Position Automaton Construction for Regular Expressions with Intersection[⋆]

Sabine Broda, António Machiavelo, Nelma Moreira, and Rogério Reis

CMUP, Faculdade de Ciências da Universidade do Porto, Portugal
sbb@dcc.fc.up.pt,ajmachia@fc.up.pt,{nam,rvr}@dcc.fc.up.pt

Abstract. Positions and derivatives are two essential notions in the conversion methods from regular expressions to equivalent finite automata. Partial derivative based methods have recently been extended to regular expressions with intersection. In this paper, we present a position automaton construction for those expressions. This construction generalizes the notion of position making it compatible with intersection. The resulting automaton is homogeneous and has the partial derivative automaton as its quotient.

1 Introduction

The position automaton, introduced by Glushkov [12], permits the conversion of a simple regular expression (involving only the sum, concatenation and star operations) into an equivalent nondeterministic finite automaton (NFA) without ε-transitions. The states in the position automaton ($\mathcal{A}_{\mathsf{pos}}$) correspond to the positions of letters in the corresponding regular expression plus an additional initial state. McNaughton and Yamada [15] also used the positions of a regular expression to define an automaton, however they directly computed a deterministic version of the position automaton. The position automaton has been well studied [8,3] and is considered the *standard* automaton simulation of a regular expression [16]. Some of its interesting properties are: homogeneity, i.e. for each state, all in-transitions have the same label (letter); whenever deterministic, these automata characterize certain families of unambiguous regular expressions, and can be computed in quadratic time [4]; other automata simulations of regular expressions are quotients of the $\mathcal{A}_{\mathsf{pos}}$, e.g. partial derivative automata ($\mathcal{A}_{\mathsf{pd}}$) [9] and follow automata [14].

Many authors observed that the position automaton construction could not directly be extended to regular expressions with intersection [3,6], as intersection (and also complementation) is not compatible with the notion of position. In fact, considering positions of letters in the expression $(ab^\star) \cap a$, whose language is $\{a\}$, we obtain the regular expression $(a_1 b_2^\star) \cap a_3$. Interpreting a_1 and a_3 as distinct alphabet symbols, the language described by this expression is empty and there is

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

no longer a correspondence between the languages of $(ab^\star)\cap a$ and $(a_1b_2^\star)\cap a_3$, as it is the case for expressions without intersection. However, the conversions from expressions to automata based on the notion of derivative or partial derivative can still be extended to regular expressions with intersection [5,7,2]. In this paper, we present a position automaton construction for regular expressions with intersection by generalizing the notion of position. Instead of positions, sets of positions are considered in such a way that marking a regular expression is made compatible with the intersection operation. We also show that the partial derivative automaton is a quotient of the position automaton.

2 Preliminaries

In this section we recall the basic definitions we use throughout this paper and the notation. For further details we refer to [13,17].

Let Σ be an *alphabet* (set of letters). A *word* over Σ is a finite sequence of letters, where ε is the empty word. The size of a word x, $|x|$, is the number of alphabet symbols in x. $\Sigma^\star$ denotes the set of all words over Σ, and a *language* over Σ is any subset of $\Sigma^\star$. The *concatenation* of two languages L_1 and L_2 is defined by $L_1 \cdot L_2 = \{\ xy \mid x \in L_1, y \in L_2\ \}$, and $L^\star$ denotes the set $\{\ x_1x_2\cdots x_n \mid n \geq 0,\ x_i \in L\ \}$. The *left quotient* of a language $L \subseteq \Sigma^\star$ w.r.t. a word $x \in \Sigma^\star$ is the language $x^{-1}L = \{\ y \mid xy \in L\ \}$.

The set $\mathsf{RE}_\cap$ of *regular expressions with intersection* over Σ is defined by the following grammar

$$\alpha, \beta := \emptyset \mid \varepsilon \mid a \in \Sigma \mid (\alpha + \beta) \mid (\alpha \cdot \beta) \mid (\alpha^\star) \mid (\alpha \cap \beta), \tag{1}$$

where the concatenation operator $\cdot$ is often omitted. We consider $\mathsf{RE}_\cap$ expressions modulo the standard equations for $\emptyset$ and ε, i.e. $\alpha+\emptyset = \emptyset+\alpha = \alpha\cdot\varepsilon = \varepsilon\cdot\alpha = \alpha$, $\alpha\cdot\emptyset = \emptyset\cdot\alpha = \alpha\cap\emptyset = \emptyset\cap\alpha = \emptyset$, and $\emptyset^\star = \varepsilon$. Throughout this paper we often refer to regular expressions with intersection just as regular expressions. The set of alphabet symbols with occurrences in α is denoted by Σ_α. Expressions containing no occurrence of the operator $\cap$ are called *simple regular expressions.* A *linear regular expression* is a regular expression in which every alphabet symbol occurs at most once. We let $|\alpha|$, $|\alpha|_\Sigma$ and $|\alpha|_\cap$ denote for $\alpha \in \mathsf{RE}_\cap$ the number of symbols, the number of occurrences of alphabet symbols and the number of occurrences of the binary operator $\cap$, respectively. The language $\mathcal{L}(\alpha)$ for $\alpha \in \mathsf{RE}_\cap$ is defined as usual, with $\mathcal{L}(\alpha\cap\beta) = \mathcal{L}(\alpha)\cap\mathcal{L}(\beta)$. The language of $S \subseteq \mathsf{RE}_\cap$ is $\mathcal{L}(S) = \cup_{\alpha\in S}\mathcal{L}(\alpha)$. Given an expression $\alpha \in \mathsf{RE}_\cap$, we define $\varepsilon(\alpha) = \varepsilon$ if $\varepsilon \in \mathcal{L}(\alpha)$, and $\varepsilon(\alpha) = \emptyset$ otherwise. A recursive definition of $\varepsilon : \mathsf{RE}_\cap \longrightarrow \{\emptyset, \varepsilon\}$ is given by the following: $\varepsilon(a) = \varepsilon(\emptyset) = \emptyset$, $\varepsilon(\varepsilon) = \varepsilon(\alpha^\star) = \varepsilon$, $\varepsilon(\alpha+\beta) = \varepsilon(\alpha) + \varepsilon(\beta)$, and $\varepsilon(\alpha\beta) = \varepsilon(\alpha\cap\beta) = \varepsilon(\alpha)\cdot\varepsilon(\beta)$.

