},\{4,5\}\},\\
&\mathsf{Last}(\alpha) = \{\{3,6\},\{4,6\}\},\\
&\mathsf{Follow}(\alpha,0) = \mathsf{Follow}(\alpha,\{3,6\}) = \mathsf{Follow}(\alpha,\{4,6\}) = \{\{1,5\},\{4,5\}\},\\
&\mathsf{Follow}(\alpha,\{1,5\}) = \{\{3,6\}\},\\
&\mathsf{Follow}(\alpha,\{1,6\}) = \mathsf{Follow}(\alpha,\{2,7\}) = \{\{2,7\},\{3,5\}\},\\
&\mathsf{Follow}(\alpha,\{4,5\}) = \mathsf{Follow}(\alpha,\{3,5\}) = \{\{4,6\},\{1,6\}\}.
\end{aligned}$$

We recall that $\mathcal{A}_{\mathrm{PD}}$ was extended to expressions with intersection [2] by considering, for $\sigma \in \Sigma$

$$\partial_\sigma(\alpha_1 \cap \alpha_2) = \{\, (\alpha \cap \alpha') \mid \alpha \in \partial_\sigma(\alpha_1) \wedge \alpha' \in \partial_\sigma(\alpha_2)\,\}.$$

To define $\mathcal{A}_{\mathrm{Pre}}$ for regular expressions with intersection we extend the function R as follows:

$$\mathsf{R}(\alpha_1 \cap \alpha_2) = \{\, (\alpha \cap \alpha')\sigma \mid \alpha\sigma \in \mathsf{R}(\alpha_1) \wedge \alpha'\sigma \in \mathsf{R}(\alpha_2)\,\}. \tag{13}$$

With this definition, we still have that $\mathcal{L}(\alpha) = \mathcal{L}(\mathsf{R}_\varepsilon(\alpha))$. Consequently, $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ defined as in (7) is equivalent to α. Note that the set S defined in (2) can be also extended for expressions with intersection considering that

$$\mathsf{S}(\alpha_1 \cap \alpha_2) = \{\, (\alpha \cap \alpha')\sigma \mid \alpha\sigma \in \mathsf{S}(\alpha_1) \wedge \alpha'\sigma \in \mathsf{S}(\alpha_2)\,\}. \tag{14}$$

However, as in the case of $\mathcal{A}_{\mathrm{PD}}$ for the sets π and PD^+, one has $\mathsf{Pre}(\alpha) \subseteq \mathsf{S}(\alpha)$ but one can have $\mathsf{S}(\alpha) \nsubseteq \mathsf{Pre}(\alpha)$ [2]. For instance, $\mathsf{Pre}(a(b\cap c)) = \emptyset$, but $\mathsf{S}(a(b\cap c)) = \{a\}$.

In the following we establish that $\mathcal{A}_{\mathrm{Pre}}$ is also a quotient of $\mathcal{A}_{\mathrm{POS}}$. Note that marked expressions of the form $\alpha\cap\beta$ are always either equivalent to the empty language $\emptyset$, or to $\{\varepsilon\}$. Consequently, the result in Propositions 23 does not hold and we have to use a different approach here. Consider function R^{I} that applies to marked expressions and is defined as R in (3) except that $\mathsf{R}^{\mathsf{I}}(\sigma_i) = \{(\varepsilon, \{i\})\}$, and

$$\begin{aligned}\mathsf{R}^{\mathsf{I}}(\alpha_1 \cap \alpha_2) = \{\, & (\alpha \cap \alpha', I_1 \cup I_2) \mid \\ & (\alpha, I_1) \in \mathsf{R}^{\mathsf{I}}(\alpha_1) \wedge (\alpha', I_2) \in \mathsf{R}^{\mathsf{I}}(\alpha_2) \wedge \ell(I_1) = \ell(I_2)\,\},\end{aligned}$$

with the convention that $\alpha(\beta, I) = (\alpha\beta, I)$ and $(\beta, I)\alpha = (\beta\alpha, I)$, and which extends to sets of pairs (β, I). Given a marked expression α and $w \in \mathsf{Ind}(\alpha)^\star$ consider $\mathsf{p}^{\mathsf{I}}{}_w(\alpha)$ such that:

$$\mathsf{p}^{\mathsf{I}}{}_\varepsilon(\alpha) = \mathsf{R}^{\mathsf{I}}{}_\varepsilon(\alpha), \qquad \mathsf{p}^{\mathsf{I}}{}_{Iw}(\alpha) = \bigcup_{(\alpha', I)\ \in\ \mathsf{p}^{\mathsf{I}}{}_w(\alpha)} \mathsf{R}^{\mathsf{I}}{}_\varepsilon(\alpha'),$$

where $\mathsf{R}^{\mathsf{I}}{}_\varepsilon(\alpha) = \mathsf{R}^{\mathsf{I}}(\alpha) \cup \varepsilon(\alpha)$.

For an (unmarked) expression $\alpha \in \mathsf{RE}_\cap$ let,

$$\mathsf{Pre}^{\mathsf{I}}(\overline{\alpha}) = \bigcup_{w \in (\mathsf{Ind}(\alpha))^\star} \mathsf{p}^{\mathsf{I}}{}_w(\overline{\alpha}).$$

Similar to the case without intersection, for each $I \in \mathsf{Ind}(\alpha)$, there is exactly one pair $(\alpha_I, I) \in \mathsf{Pre}^{\mathsf{I}}(\overline{\alpha})$. For $\alpha \in \mathsf{RE}$, we have that $\mathsf{Pre}^{\mathsf{I}}(\overline{\alpha})$ satisfies (2) considering that $\mathsf{Pre}^{\mathsf{I}}(\sigma_i) = \{(\varepsilon, \{i\})\}$. The following lemma caracterizes $\mathsf{Pre}^{\mathsf{I}}$ for expressions with intersection.