A *nondeterministic finite automaton* (NFA) is a tuple $\mathcal{A} = \langle S, \Sigma, S_0, \delta, F\rangle$, where S is a finite set of states, Σ is a finite alphabet, $S_0 \subseteq S$ a set of initial states, $\delta : S \times \Sigma \longrightarrow \mathcal{P}(S)$ the transition function, and $F \subseteq S$ a set of final states. The extension of δ to sets of states and words is defined by $\delta(X, \varepsilon) = X$ and $\delta(X, ax) = \delta(\cup_{s\in X}\delta(s, a), x)$. A word $x \in \Sigma^\star$ is accepted by $\mathcal{A}$ if and only if $\delta(S_0, x) \cap F \neq \emptyset$. The *language of* $\mathcal{A}$, $\mathcal{L}(\mathcal{A})$, is the set of words accepted by $\mathcal{A}$. The *right language of a state* s, $\mathcal{L}_s$, is the language accepted by $\mathcal{A}$ if we take $S_0 = \{s\}$. An NFA is *initially connected* or *accessible* if each state is reachable from an initial state and it is *trimmed* if, moreover, the right language of each state is non-empty. Given $\mathcal{A}$, we denote by $\mathcal{A}^{\mathsf{ac}}$ and $\mathcal{A}^{\mathsf{t}}$ the result of removing unreachable states from $\mathcal{A}$ and trimming $\mathcal{A}$, respectively. It is clear that $\mathcal{L}(\mathcal{A}) = \mathcal{L}(\mathcal{A}^{\mathsf{ac}}) = \mathcal{L}(\mathcal{A}^{\mathsf{t}})$.

We say that an equivalence relation $\equiv$ over S is right invariant w.r.t. $\mathcal{A}$ iff

1. $\forall s, t \in S,\ s \equiv t \wedge s \in F \implies t \in F$
2. $\forall s, t \in S, \forall a \in \Sigma,\ s \equiv t \implies \forall s_1 \in \delta(s, a)\ \exists t_1 \in \delta(t, a), s_1 \equiv t_1.$

If $\equiv$ is right invariant, then we can define the quotient automaton $\mathcal{A}/_{\equiv}$ in the usual way, and $\mathcal{L}(\mathcal{A}/_{\equiv}) = \mathcal{L}(\mathcal{A})$.

The notions of partial derivatives and partial derivative automata were introduced by Antimirov [1] for simple regular expressions. Bastos et al. [2] presented an extension of the Antimirov construction from $\mathsf{RE}_{\cap}$ expressions.

Definition 1. *For* $\alpha \in \mathsf{RE}_{\cap}$ *and* $a \in \Sigma$*, the set* $\partial_a(\alpha)$ *of* partial derivatives *of* α *w.r.t.* a *is defined by:*

$$\partial_a(\emptyset) = \partial_a(\varepsilon) = \emptyset \qquad \partial_a(\alpha + \beta) = \partial_a(\alpha) \cup \partial_a(\beta)$$

$$\partial_a(b) = \begin{cases} \{\varepsilon\}, & \textit{if } a = b \\ \emptyset & \textit{otherwise} \end{cases} \qquad \partial_a(\alpha\beta) = \begin{cases} \partial_a(\alpha) \odot \beta \cup \partial_a(\beta), & \textit{if } \varepsilon(\alpha) = \varepsilon \\ \partial_a(\alpha)\beta, & \textit{otherwise} \end{cases}$$

$$\partial_a(\alpha^\star) = \partial_a(\alpha) \odot \alpha^\star \qquad \partial_a(\alpha \cap \beta) = \partial_a(\alpha) \Cap \partial_a(\beta),$$

where for $S, T \subseteq \mathsf{RE}_{\cap}$ *and* $\beta \in \mathsf{RE}_{\cap}$, $S \odot \beta = \{\ \alpha\beta \mid \alpha \in S\ \}$, $\beta \odot S = \{\ \beta\alpha \mid \alpha \in S\ \}$*, and* $S \Cap T = \{\ \alpha \cap \beta \mid \alpha \in S, \beta \in T\ \}$.

This definition is extended to any word w by $\partial_\varepsilon(\alpha) = \{\alpha\}$, $\partial_{wa}(\alpha) = \bigcup_{\alpha_i \in \partial_w(\alpha)} \partial_a(\alpha_i)$, and $\partial_w(R) = \bigcup_{\alpha_i \in R} \partial_w(\alpha_i)$, where $R \subseteq \mathsf{RE}_{\cap}$. The set of partial derivatives of an expression α is $\partial(\alpha) = \bigcup_{w \in \Sigma^\star} \partial_w(\alpha)$. As for simple regular expressions, the partial derivative automaton of an expression $\alpha \in \mathsf{RE}_{\cap}$ is defined by $\mathcal{A}_{\mathsf{pd}}(\alpha) = \langle \partial(\alpha), \Sigma, \{\alpha\}, \delta_{\mathsf{pd}}, F_{\mathsf{pd}} \rangle$, where $F_{\mathsf{pd}} = \{\ \gamma \in \partial(\alpha) \mid \varepsilon(\gamma) = \varepsilon\ \}$ and $\delta_{\mathsf{pd}}(\gamma, a) = \partial_a(\gamma)$. It follows that $\mathcal{L}(\mathcal{A}_{\mathsf{pd}}(\alpha))$ is exactly $\mathcal{L}(\alpha)$ and by construction $\mathcal{A}_{\mathsf{pd}}(\alpha)$ is accessible. Bastos et al. [2] showed also that $|\partial(\alpha)| \leq 2^{|\alpha|_\Sigma - |\alpha|_\cap - 1} + 1$ and asymptotically and on average an upper bound for the number of states is $(1.056 + o(1))^n$, where n is the size of the expression.