Lemma 34.

$$\begin{aligned}\mathsf{Pre}^{\mathsf{I}}(\alpha_1 \cap \alpha_2) \setminus \{\varepsilon\} \;&= \{\, (\alpha \cap \alpha', I_1 \cup I_2) \mid \\ & (\alpha, I_1) \in \mathsf{Pre}^{\mathsf{I}}(\alpha_1),\ (\alpha', I_2) \in \mathsf{Pre}^{\mathsf{I}}(\alpha_2),\ \ell(I_1) = \ell(I_2)\,\}.\end{aligned}$$

Proof. We show that for all $w \in \mathsf{Ind}(\alpha_1 \cap \alpha_2)^\star$ one has $(\beta, I) \in \mathsf{p}^{\mathsf{I}}{}_w(\alpha_1 \cap \alpha_2)$ if and only if $\beta = \beta_1 \cap \beta_2$, $I = I_1 \cup I_2$, $(\beta_1, I_1) \in \mathsf{p}^{\mathsf{I}}{}_{w_1}(\alpha_1)$, $(\beta_2, I_2) \in \mathsf{p}^{\mathsf{I}}{}_{w_2}(\alpha_2)$, and $\ell(I_1) = \ell(I_2)$, where $I_1 = I\big|_{\beta_1}$, $I_2 = I\big|_{\beta_2}$, $w_1 = w\big|_{\beta_1}$, and $w_2 = w\big|_{\beta_2}$. The proof is by induction on $|w|$. For $w = \varepsilon$ the result follows from the definition of R^{I}. Furthermore, we have $(\beta, I) \in \mathsf{p}^{\mathsf{I}}{}_{Jw}(\alpha_1 \cap \alpha_2)$ iff there is a pair $(\alpha', J) \in \mathsf{p}^{\mathsf{I}}{}_w(\alpha_1 \cap \alpha_2)$ such that $(\beta, I) \in \mathsf{R}^{\mathsf{I}}{}_\varepsilon(\alpha')$. Thus, by the induction hypothesis $\alpha' = \alpha'_1 \cap \alpha'_2$, $J = J_1 \cup J_2$, $(\alpha'_1, J_1) \in \mathsf{p}^{\mathsf{I}}{}_{w_1}(\alpha_1)$, $(\alpha'_2, J_2) \in \mathsf{p}^{\mathsf{I}}{}_{w_2}(\alpha_2)$, $\ell(J_1) = \ell(J_2)$, where $J_1 = J\big|_{\alpha_1}$, $J_2 = J\big|_{\alpha_2}$, $w_1 = w\big|_{\alpha_1}$, and $w_2 = w\big|_{\alpha_2}$. But $(\beta, I) \in \mathsf{R}^{\mathsf{I}}{}_\varepsilon(\alpha'_1 \cap \alpha'_2)$ means that $\beta = \beta_1 \cap \beta_2$, $I = I_1 \cup I_2$, $(\beta_1, I_1) \in \mathsf{R}^{\mathsf{I}}(\alpha'_1)$, $(\beta_2, I_2) \in \mathsf{R}^{\mathsf{I}}(\alpha'_2)$, and $\ell(I_1) = \ell(I_2)$. And, finally, that $(\beta_1, I_1) \in \mathsf{p}^{\mathsf{I}}{}_{J_1 w_1}(\alpha_1)$ and $(\beta_2, I_2) \in \mathsf{p}^{\mathsf{I}}{}_{J_2 w_2}(\alpha_2)$. □

The next lemma establishes the connection of $\mathsf{Pre}^{\mathsf{I}}$ with Pre and can be proved by induction on the structure of α.

Lemma 35. $\mathsf{Pre}^+(\alpha) = \{\, \overline{\alpha'}\sigma \mid (\alpha', I) \in \mathsf{Pre}^{\mathsf{I}}(\overline{\alpha}) \wedge \sigma = \ell(I)\,\}$.

The following result, which is analogous to Lemma 21, relates $\mathcal{A}_{\mathrm{POS}}$ with $\mathcal{A}_{\mathrm{Pre}}$.

Lemma 36. *For $\alpha \in RE_\cap$ we have*

(i) $\mathsf{First}(\alpha) = \{\, I \mid (\beta, I) \in \mathsf{Pre}^{\mathsf{I}}(\overline{\alpha}) \wedge \varepsilon(\beta) = \varepsilon \,\}$,

(ii) $\mathsf{Last}(\alpha) = \{\, I \mid (\beta, I) \in \mathsf{R}^{\mathsf{I}}(\overline{\alpha}) \,\}$,

(iii) *For all* $(\beta, I), (\gamma, J) \in \mathsf{Pre}^{\mathsf{I}}(\overline{\alpha})$, $\sigma^I = \ell(I)$ *and* $\sigma^J = \ell(J)$, *one has* $(\overline{\beta}\sigma^I, \sigma^J, \overline{\gamma}\sigma^J) \in \delta_{\mathrm{Pre}}$ *if and only if* $J \in \mathsf{Follow}(\alpha, I)$.

Proof. By induction on $\overline{\alpha} \in \mathsf{RE}_{\cap}$ (as defined in the beginning of this section). If α does not contain the intersection operator, then the result follows by Lemma 21. In fact, the definitions are identical except for the case of σ_i, where instead of a position i we have the singleton $I = \{i\}$. We will check the results for the case of an intersection expression $\alpha_1 \cap \alpha_2$.