3 Indexed Expressions

Given an alphabet Σ and a nonempty set of indexes $J \subseteq \mathbb{N}$, let $\Sigma_J = \{\ a_j \mid a \in \Sigma, j \in J\ \}$. An *indexed regular expression* is a regular expression over the alphabet Σ_J such that for all $a_i, b_j \in \Sigma_J$ occurring in the expression, $a \neq b$ implies $i \neq j$. We let $\rho, \rho_1, \rho_2, \ldots$ denote indexed regular expressions. If ρ is an indexed expression, then $\overline{\rho}$ is the regular expression over the alphabet Σ obtained by removing the indexes. The set of all indexes occurring in ρ is denoted by

$\mathsf{ind}(\rho) = \{\ i \mid a_i \in \Sigma_\rho\ \}$. Given an indexed expression ρ and $i \in \mathsf{ind}(\rho)$, $\ell_\rho(i)$ is the letter indexed by i in ρ. From now on, we will simply write $\ell(i)$ for $\ell_\rho(i)$ since it will always be clear that we are referring to a specific expression ρ. Given an indexed expression ρ, let

$$\mathcal{I}_\rho = \{\ \mathrm{I} \subseteq \mathsf{ind}(\rho) \mid \mathrm{I} \neq \emptyset \text{ and } \forall i_1, i_2 \in \mathrm{I}, \ell(i_1) = \ell(i_2)\ \}.$$

For $\mathrm{I} \in \mathcal{I}_\rho$ we extend the definition of ℓ by $\ell(\mathrm{I}) = \ell(i)$, $i \in \mathrm{I}$. Finally, we say that ρ is *well-indexed* if for all subterms of ρ of the form $\rho_1 \cap \rho_2$ one has $\mathsf{ind}(\rho_1) \cap \mathsf{ind}(\rho_2) = \emptyset$.

Example 2. For $\rho = a_1(a_4 b_5^\star \cap a_4)$ one has $\overline{\rho} = a(ab^\star \cap a)$, $\mathsf{ind}(\rho) = \{1, 4, 5\}$, $\ell(4) = \ell(\{1, 4\}) = a$ and $\mathcal{I}_\rho = \{\{1\}, \{4\}, \{5\}, \{1, 4\}\}$. However, this expression is not well-indexed, since a_4 occurs on both sides of an intersection.

Definition 3. *Consider an indexed expression ρ. For $L \subseteq \mathcal{I}_\rho^\star$ and $x = \mathrm{I}_1 \cdots \mathrm{I}_n \in L$, we define $\ell(x) = \ell(\mathrm{I}_1) \cdots \ell(\mathrm{I}_n)$ and $\ell(L) = \{\ \ell(x) \mid x \in L\ \}$. The* indexed intersection *of two words $x = \mathrm{I}_1 \cdots \mathrm{I}_m, y = \mathrm{J}_1 \cdots \mathrm{J}_n \in \mathcal{I}_\rho^\star$ is defined by $x \cap_\mathcal{I} y = (\mathrm{I}_1 \cup \mathrm{J}_1) \cdots (\mathrm{I}_n \cup \mathrm{J}_n)$ if $\ell(x) = \ell(y)$[1], and undefined otherwise. Then, the* indexed intersection *of two languages $L_1, L_2 \in \mathcal{I}_\rho^\star$ is defined as follows:*

$$L_1 \cap_\mathcal{I} L_2 = \{\ x \cap_\mathcal{I} y \mid x \in L_1, y \in L_2\ \}.$$

We define the index-language *$\mathcal{L}_\mathcal{I}(\rho) \subseteq \mathcal{I}_\rho^\star$ associated to ρ as follows.*

$$\begin{array}{lll} \mathcal{L}_\mathcal{I}(\emptyset) = \emptyset, & \mathcal{L}_\mathcal{I}(a_i) = \{\{i\}\}, & \mathcal{L}_\mathcal{I}(\rho_1 + \rho_2) = \mathcal{L}_\mathcal{I}(\rho_1) \cup \mathcal{L}_\mathcal{I}(\rho_2), \\ \mathcal{L}_\mathcal{I}(\varepsilon) = \{\varepsilon\}, & \mathcal{L}_\mathcal{I}(\rho^\star) = \mathcal{L}_\mathcal{I}(\rho)^\star, & \mathcal{L}_\mathcal{I}(\rho_1 \cdot \rho_2) = \mathcal{L}_\mathcal{I}(\rho_1) \cdot \mathcal{L}_\mathcal{I}(\rho_2), \\ & & \mathcal{L}_\mathcal{I}(\rho_1 \cap \rho_2) = \mathcal{L}_\mathcal{I}(\rho_1) \cap_\mathcal{I} \mathcal{L}_\mathcal{I}(\rho_2). \end{array}$$

Example 4. For $\rho = (a_1 a_2 + b_3 + a_4)^\star \cap (a_5 + b_6)^\star$, we have $\mathcal{L}_\mathcal{I}(\rho) = \{\{4, 5\}, \{3, 6\}, \{1, 5\}\{2, 5\}, \{4, 5\}\{4, 5\}, \{4, 5\}\{3, 6\}, \ldots\}$, and $\ell(\mathcal{L}_\mathcal{I}(\rho)) = \{a, b, aa, ab, \ldots\}$ (since $\ell(\{1, 5\}\{2, 5\}) = \ell(\{4, 5\}\{4, 5\}) = aa$).