(i) By Lemma 34 we have

$$\begin{aligned}\mathsf{First}(\alpha_1 \cap \alpha_2) &= \{\, I_1 \cup I_2 \mid \ell(I_1) = \ell(I_2) \wedge I_1 \in \mathsf{First}(\alpha_1) \wedge I_2 \in \mathsf{First}(\alpha_2) \,\} \\ &= \{\, I_1 \cup I_2 \mid (\alpha, I_1) \in \mathsf{Pre}^{\mathsf{I}}(\alpha_1) \wedge (\alpha', I_2) \in \mathsf{Pre}^{\mathsf{I}}(\alpha_2) \\ &\qquad \wedge \ell(I_1) = \ell(I_2) \wedge \varepsilon(\alpha) = \varepsilon(\alpha') = \varepsilon \,\} \\ &= \{\, I_1 \cup I_2 \mid (\alpha \cap \alpha', I_1 \cup I_2) \in \mathsf{Pre}^{\mathsf{I}}(\alpha_1 \cap \alpha_2) \wedge \varepsilon(\alpha \cap \alpha') = \varepsilon \,\}.\end{aligned}$$

(ii) We have

$$\begin{aligned}\mathsf{Last}(\alpha_1 \cap \alpha_2) &= \mathsf{Last}(\alpha_1) \otimes \mathsf{Last}(\alpha_2) \\ &= \{\, I_1 \mid (\beta_1, I_1) \in \mathsf{R}^{\mathsf{I}}(\alpha_1) \,\} \otimes \{\, I_2 \mid (\beta_2, I_2) \in \mathsf{R}^{\mathsf{I}}(\alpha_2) \,\} \\ &= \{\, I_1 \cup I_2 \mid \ell(I_1) = \ell(I_2) \wedge (\beta_1, I_1) \in \mathsf{R}^{\mathsf{I}}(\alpha_1) \wedge (\beta_2, I_2) \in \mathsf{R}^{\mathsf{I}}(\alpha_2) \,\} \\ &= \{\, I_1 \cup I_2 \mid (\beta_1 \cap \beta_2, I_1 \cup I_2) \in \mathsf{R}^{\mathsf{I}}(\alpha_1 \cap \alpha_2) \,\}.\end{aligned}$$

(iii) Here we use Lemma 35. Let $(\beta_1 \cap \beta_2, I_1 \cup I_2)$, $(\gamma_1 \cap \gamma_2, J_1 \cup J_2) \in \mathsf{Pre}^{\mathsf{I}}(\alpha_1 \cap \alpha_2)$. Let $\sigma^I = \ell(I_1) = \ell(I_2)$ and let $\sigma^J = \ell(J_1) = \ell(J_2)$. Then,

$$\begin{aligned}&(\overline{\beta_1 \cap \beta_2}\sigma^I, \sigma^J, \overline{\gamma_1 \cap \gamma_2}\sigma^J) \in \delta_{\mathrm{Pre}}(\alpha_1 \cap \alpha_2) \\ &\iff \overline{\gamma_1 \cap \gamma_2}\sigma^J \in \mathsf{Pre}^{+}(\alpha_1 \cap \alpha_2) \wedge \overline{\beta_1 \cap \beta_2}\sigma^I \in \mathsf{R}_{\varepsilon}(\overline{\gamma_1 \cap \gamma_2}) \\ &\iff (\gamma_1 \cap \gamma_2, J) \in \mathsf{Pre}^{\mathsf{I}}(\alpha_1 \cap \alpha_2) \wedge \overline{\beta_1}\sigma^I \in \mathsf{R}_{\varepsilon}(\overline{\gamma_1}) \wedge \overline{\beta_2}\sigma^I \in \mathsf{R}_{\varepsilon}(\overline{\gamma_2}) \\ &\iff (\gamma_1, J_1) \in \mathsf{Pre}^{\mathsf{I}}(\alpha_1) \wedge (\gamma_2, J_2) \in \mathsf{Pre}^{\mathsf{I}}(\alpha_2) \wedge \overline{\beta_1}\sigma^I \in \mathsf{R}_{\varepsilon}(\overline{\gamma_1}) \wedge \overline{\beta_2}\sigma^I \in \mathsf{R}_{\varepsilon}(\overline{\gamma_2}) \\ &\iff (\overline{\beta_1}\sigma^I, \sigma^J, \overline{\gamma_1}\sigma^J) \in \delta_{\mathrm{Pre}}(\alpha_1) \wedge (\overline{\beta_2}\sigma^I, \sigma^J, \overline{\gamma_2}\sigma^J) \in \delta_{\mathrm{Pre}}(\alpha_2) \\ &\iff J_1 \in \mathsf{Follow}(\alpha_1, I_1) \wedge J_2 \in \mathsf{Follow}(\alpha_2, I_2) \\ &\iff J_1 \cup J_2 \in \mathsf{Follow}(\alpha_1 \cap \alpha_2, I_1 \cup I_2).\end{aligned}$$

□

Finally, consider the relation $\equiv_\ell$ defined in $\mathsf{Ind}(\alpha) \cup \{\{0\}\}$ such that $\{0\} \equiv_\ell \{0\}$ and for $I, J \in \mathsf{Ind}(\alpha)$,

$$I \equiv_\ell J \iff \ell(I) = \ell(J) \text{ and } \overline{\alpha_I} \doteq \overline{\alpha_J}.$$

This relation is left-invariant w.r.t. $\mathcal{A}_{\mathrm{POS}}(\alpha)$. We conclude that, also for expressions with intersection, one has the following result.

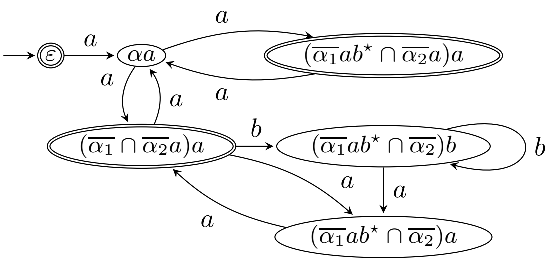

ε
a
αa
a
$(\overline{\alpha_1}ab^\star \cap \overline{\alpha_2}a)a$
a
a
a
$(\overline{\alpha_1} \cap \overline{\alpha_2}a)a$
b
$(\overline{\alpha_1}ab^\star \cap \overline{\alpha_2})b$
b
a
a
a
$(\overline{\alpha_1}ab^\star \cap \overline{\alpha_2})a$

Figure 8: $\mathcal{A}_{\mathrm{Pre}}((ab^\star a + a)^\star \cap (aa+b)^\star)$

Proposition 37. $\mathcal{A}_{\mathrm{Pre}}(\alpha) \simeq \mathcal{A}_{\mathrm{POS}}(\alpha)/{\equiv_\ell}$.