Proposition 5. *Given an indexed expression ρ, one has $\ell(\mathcal{L}_\mathcal{I}(\rho)) = \mathcal{L}(\overline{\rho})$.*

4 A Position Automaton for $\mathsf{RE}_\cap$ Expressions

Let $\alpha \in \mathsf{RE}_\cap$. We define the set of positions in α by $\mathsf{pos}(\alpha) = \{1, \ldots, |\alpha|_\Sigma\}$. As usual, we let $\overline{\alpha}$ denote the expression obtained from α by indexing each letter with its position in α. The same notation is used to remove the indexes, as already stated, thus, $\overline{\overline{\alpha}} = \alpha$. Note that for $\alpha \in \mathsf{RE}_\cap$, the indexed expression $\overline{\alpha}$ is always linear (thus well-indexed), and also $\mathsf{pos}(\alpha) = \mathsf{ind}(\overline{\alpha})$.

Given an indexed linear expression ρ we define the following sets:

$$\begin{aligned} \mathsf{First}'(\rho) &= \{\ \mathrm{I} \mid \mathrm{I}x \in \mathcal{L}_\mathcal{I}(\rho)\ \}, \\ \mathsf{Last}'(\rho) &= \{\ \mathrm{I} \mid x\mathrm{I} \in \mathcal{L}_\mathcal{I}(\rho)\ \}, \\ \mathsf{Follow}'(\rho) &= \{\ (\mathrm{I}, \mathrm{J}) \mid x\mathrm{I}\mathrm{J}y \in \mathcal{L}_\mathcal{I}(\rho)\ \}. \end{aligned}$$

[1] Note that $\ell(x) = \ell(y)$ implies that $m = n$ and that $\ell(x \cap_\mathcal{I} y) = \ell(x) = \ell(y)$.

Then, given $\alpha \in \mathsf{RE}_\cap$, we define $\mathsf{First}(\alpha) = \mathsf{First}'(\overline{\alpha}), \mathsf{Last}(\alpha) = \mathsf{Last}'(\overline{\alpha})$, and $\mathsf{Follow}(\alpha) = \mathsf{Follow}'(\overline{\alpha})$.

Definition 6. *The* position automaton *of an expression $\alpha \in \mathsf{RE}_\cap$ is*

$$\mathcal{A}_{\mathsf{pos}}(\alpha) = \langle S_{\mathsf{pos}}, \Sigma, \{\{0\}\}, \delta_{\mathsf{pos}}, F_{\mathsf{pos}} \rangle,$$

$$\begin{aligned} \textit{where } S_{\mathsf{pos}} &= \{\{0\}\} \cup \{\mathrm{I} \in \mathcal{I}_{\overline{\alpha}} \mid x\mathrm{I}y \in \mathcal{L}_{\mathcal{I}}(\overline{\alpha}) \textit{ for some } x, y \in \mathcal{I}^\star_{\overline{\alpha}}\ \}, \\ \delta_{\mathsf{pos}} &= \{\ (\mathrm{I}, \ell(\mathrm{J}), \mathrm{J}) \mid (\mathrm{I}, \mathrm{J}) \in \mathsf{Follow}(\alpha)\ \} \cup \{\ (\{0\}, \ell(\mathrm{I}), \mathrm{I}) \mid \mathrm{I} \in \mathsf{First}(\alpha)\ \}, \\ F_{\mathsf{pos}} &= \begin{cases} \mathsf{Last}(\alpha) \cup \{\{0\}\}, & \textit{if } \varepsilon(\alpha) = \varepsilon; \\ F_{\mathsf{pos}} = \mathsf{Last}(\alpha), & \textit{otherwise.} \end{cases} \end{aligned}$$

Proposition 7. *Given an expression $\alpha \in \mathsf{RE}_\cap$, one has $\mathcal{L}(\mathcal{A}_{\mathsf{pos}}(\alpha)) = \mathcal{L}(\alpha)$.*

Note that for regular expressions without intersection (simple regular expressions) the automaton is, by the definition of $\mathcal{L}_{\mathcal{I}}$, isomorphic to the classic position automaton, with the difference that now states are labelled with singletons $\{i\}$ instead of $i \in \mathsf{pos}(\alpha) \cup \{0\}$. We now give definitions for recursively computing sets corresponding to First, Last and Follow. These definitions lead to supersets of the corresponding sets but we will proof that extra elements can be discarded and if we trim the resulting NFA we obtain $\mathcal{A}_{\mathsf{pos}}$.

Definition 8. *Given a indexed well-indexed expression ρ, let $\mathsf{Fst}(\rho) \subseteq \mathcal{I}_\rho$ be inductively defined as follows,*

$$\begin{aligned} \mathsf{Fst}(\emptyset) = \mathsf{Fst}(\varepsilon) &= \emptyset \\ \mathsf{Fst}(a_i) &= \{\{i\}\} \\ \mathsf{Fst}(\rho^\star) &= \mathsf{Fst}(\rho) \end{aligned} \qquad \begin{aligned} \mathsf{Fst}(\rho_1 + \rho_2) &= \mathsf{Fst}(\rho_1) \cup \mathsf{Fst}(\rho_2) \\ \mathsf{Fst}(\rho_1 \cdot \rho_2) &= \begin{cases} \mathsf{Fst}(\rho_1) \cup \mathsf{Fst}(\rho_2), & \textit{if } \varepsilon(\rho_1) = \varepsilon \\ \mathsf{Fst}(\rho_1), & \textit{otherwise} \end{cases} \\ \mathsf{Fst}(\rho_1 \cap \rho_2) &= \mathsf{Fst}(\rho_1) \otimes \mathsf{Fst}(\rho_2). \end{aligned}$$

where for $F_1, F_2 \subseteq \mathcal{I}_\rho$, $F_1 \otimes F_2 = \{\ \mathrm{I}_1 \cup \mathrm{I}_2 \mid \ell(\mathrm{I}_1) = \ell(\mathrm{I}_2)$ and $\mathrm{I}_1 \in F_1$, $\mathrm{I}_2 \in F_2\ \}$.