Proof. Let $\varphi_\ell : (\mathsf{Ind}(\alpha) \cup \{\{0\}\})/{\equiv_\ell} \to \mathsf{Pre}(\alpha)$ defined by $\varphi_\ell([\{0\}]) = \varepsilon$ and let $\varphi_\ell([I]) = \overline{\alpha_I}\sigma$, where $\sigma = \ell(I)$. It is obvious that φ_ℓ is a bijection and, using Lemma 36, defines an isomorphism between $\mathcal{A}_{\mathrm{POS}}(\alpha)/{\equiv_\ell}$ and $\mathcal{A}_{\mathrm{Pre}}(\alpha)$. □

Example 38. For α from Example 33, with $\overline{\alpha} = \alpha_1 \cap \alpha_2$, for $\alpha_1 = (a_1b_2^\star a_3 + a_4)^\star$ and $\alpha_2 = (a_5a_6 + b_7)^\star$ we have

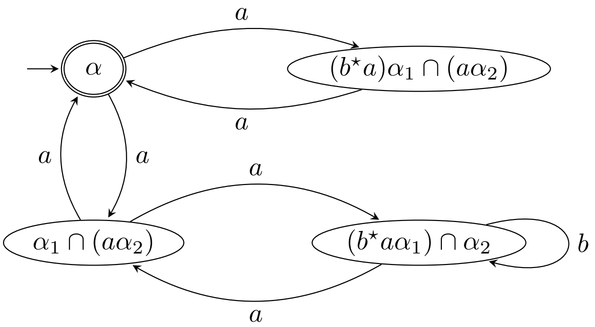
$$\begin{aligned}
\mathsf{Pre}^{\mathsf{I}}(\alpha_1) &= \{\varepsilon, (\alpha_1, \{1\}), (\alpha_1, \{4\}), (\alpha_1 a_1 b_2^\star, \{3\}), (\alpha_1 a_1 b_2^\star, \{2\})\},\\
\mathsf{Pre}^{\mathsf{I}}(\alpha_2) &= \{\varepsilon, (\alpha_2, \{5\}), (\alpha_2, \{7\}), (\alpha_2 a_5, \{6\})\},\\
\mathsf{Pre}^{\mathsf{I}}(\overline{\alpha}) &= \{\varepsilon, (\alpha_1 \cap \alpha_2, \{1,5\}), (\alpha_1 \cap \alpha_2, \{4,5\}),\\
&\quad (\alpha_1 \cap \alpha_2 a_5, \{1,6\}), (\alpha_1 \cap \alpha_2 a_5, \{4,6\}),\\
&\quad (\alpha_1 a_1 b_2^\star \cap \alpha_2 a_5, \{3,6\}),\\
&\quad (\alpha_1 a_1 b_2^\star \cap \alpha_2, \{3,5\}), (\alpha_1 a_1 b_2^\star \cap \alpha_2, \{2,7\})\}.
\end{aligned}$$

By inspection of the expressions in these pairs and since $\ell(\{1,5\}) = \ell(\{4,5\}) = a$ and also $\ell(\{1,6\}) = \ell(\{4,6\}) = a$, we conclude that an automaton isomorphic to $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ can be obtained from $\mathcal{A}_{\mathrm{POS}}(\alpha)$, merging the states with labels $\{1,5\}$ and $\{4,5\}$, as well as the states with labels $\{1,6\}$ and $\{4,6\}$. The first two states merge to a non final state with label αa, while the latter merge to a final state with label $(\overline{\alpha_1} \cap \overline{\alpha_2}a)a$. The automaton $\mathcal{A}_{\mathrm{Pre}}(\alpha)$ is depicted in Figure 8.

In a symmetric way an extension of L to expressions with intersection can be given, as well as a function L^{I}, that applies to marked expressions. Using these definitions, one shows that also for intersection the relationship between $\mathcal{A}_{\overleftarrow{\mathrm{POS}}}$ and $\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}$ is $\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}(\alpha) \simeq \mathcal{A}_{\overleftarrow{\mathrm{POS}}}(\alpha)/{\equiv_\ell}$. Using the counterpart of (13) for L we have that Lemma 29(ii) is also true stablishing a relation between $\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}$ and $\mathcal{A}_{\mathrm{PD}}$. Finaly, the right-partial derivative automaton $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ can be extended exactly as $\mathcal{A}_{\mathrm{PD}}$ for expressions with intersection, and we conclude that Lemma 29(i) also holds in this case.

Example 39. In Figure 9 is depicted the $\mathcal{A}_{\mathrm{PD}}(\alpha)$ for $\alpha = (ab^\star a + a)^\star \cap (aa+b)^\star$. As $\alpha = \alpha^{\mathrm{R}}$,we have $\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}(\alpha) \simeq \mathcal{A}_{\mathrm{Pre}}(\alpha)^{\mathrm{R}}$ and $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha) \simeq \mathcal{A}_{\mathrm{PD}}(\alpha)^{\mathrm{R}}$.

Figure 9: $\mathcal{A}_{\mathrm{PD}}((ab^\star a + a)^\star \cap (aa+b)^\star)$ where $\alpha_1 = (ab^\star a + a)^\star$ and $\alpha_2 = (aa+b)^\star$.