By construction, all elements $\mathrm{I} \in \mathsf{Fst}(\rho)$ are non-empty and such that $\ell(i_1) = \ell(i_2)$ for all $i_1, i_2 \in \mathrm{I}$, guaranting that $\otimes$ is well defined and $\mathsf{Fst}(\rho) \subseteq \mathcal{I}_\rho$.

Example 9. We have $\mathsf{Fst}(a_1^\star b_2^\star \cap a_3) = \mathsf{Fst}(a_1^\star b_2^\star) \otimes \mathsf{Fst}(a_3) = \{\{1\},\{2\}\} \otimes \{\{3\}\} = \{\{1,3\}\}$.

Definition 10. *Given a well-indexed expression ρ, the set $\mathsf{Lst}(\rho) \subseteq \mathcal{I}_\rho$ is defined as $\mathsf{Fst}(\rho)$, with the difference that for concatenation we have:*

$$\mathsf{Lst}(\rho_1 \cdot \rho_2) = \begin{cases} \mathsf{Lst}(\rho_1) \cup \mathsf{Lst}(\rho_2), & \textit{if } \varepsilon(\rho_2) = \varepsilon \\ \mathsf{Lst}(\rho_2), & \textit{otherwise.} \end{cases}$$

The set $\mathsf{Fol}(\rho) \subseteq \mathcal{I}_\rho \times \mathcal{I}_\rho$ *is inductively defined as follows,*

$$\begin{array}{ll} \mathsf{Fol}(\emptyset) \;\;= \mathsf{Fol}(\varepsilon) = \mathsf{Fol}(a_i) = \emptyset & \mathsf{Fol}(\rho_1 + \rho_2) = \mathsf{Fol}(\rho_1) \cup \mathsf{Fol}(\rho_2) \\ \mathsf{Fol}(\rho^\star) = \mathsf{Fol}(\rho) \cup \mathsf{Lst}(\rho) \times \mathsf{Fst}(\rho) & \mathsf{Fol}(\rho_1 \cap \rho_2) = \mathsf{Fol}(\rho_1) \otimes \mathsf{Fol}(\rho_2) \end{array}$$
$$\mathsf{Fol}(\rho_1 \cdot \rho_2) = \mathsf{Fol}(\rho_1) \cup \mathsf{Fol}(\rho_2) \cup \mathsf{Lst}(\rho_1) \times \mathsf{Fst}(\rho_2).$$

where, for $S_1, S_2 \subseteq \mathcal{I}_\rho \times \mathcal{I}_\rho$,

$$S_1 \otimes S_2 = \{\ (\mathrm{I}_1 \cup \mathrm{I}_2, \mathrm{J}_1 \cup \mathrm{J}_2) \mid (\mathrm{I}_1, \mathrm{J}_1) \in S_1, (\mathrm{I}_2, \mathrm{J}_2) \in S_2 \text{ and } \ell(\mathrm{I}_1) = \ell(\mathrm{I}_2), \ell(\mathrm{J}_1) = \ell(\mathrm{J}_2)\ \}.$$

In the next definition we will use the projection functions on the first and second coordinates, π_1 and π_2, respectively.

Definition 11. *Given* $\alpha \in \mathsf{RE}_\cap$, *let* $\mathcal{A}_{\mathsf{posi}}(\alpha) = \langle S_{\mathsf{posi}}, \Sigma, \{\{0\}\}, \delta_{\mathsf{posi}}, F_{\mathsf{posi}} \rangle$ *be the NFA where* $S_{\mathsf{posi}} = \{\{0\}\} \cup \mathsf{Fst}(\overline{\alpha}) \cup \mathsf{Lst}(\overline{\alpha}) \cup \pi_1(\mathsf{Fol}(\overline{\alpha})) \cup \pi_2(\mathsf{Fol}(\overline{\alpha}))$, *and* δ_{posi} *and* F_{posi} *are defined as* δ_{pos} *and* F_{pos}, *substituting the functions* First, Last *and* Follow, *by* Fst, Lst *and* Fol, *respectively.*

We will now show that $\mathcal{L}(\mathcal{A}_{\mathsf{pos}}(\alpha)) = \mathcal{L}(\mathcal{A}_{\mathsf{posi}}(\alpha))$, and that $\mathcal{A}_{\mathsf{pos}}(\alpha)$ is obtained by trimming $\mathcal{A}_{\mathsf{posi}}(\alpha)$, as the result of the two following lemmas. An example is presented at the end of this section.

Lemma 12. *Given an indexed linear expression* ρ, *one has:* 1) $\mathsf{First}'(\rho) \subseteq \mathsf{Fst}(\rho)$; 2) $\mathsf{Last}'(\rho) \subseteq \mathsf{Lst}(\rho)$; 3) $\mathsf{Follow}'(\rho) \subseteq \mathsf{Fol}(\rho)$.

Example 13. For $\rho = (a_1 \cap b_2)c_3d_4$, we have $(\{3\}, \{4\}) \in \mathsf{Fol}(\rho)$, but $(\{3\}, \{4\}) \notin \mathsf{Follow}(\rho)$. Thus, $\mathsf{Fol}(\rho) \not\subseteq \mathsf{Follow}(\rho)$.

The previous Lemma shows that for any $\alpha \in \mathsf{RE}_\cap$, $\mathcal{A}_{\mathsf{pos}}(\alpha)$ is a subautomaton of $\mathcal{A}_{\mathsf{posi}}(\alpha)$, and thus $\mathcal{L}(\mathcal{A}_{\mathsf{pos}}(\alpha)) \subseteq \mathcal{L}(\mathcal{A}_{\mathsf{posi}}(\alpha))$. We now show that both recognize the same language and can be made isomorphic by trimming $\mathcal{A}_{\mathsf{posi}}$.