α
$(b^\star a)\alpha_1 \cap (a\alpha_2)$
$\alpha_1 \cap (a\alpha_2)$
$(b^\star a\alpha_1) \cap \alpha_2$
a
a
a
a
a
a
b

7. Conclusion

An upper-bound for the asymptotical average number of states of $\mathcal{A}_{\mathrm{PD}}(\alpha)$ for $\alpha \in \mathsf{RE}_\cap$ of size n is $(1.056 + o(1))^n$ (cf. Bastos et al. [2]). The same upper-bound holds for $\mathcal{A}_{\mathrm{POS}}$ and thus also for $\mathcal{A}_{\mathrm{Pre}}$. In Broda et.al [3] several automaton constructions were based on the Follow set and the Select function. Considering the definition of the set Follow for expressions in $\mathsf{RE}_\cap$, all those constructions can be extended and the same relationships hold. With the extension of $\mathcal{A}_{\mathrm{Pre}}$ to $\mathsf{RE}_\cap$ and the results here presented we conclude that the relationships presented in that taxonomy also hold for $\mathcal{A}_{\mathrm{Pre}}$, $\mathcal{A}_{\overleftarrow{\mathrm{Pre}}}$ and their determinisations.

Acknowledgement

The authors thank the anonymous referees for their kind suggestions.

References

[1] V. M. Antimirov, Partial Derivatives of Regular Expressions and Finite Automaton Constructions. *Theoret. Comput. Sci.* **155** (1996) 2, 291–319.

[2] R. Bastos, S. Broda, A. Machiavelo, N. Moreira, R. Reis, On the Average Complexity of Partial Derivative Automata for Semi-extended Expressions. *Journal of Automata, Languages and Combinatorics* **22** (2017) 1–3, 5–28.

[3] S. Broda, M. Holzer, E. Maia, N. Moreira, R. Reis, A Mesh of Automata. *Inf. Comput.* **265** (2019), 94–111.

[4] S. Broda, A. Machiavelo, N. Moreira, R. Reis, On the Average State Complexity of Partial Derivative Automata: an Analytic Combinatorics Approach. *Internat. J. Found. Comput. Sci.* **22** (2011) 7, 1593–1606.

[5] S. Broda, A. Machiavelo, N. Moreira, R. Reis, On the Average Size of Glushkov and Partial Derivative Automata. *Internat. J. Found. Comput. Sci.* **23** (2012) 5, 969–984.

[6] S. Broda, A. Machiavelo, N. Moreira, R. Reis, Position automaton construction for regular expressions with intersection. In: S. Brlek, C. Reutenauer (eds.), *20th DLT*. LNCS 9840, Springer, 2016, 51–63.

[7] S. Broda, A. Machiavelo, N. Moreira, R. Reis, Automata for regular expressions with shuffle. *Inf. Comput.* **259** (2018) 2, 162–173.

[8] S. Broda, A. Machiavelo, N. Moreira, R. Reis, Position Automata for Semi-extended Expressions. *Journal of Automata, Languages and Combinatorics* **23** (2018) 1–3, 39–65.

[9] J. M. Champarnaud, D. Ziadi, From Mirkin's Prebases to Antimirov's Word Partial Derivatives. *Fundam. Inform.* **45** (2001) 3, 195–205.

[10] J. M. Champarnaud, D. Ziadi, Canonical derivatives, partial derivatives and finite automaton constructions. *Theoret. Comput. Sci.* **289** (2002), 137–163.

[11] W. G. Cochran, *Sampling Techniques*. 3rd edition, John Wiley and Sons, 1977.

[12] P. Flajolet, R.Sedgewick, *Analytic Combinatorics*. CUP, 2008.

[13] D. Giammarresi, J.-L. Ponty, D. Wood, *Glushkov and Thompson Constructions: A Synthesis*. HKUST TCSC-98-11, The Department of Science & Engineering, Theoretical Cmputer Science Group, The Hong Kong University of Science and Technology, 1998.

[14] V. M. Glushkov, The abstract theory of automata. *Russian Math. Surveys* **16** (1961), 1–53.

[15] L. Ilie, S. Yu, Follow automata. *Inf. Comput.* **186** (2003) 1, 140–162.

[16] E. Maia, *On the Descriptional Complexity of Some Operations and Simulations of Regular Models*. Ph.D. thesis, Faculdade de Ciências da Universidade do Porto, 2015.

[17] E. Maia, N. Moreira, R. Reis, Prefix and Right-Partial Derivative Automata. In: M. Soskova, V. Mitrana (eds.), *11th CiE*. LNCS 9136, Springer, 2015, 258–267.

[18] B. G. Mirkin, An Algorithm for Constructing a base in a Language of Regular Expressions. *Engineering Cybernetics* **5** (1966), 51—57.

[19] C. Nicaud, On the Average Size of Glushkov's Automata. In: A. Dediu, A.-M. Ionescu, C. M. Vide (eds.), *3rd LATA*. LNCS 5457, Springer, 2009, 626–637.

[20] K. Thompson, Regular expression search algorithm. *Commun. ACM* **11** (1968) 6, 410–422.

[21] H. Yamamoto, A new finite automaton construction for regular expressions. In: S. Bensch, R. Freund, F. Otto (eds.), *6th NCMA*. books@ocg.at 304, Österreichische Computer Gesellschaft, 2014, 249–264.

A. Inductive Characterization of $\mathcal{A}_{\mathrm{PD}}$

Mirkin's construction of the $\mathcal{A}_{\mathrm{PD}}(\alpha)$ is based on the existence of a set of expressions $\pi(\alpha) = \{\alpha_1, \dots, \alpha_n\}$ that satisfies a system of equations

$$\alpha_i = \sigma_1\alpha_{i1} + \cdots + \sigma_k\alpha_{ik} + \varepsilon(\alpha_i),$$

with $\alpha_0 \doteq \alpha$ and such that α_{ij} are linear combinations of elements of $\pi(\alpha)$, for all $i \in [1, n]$ and $j \in [1, k]$. The set $\pi(\alpha)$ can be obtained inductively on the structure of

α as follows:

$$\begin{aligned} \pi(\emptyset) &= \emptyset, & \pi(\alpha+\beta) &= \pi(\alpha) \cup \pi(\beta), \\ \pi(\varepsilon) &= \emptyset, & \pi(\alpha\beta) &= \pi(\alpha)\beta \cup \pi(\beta), \\ \pi(\sigma) &= \{\varepsilon\}, & \pi(\alpha^\star) &= \pi(\alpha)\alpha^\star. \end{aligned} \tag{15}$$

Champarnaud and Ziadi [9] proved that $\mathsf{PD}(\alpha) = \pi(\alpha) \cup \{\alpha\}$ and that the Antimirov and the Mirkin constructions lead to the same automaton. As noted by Broda et al. [5], Mirkin's algorithm to compute $\pi(\alpha)$ also provides an inductive definition of the set of transitions of $\mathcal{A}_{\mathrm{PD}}(\alpha)$. Let $\varphi(\alpha) = \{\, (\sigma,\gamma) \mid \gamma \in \partial_\sigma(\alpha) \wedge \sigma \in \Sigma \,\}$ and $\lambda(\alpha) = \{\, \alpha' \mid \alpha' \in \pi(\alpha) \wedge \varepsilon(\alpha') = \varepsilon \,\}$, where both sets can be inductively defined as follows:

$$\begin{aligned} \varphi(\emptyset) &= \emptyset, & \varphi(\alpha+\beta) &= \varphi(\alpha) \cup \varphi(\beta), \\ \varphi(\varepsilon) &= \emptyset, & \varphi(\alpha\beta) &= \varphi(\alpha)\beta \cup \varepsilon(\alpha)\varphi(\beta), \\ \varphi(\sigma) &= \{(\sigma,\varepsilon)\} & \varphi(\alpha^\star) &= \varphi(\alpha)\alpha^\star, \end{aligned} \tag{16}$$

$$\begin{aligned} \lambda(\emptyset) &= \emptyset, & \lambda(\alpha+\beta) &= \lambda(\alpha) \cup \lambda(\beta), \\ \lambda(\varepsilon) &= \emptyset, & \lambda(\alpha\beta) &= \varepsilon(\beta)\lambda(\alpha)\beta \cup \lambda(\beta), \\ \lambda(\sigma) &= \{\varepsilon\}, & \lambda(\alpha^\star) &= \lambda(\alpha)\alpha^\star. \end{aligned} \tag{17}$$

In the above definitions, for any tuple (σ,τ) and expression β, $(\sigma,\tau)\beta = (\sigma,\tau\beta)$, $\beta(\sigma,\tau) = (\sigma,\beta\tau)$ and these also extend to sets of tuples. The set $\mathsf{Tr}(\alpha)$ of transitions is inductively defined by:

$$\begin{aligned} \mathsf{Tr}(\emptyset) = \mathsf{Tr}(\varepsilon) = \mathsf{Tr}(\sigma) &= \emptyset,\ \sigma \in \Sigma, \\ \mathsf{Tr}(\alpha+\beta) &= \mathsf{Tr}(\alpha) \cup \mathsf{Tr}(\beta), \\ \mathsf{Tr}(\alpha\beta) &= \mathsf{Tr}(\alpha)\beta \cup \mathsf{Tr}(\beta) \cup (\lambda(\alpha)\beta \times \varphi(\beta)), \\ \mathsf{Tr}(\alpha^\star) &= \mathsf{Tr}(\alpha)\alpha^\star \cup (\lambda(\alpha) \times \varphi(\alpha))\alpha^\star, \end{aligned}$$

where the result of the × operation is seen as a set of triples (α',σ,β') and the concatenation of a transition (α,σ,β) with a regular expression γ is defined by $(\alpha,\sigma,\beta)\gamma = (\alpha\gamma,\sigma,\beta\gamma)$ and $\gamma(\alpha,\sigma,\beta) = (\gamma\alpha,\sigma,\gamma\beta)$. These also extends to sets of transitions. Then we can inductively construct the partial derivative automaton of α using the following results.

Proposition 40. $\mathsf{Tr}(\alpha) = \{(\tau,\sigma,\tau') \mid \tau \in \mathsf{PD}^+(\alpha) \wedge \tau' \in \partial_\sigma(\tau) \wedge \sigma \in \Sigma\}$.

Proof. We know [7, 9] that $\mathsf{PD}^+(\alpha) = \pi(\alpha)$. Thus we want to prove that $\mathsf{Tr}(\alpha) = \{(\tau,\sigma,\tau') \mid \tau \in \pi(\alpha) \wedge \tau' \in \partial_\sigma(\tau) \wedge \sigma \in \Sigma\}$. Let us proceed by induction on the structure of α. For the base cases the equality is obvious. We assume that $\sigma \in \Sigma$.

Let $\alpha \doteq \alpha_1 + \alpha_2$. Then

$$\begin{aligned}
&\{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1+\alpha_2) \wedge \tau' \in \partial_\sigma(\tau) \,\}\\
&= \{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1) \cup \pi(\alpha_2) \wedge \tau' \in \partial_\sigma(\tau) \,\}\\
&= \{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1) \wedge \tau' \in \partial_\sigma(\tau) \,\} \cup \{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_2) \wedge \tau' \in \partial_\sigma(\tau) \,\}\\
&= \mathsf{Tr}(\alpha_1) \cup \mathsf{Tr}(\alpha_2) = \mathsf{Tr}(\alpha_1+\alpha_2).
\end{aligned}$$

Let $\alpha \doteq \alpha_1\alpha_2$. Then

$$\begin{aligned}
&\{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1\alpha_2) \wedge \tau' \in \partial_\sigma(\tau) \,\}\\
&= \{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1)\alpha_2 \cup \pi(\alpha_2) \wedge \tau' \in \partial_\sigma(\tau) \,\}\\
&= \{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1)\alpha_2 \wedge \tau' \in \partial_\sigma(\tau) \,\} \cup \{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_2) \wedge \tau' \in \partial_\sigma(\tau) \,\}.
\end{aligned}$$