Lemma 14. *Given an indexed linear expression* ρ *and some* $n \geq 1$, *if* $\mathrm{I}_n \in \mathsf{Lst}(\rho)$ *and there exist* $\mathrm{I}_1, \ldots, \mathrm{I}_n \in \mathcal{I}_\rho$ *such that*

$$(\{0\}, \ell(\mathrm{I}_1), \mathrm{I}_1), (\mathrm{I}_1, \ell(\mathrm{I}_2), \mathrm{I}_2), \ldots, (\mathrm{I}_{n-1}, \ell(\mathrm{I}_n), \mathrm{I}_n) \in \delta_{\mathsf{posi}},$$

then $\mathrm{I}_1 \cdots \mathrm{I}_n \in \mathcal{L}_{\mathcal{I}}(\rho)$.

Theorem 15. *For any* $\alpha \in \mathsf{RE}_\cap$, $\mathcal{L}(\mathcal{A}_{\mathsf{pos}}(\alpha)) = \mathcal{L}(\mathcal{A}_{\mathsf{posi}}(\alpha))$.

From these results, it follows that if we trim the automaton $\mathcal{A}_{\mathsf{posi}}$ we obtain exactly $\mathcal{A}_{\mathsf{pos}}$.

Example 16. Consider $\alpha = (ba^\star b + a) \cap (aa + b)^\star$. Then $\overline{\alpha} = (b_1a_2^\star b_3 + a_4) \cap (a_5a_6 + b_7)^\star$, $\mathsf{Fst}(\overline{\alpha}) = \{\{1,7\},\{4,5\}\}$, $\mathsf{Lst}(\overline{\alpha}) = \{\{3,7\},\{4,6\}\}$, and $\mathsf{Fol}(\overline{\alpha}) = \{(\{2,5\},\{2,6\}),(\{2,6\},\{2,5\}),(\{2,6\},\{3,7\}),(\{1,7\},\{2,5\}),(\{1,7\},\{3,7\})\}$.

The automaton $\mathcal{A}_{\mathsf{posi}}(\alpha)$ is represented in Figure 1. The trimmed automaton, $\mathcal{A}_{\mathsf{posi}}(\alpha)^{\mathsf{t}}$, is obtained removing the states labeled by $\{4,5\}$ and $\{4,6\}$, and the correspondent transitions.

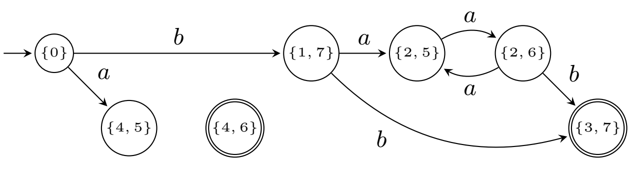

Fig. 1. $\mathcal{A}_{\mathsf{posi}}((ba^{\star}b + a) \cap (aa + b)^{\star})$

5 A c-Continuation Automaton for $\mathsf{RE}_{\cap}$ Expressions

In the case of simple regular expressions, Champarnaud and Ziadi [9] defined a nondeterministic automaton isomorphic to the position automaton, called the c-continuation automaton, in order to show that the partial derivative automaton can be seen as a quotient of the position automaton. With the same purpose, in this section, we present a c-continuation automaton for expressions with intersection. Moreover, instead of considering derivatives of regular expressions [5], we use partial derivatives to restate some known results for simple regular expressions.

The notion of continuation was defined by Berry and Sethy [3], and developed by Champarnaud and Ziadi [9], by Ilie and Yu [14], and by Chen and Yu [10]. Given $a \in \Sigma$ and a linear simple expression α, the set of partial derivatives $\partial_{xa}(\alpha)$, for any word $x \in \Sigma^{\star}$, is either $\emptyset$ or has a unique element γ called the *continuation* of a in α. Note that using partial derivatives, *continuations* and non-null c-*continuations* coincide. Furthermore, the continuation can be obtained by some refinement of the inductive definition of partial derivatives, exploring the linearity of α. In order to establish similar results for linear well-indexed expressions, we introduce the notion of *partial index-derivative* of a well-indexed expression ρ w.r.t. an index $\mathrm{I} \in \mathcal{I}_{\rho}$.

Given a well-indexed expression ρ, a subexpression τ of ρ, and a set of indexes $\mathrm{I} \in \mathcal{I}_{\rho}$, let $\mathrm{I}\big|_{\tau}$ denote the set of indexes in I that occur in τ. This definition is naturally extended to words $x = \mathrm{I}_1 \cdots \mathrm{I}_n \in \mathcal{I}_{\rho}^{\star}$ by $x\big|_{\tau} = \mathrm{I}_1\big|_{\tau} \cdots \mathrm{I}_n\big|_{\tau}$, for $n \geq 0$.

Definition 17. *The set of partial index-derivatives of a well-indexed expression ρ by $\mathrm{I} \in \mathcal{I}_{\rho} \cup \{\emptyset\}$, $\partial_{\mathrm{I}}(\rho)$, is defined by*

$$\begin{aligned} \partial_{\mathrm{I}}(\emptyset) &= \partial_{\mathrm{I}}(\varepsilon) = \emptyset & \partial_{\mathrm{I}}(\rho^{\star}) &= \partial_{\mathrm{I}}(\rho) \odot \rho^{\star} \\ \partial_{\mathrm{I}}(a_i) &= \begin{cases} \{\varepsilon\}, & \textit{if } \mathrm{I} = \{i\} \\ \emptyset, & \textit{otherwise} \end{cases} & \partial_{\mathrm{I}}(\rho_1 + \rho_2) &= \partial_{\mathrm{I}}(\rho_1) \cup \partial_{\mathrm{I}}(\rho_2) \end{aligned}$$

$$\begin{aligned} \partial_{\mathrm{I}}(\rho_1 \cdot \rho_2) &= \begin{cases} \partial_{\mathrm{I}}(\rho_1) \odot \rho_2 \cup \partial_{\mathrm{I}}(\rho_2), & \textit{if } \varepsilon(\rho_1) = \varepsilon \\ \partial_{\mathrm{I}}(\rho_1) \odot \rho_2, & \textit{otherwise} \end{cases} \\ \partial_{\mathrm{I}}(\rho_1 \cap \rho_2) &= \begin{cases} \partial_{\mathrm{I}|_{\rho_1}}(\rho_1) \mathbin{\dot{\cap}} \partial_{\mathrm{I}|_{\rho_2}}(\rho_2), & \textit{if } \mathrm{I} = \mathrm{I}\big|_{\rho_1} \cup \mathrm{I}\big|_{\rho_2} \\ \emptyset, & \textit{otherwise.} \end{cases} \end{aligned}$$