By the induction hypothesis, we have $\{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_2) \wedge \tau' \in \partial_\sigma(\tau) \,\} = \mathsf{Tr}(\alpha_2)$. On the other hand,

$$\begin{aligned}
&\{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1)\alpha_2 \wedge \tau' \in \partial_\sigma(\tau) \,\}\\
&= \{\, (\tau,\sigma,\tau') \mid \tau \doteq \alpha_1'\alpha_2 \wedge \alpha_1' \in \pi(\alpha_1) \wedge \tau' \in \partial_\sigma(\alpha_1'\alpha_2) \,\}\\
&= \{\, (\tau,\sigma,\tau') \mid \tau \doteq \alpha_1'\alpha_2 \wedge \alpha_1' \in \pi(\alpha_1) \wedge \tau' \in \partial_\sigma(\alpha_1')\alpha_2 \,\}\\
&\cup \{\, (\tau,\sigma,\tau') \mid \tau \doteq \alpha_1'\alpha_2 \wedge \alpha_1' \in \pi(\alpha_1) \wedge \varepsilon(\alpha_1') = \varepsilon \wedge \tau' \in \partial_\sigma(\alpha_2) \,\}\\
&= \mathsf{Tr}(\alpha_1)\alpha_2 \cup (\lambda(\alpha_1)\alpha_2 \times \varphi(\alpha_2)).
\end{aligned}$$

We conclude that $\{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1\alpha_2) \wedge \tau' \in \partial_\sigma(\tau) \,\} = \mathsf{Tr}(\alpha_1\alpha_2)$. Let $\alpha \doteq \alpha_1^\star$. Then

$$\begin{aligned}
&\{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1^\star) \wedge \tau' \in \partial_\sigma(\tau) \,\} = \{\, (\tau,\sigma,\tau') \mid \tau \in \pi(\alpha_1)\alpha_1^\star \wedge \tau' \in \partial_\sigma(\tau) \,\}\\
&= \{\, (\tau,\sigma,\tau') \mid \tau \doteq \alpha_1'\alpha_1^\star \wedge \alpha_1' \in \pi(\alpha_1) \wedge \tau' \in \partial_\sigma(\tau) \,\}\\
&= \{\, (\tau,\sigma,\tau') \mid \tau \doteq \alpha_1'\alpha_1^\star \wedge \alpha_1' \in \pi(\alpha_1) \wedge \tau' \in \partial_\sigma(\alpha_1')\alpha_1^\star \,\}\\
&\cup \{\, (\tau,\sigma,\tau') \mid \tau \doteq \alpha_1'\alpha_1^\star \wedge \alpha_1' \in \pi(\alpha_1) \wedge \varepsilon(\alpha_1') = \varepsilon \wedge \tau' \in \partial_\sigma(\alpha_1^\star) \,\}\\
&= \mathsf{Tr}(\alpha_1)\alpha_1^\star \cup (\lambda(\alpha_1) \times \varphi(\alpha_1))\alpha_1^\star = \mathsf{Tr}(\alpha_1^\star).
\end{aligned}$$

□

Proposition 41. $\mathcal{A}_{\mathrm{PD}}(\alpha) = \langle \pi(\alpha) \cup \{\alpha\}, \Sigma, \{\alpha\} \times \varphi(\alpha) \cup \mathsf{Tr}(\alpha), \alpha, \lambda(\alpha) \cup \varepsilon(\alpha)\{\alpha\} \rangle$.

Proof. We want to prove that the right-hand side of this equality corresponds to the definition of $\mathcal{A}_{\mathrm{PD}}$ previously presented. For the set of states in both definitions of the automata note that $\mathsf{PD}(\alpha) = \pi(\alpha) \cup \{\alpha\}$. Also the initial and final states coincide. The transition function

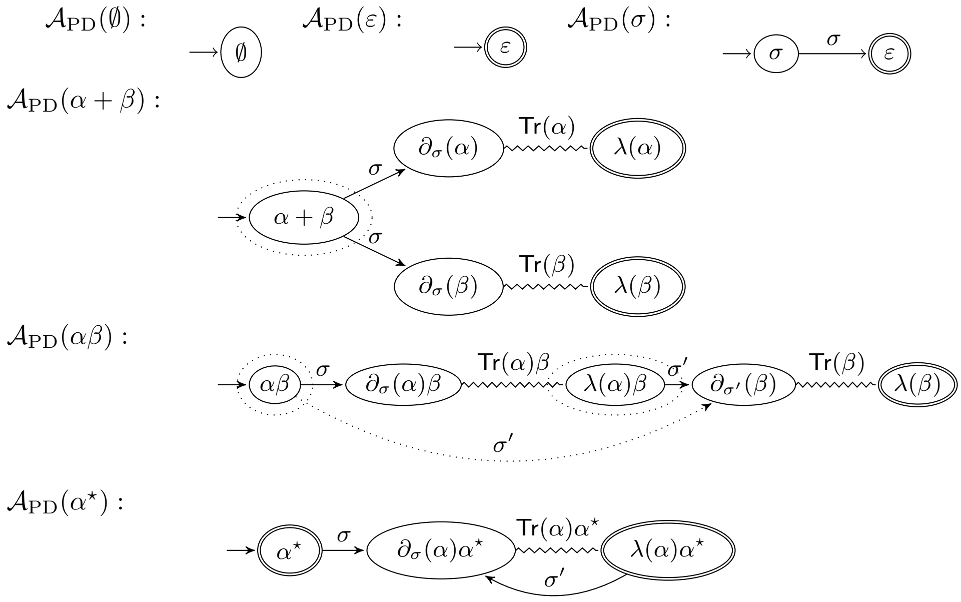
$$\delta_{\mathrm{PD}} = \{\, (\tau,\sigma,\tau') \mid \tau \in \mathsf{PD}(\alpha) \wedge \tau' \in \partial_\sigma(\tau) \wedge \sigma \in \Sigma \,\}$$

can be written as the following union:

$$\{\, (\alpha,\sigma,\tau') \mid \tau' \in \partial_\sigma(\alpha) \wedge \sigma \in \Sigma \,\} \cup \{\, (\tau,\sigma,\tau') \mid \tau \in \mathsf{PD}^+(\alpha) \wedge \tau' \in \partial_\sigma(\tau) \wedge \sigma \in \Sigma \,\}.$$

The first set is clearly equal to $\{\alpha\} \times \varphi(\alpha)$, and by Proposition 40 the second set is equal to $\mathsf{Tr}(\alpha)$. Figure 10 illustrates this inductive construction, where we assume that states are merged whenever they correspond to equal REs. □

Figure 10: Inductive Construction of $\mathcal{A}_{\mathrm{PD}}$. Dotted states are final if ε belongs to the language of their label. If $\varepsilon(\beta) = \varepsilon$ the dotted arrow exists.


$\mathcal{A}_{\mathrm{PD}}(\emptyset)$:
$\emptyset$
$\mathcal{A}_{\mathrm{PD}}(\varepsilon)$:
ε
$\mathcal{A}_{\mathrm{PD}}(\sigma)$:
σ
σ
ε
$\mathcal{A}_{\mathrm{PD}}(\alpha+\beta)$:
$\alpha+\beta$
σ
σ
$\partial_\sigma(\alpha)$
$\mathsf{Tr}(\alpha)$
$\lambda(\alpha)$
$\partial_\sigma(\beta)$
$\mathsf{Tr}(\beta)$
$\lambda(\beta)$
$\mathcal{A}_{\mathrm{PD}}(\alpha\beta)$:
$\alpha\beta$
σ
$\partial_\sigma(\alpha)\beta$
$\mathsf{Tr}(\alpha)\beta$
$\lambda(\alpha)\beta$
σ'
$\partial_{\sigma'}(\beta)$
$\mathsf{Tr}(\beta)$
$\lambda(\beta)$
σ'
$\mathcal{A}_{\mathrm{PD}}(\alpha^\star)$:
$\alpha^\star$
σ
$\partial_\sigma(\alpha)\alpha^\star$
$\mathsf{Tr}(\alpha)\alpha^\star$
$\lambda(\alpha)\alpha^\star$
σ'

B. Inductive Construction of $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$

The set $\overleftarrow{\pi}(\alpha)$ can be defined inductively as in (15) except for the following cases:
$\overleftarrow{\pi}(\alpha\beta) = \alpha\overleftarrow{\pi}(\beta) \cup \overleftarrow{\pi}(\alpha)$ and $\overleftarrow{\pi}(\alpha^\star) = \alpha^\star\overleftarrow{\pi}(\alpha)$.
The solution of the system of equations also allows to inductively define the transition function and the set of initial states of $\mathcal{A}_{\overleftarrow{\mathrm{PD}}}$ [17, 16]. As before, we consider the sets $\overleftarrow{\varphi}(\alpha) = \{(\gamma,\sigma) \mid \gamma \in \overleftarrow{\partial}_\sigma(\alpha), \sigma \in \Sigma\}$ and $\overleftarrow{\lambda}(\alpha) = \{\alpha' \mid \alpha' \in \overleftarrow{\pi}(\alpha),\ \varepsilon(\alpha') = \varepsilon\}$, which are inductively defined as in (16) and (17), respectively, except for following cases:

$$\begin{aligned}
&\overleftarrow{\varphi}(\sigma) = \{(\varepsilon,\sigma)\}, \\
&\overleftarrow{\varphi}(\alpha^\star) = \alpha^\star\overleftarrow{\varphi}(\alpha), && \overleftarrow{\lambda}(\alpha^\star) = \alpha^\star\overleftarrow{\lambda}(\alpha), \\
&\overleftarrow{\varphi}(\alpha\beta) = \alpha\overleftarrow{\varphi}(\beta) \cup \varepsilon(\beta)\overleftarrow{\varphi}(\alpha), && \overleftarrow{\lambda}(\alpha\beta) = \varepsilon(\alpha)\alpha\overleftarrow{\lambda}(\beta) \cup \overleftarrow{\lambda}(\alpha).
\end{aligned}$$

The set of transitions is $\overleftarrow{\varphi}(\alpha) \times \{\alpha\} \cup \overleftarrow{\mathsf{Tr}}(\alpha)$ where $\overleftarrow{\mathsf{Tr}}$ is defined as Tr except for:

$$\begin{aligned}
&\overleftarrow{\mathsf{Tr}}(\alpha\beta) = \overleftarrow{\mathsf{Tr}}(\alpha) \cup \alpha\overleftarrow{\mathsf{Tr}}(\beta) \cup (\overleftarrow{\varphi}(\alpha) \times (\alpha\overleftarrow{\lambda}(\beta))), \\
&\overleftarrow{\mathsf{Tr}}(\alpha^\star) = \alpha^\star\overleftarrow{\mathsf{Tr}}(\alpha) \cup \alpha^\star(\overleftarrow{\varphi}(\alpha) \times \overleftarrow{\lambda}(\alpha)).
\end{aligned}$$

Then, we have

$$\mathcal{A}_{\overleftarrow{\mathrm{PD}}}(\alpha) = \langle \overleftarrow{\pi}(\alpha) \cup \{\alpha\}, \Sigma, \overleftarrow{\varphi}(\alpha) \times \{\alpha\} \cup \overleftarrow{\mathsf{Tr}}(\alpha), \overleftarrow{\lambda}(\alpha) \cup \varepsilon(\alpha)\{\alpha\}, \alpha\rangle.$$