The set of partial index-derivatives of ρ by a word $x \in \mathcal{I}_\rho^\star$ is then inductively defined by $\partial_\varepsilon(\rho) = \{\rho\}$ and $\partial_{x\mathrm{I}}(\rho) = \bigcup_{\rho' \in \partial_x(\rho)} \partial_\mathrm{I}(\rho')$. If S is a set of well-indexed expressions, $\partial_x(S) = \bigcup_{\rho \in S} \partial_x(\rho)$.

It is straightforward to see that $\partial_\emptyset(\rho) = \emptyset$ for all ρ. Although $\emptyset \notin \mathcal{I}_\rho$, the notion of partial index-derivative includes the derivative by an empty set of indexes, in order to guarantee that the derivative of an intersection is well-defined. Also note that the partial index-derivative of a well-indexed expression is still well-indexed. Finally, the set of partial index-derivatives of ρ by all $\mathrm{I} \in \mathcal{I}_\rho$ can be calculated simultaneously using an extension of the linear form defined by Antimirov [1], i.e. considering pairs (I, ρ') where $\rho' \in \partial_\mathrm{I}(\rho)$. The following lemma is proved by induction on n.

Lemma 18. *If $x = \mathrm{I}_1 \cdots \mathrm{I}_n$ and $\partial_x(\rho) \neq \emptyset$, then $x = x\big|_\rho$.*

Example 19. We have $\partial_{\{1,3\}}(a_1^\star b_2^\star \cap a_3) = \partial_{\{1\}}(a_1^\star b_2^\star) \mathbin{\dot{\cap}} \partial_{\{3\}}(a_3) = \{a_1^\star b_2^\star \cap \varepsilon\}$.

Proposition 20. *Consider a well-indexed expression ρ and $\mathrm{I} \in \mathcal{I}_\rho$. Then,*

$$\mathrm{I}^{-1}\mathcal{L}_\mathcal{I}(\rho) = \mathcal{L}_\mathcal{I}(\partial_\mathrm{I}(\rho)) \qquad \textit{and} \qquad \mathcal{L}_\mathcal{I}(\rho) = \mathcal{L}_\mathcal{I}\left(\textstyle\bigcup_{\mathrm{I} \in \mathcal{I}_\rho} (\mathrm{I} \odot \partial_\mathrm{I}(\rho)) \cup \varepsilon(\rho)\right).$$

Corollary 21. *For every well-indexed expression $\rho \in \mathsf{RE}_\cap$ and word $x \in \mathcal{I}_\rho^\star$, one has $x^{-1}\mathcal{L}_\mathcal{I}(\rho) = \mathcal{L}_\mathcal{I}(\partial_x(\rho))$ and $\mathcal{L}_\mathcal{I}(\rho) = \mathcal{L}_\mathcal{I}(\bigcup_{x \in \mathcal{I}_\rho^\star}(x \odot \partial_x(\rho)) \cup \varepsilon(\rho))$.*

The following is an adaptation, for partial index-derivatives and intersection, of a result due to Berry and Sethi [3].

Proposition 22. *Consider a linear indexed expression $\rho \in \mathsf{RE}_\cap$ and $x\mathrm{I} \in \mathcal{I}_\rho^\star$, and let $\mathsf{suff}(x)$ denote the set of all suffixes of x. The partial index-derivative $\partial_{x\mathrm{I}}(\rho)$ of ρ satisfies:*

$$\begin{aligned}
\partial_{x\mathrm{I}}(\emptyset) = \partial_{x\mathrm{I}}(\varepsilon) = \emptyset,& \\
\partial_{x\mathrm{I}}(a_i) &= \begin{cases} \{\varepsilon\}, & \textit{if } x\mathrm{I} = \{i\}, \\ \emptyset, & \textit{otherwise,} \end{cases} \\
\partial_{x\mathrm{I}}(\rho_1 + \rho_2) &= \begin{cases} \partial_{x\mathrm{I}}(\rho_1), & \textit{if } x\mathrm{I} = (x\mathrm{I})\big|_{\rho_1}, \\ \partial_{x\mathrm{I}}(\rho_2), & \textit{if } x\mathrm{I} = (x\mathrm{I})\big|_{\rho_2}, \\ \emptyset & \textit{otherwise} \end{cases} \\
\partial_{x\mathrm{I}}(\rho_1 \cdot \rho_2) &= \begin{cases} \partial_{x\mathrm{I}}(\rho_1) \odot \rho_2, & \textit{if } x\mathrm{I} = (x\mathrm{I})\big|_{\rho_1}, \\ \partial_{z\mathrm{I}}(\rho_2), & \textit{if } x = yz, \varepsilon(\partial_y(\rho_1)) = \varepsilon, z\mathrm{I} = (z\mathrm{I})\big|_{\rho_2}, \\ \emptyset, & \textit{otherwise,} \end{cases} \\
\partial_{x\mathrm{I}}(\rho^\star) &\subseteq \bigcup_{v \in \mathsf{suff}(x)} \partial_{v\mathrm{I}}(\rho) \odot \rho^\star, \\
\partial_{x\mathrm{I}}(\rho_1 \cap \rho_2) &= \begin{cases} \partial_{(x\mathrm{I})|_{\rho_1}}(\rho_1) \mathbin{\dot{\cap}} \partial_{(x\mathrm{I})|_{\rho_2}}(\rho_2), & \textit{if } x\mathrm{I} = (x\mathrm{I})\big|_{\rho_1} \cap_\mathcal{I} (x\mathrm{I})\big|_{\rho_2}, \\ \emptyset, & \textit{otherwise.} \end{cases}
\end{aligned}$$

The previous proposition implies that if $\partial_{x\mathrm{I}}(\rho) \neq \emptyset$, then it has only one element for every $x \in \mathcal{I}_\rho^\star$. This fact is proved in Proposition 24 and the unique element (if exists) is defined below.

Definition 23. *Given a linear indexed expression ρ and a set of indexes I, the c-continuation $\mathsf{c}_{\mathrm{I}}(\rho)$ of ρ w.r.t. I is defined by the following rules.*

$$\mathsf{c}_{\mathrm{I}}(\emptyset) = \mathsf{c}_{\mathrm{I}}(\varepsilon) = \emptyset \qquad \mathsf{c}_{\mathrm{I}}(\rho^\star) = \mathsf{c}_{\mathrm{I}}(\rho)\rho^\star$$

$$\mathsf{c}_{\mathrm{I}}(a_i) = \begin{cases} \varepsilon, & \textit{if } \mathrm{I} = \{i\} \\ \emptyset, & \textit{otherwise} \end{cases} \qquad \mathsf{c}_{\mathrm{I}}(\rho_1 + \rho_2) = \begin{cases} \mathsf{c}_{\mathrm{I}}(\rho_1), & \textit{if } \mathsf{c}_{\mathrm{I}}(\rho_1) \neq \emptyset \\ \mathsf{c}_{\mathrm{I}}(\rho_2), & \textit{otherwise} \end{cases}$$

$$\mathsf{c}_{\mathrm{I}}(\rho_1 \cdot \rho_2) = \begin{cases} \mathsf{c}_{\mathrm{I}}(\rho_1) \cdot \rho_2, & \textit{if } \mathsf{c}_{\mathrm{I}}(\rho_1) \neq \emptyset \\ \mathsf{c}_{\mathrm{I}}(\rho_2), & \textit{otherwise} \end{cases}$$

$$\mathsf{c}_{\mathrm{I}}(\rho_1 \cap \rho_2) = \begin{cases} \mathsf{c}_{\mathrm{I}|_{\rho_1}}(\rho_1) \cap \mathsf{c}_{\mathrm{I}|_{\rho_2}}(\rho_2), & \textit{if } \mathrm{I} = \mathrm{I}\big|_{\rho_1} \cup \mathrm{I}\big|_{\rho_1} \\ \emptyset, & \textit{otherwise.} \end{cases}$$

It is easy to verify that $\mathsf{c}_{\mathrm{I}}(\rho) \neq \emptyset$ implies $\mathrm{I} \subseteq \mathsf{ind}(\rho)$, i.e. $\mathrm{I}\big|_\rho = \mathrm{I}$.

Proposition 24. *Consider a linear indexed expression ρ and $\mathrm{I} \in \mathcal{I}_\rho$. Then, for every $x \in \mathcal{I}_\rho^\star$ such that $\partial_{x\mathrm{I}}(\rho) \neq \emptyset$, one has $\partial_{x\mathrm{I}}(\rho) = \{\mathsf{c}_{\mathrm{I}}(\rho)\}$ and $\mathsf{c}_{\mathrm{I}}(\rho) \neq \emptyset$.*

Proof. We proceed by induction on the structure of ρ. For $\emptyset$ and ε the set of partial index-derivatives is $\emptyset$. Let ρ be a_i. We need to prove that $\forall \mathrm{I} \in \mathcal{I}_{a_i} \forall x \in \mathcal{I}_{a_i}^\star \; (\partial_{x\mathrm{I}}(a_i) \neq \emptyset \implies \partial_{x\mathrm{I}}(a_i) = \{\mathsf{c}_{\mathrm{I}}(a_i)\} \neq \{\emptyset\})$. Let $\partial_{x\mathrm{I}}(a_i) \neq \emptyset$, then by Proposition 22, $\partial_{x\mathrm{I}}(a_i) = \{\varepsilon\}$ and $x\mathrm{I} = \{i\}$. Then $\mathrm{I} = \{i\}$ and $\mathsf{c}_{\mathrm{I}}(a_i) = \varepsilon$. Thus, we conclude that $\partial_{x\mathrm{I}}(a_i) = \{\mathsf{c}_{\mathrm{I}}(a_i)\} \neq \{\emptyset\}$. Let us suppose that for ρ_i, $i = 1, 2$ we have $\forall \mathrm{I} \in \mathcal{I}_{\rho_i} \forall x \in \mathcal{I}_{\rho_i}^\star \; (\partial_{x\mathrm{I}}(\rho_i) \neq \emptyset \implies \partial_{x\mathrm{I}}(\rho_i) = \{\mathsf{c}_{\mathrm{I}}(\rho_i)\} \neq \{\emptyset\})$. Let $\rho = \rho_1 + \rho_2$ be such that $\partial_{x\mathrm{I}}(\rho_1 + \rho_2) \neq \emptyset$. Then, $\partial_{x\mathrm{I}}(\rho_1 + \rho_2) = \partial_{x\mathrm{I}}(\rho_i)$ with $x\mathrm{I} = (x\mathrm{I})\big|_{\rho_i}$, for some $i \in \{1, 2\}$. By the induction hypothesis, $\partial_{x\mathrm{I}}(\rho_i) = \{\mathsf{c}_{\mathrm{I}}(\rho_i)\} \neq \{\emptyset\}$. Thus, $\mathsf{c}_{\mathrm{I}}(\rho_i) \neq \emptyset$ and $\mathsf{c}_{\mathrm{I}}(\rho_1 + \rho_2) = \mathsf{c}_{\mathrm{I}}(\rho_i)$. Let $\rho = \rho_1\rho_2$. If $\partial_{x\mathrm{I}}(\rho_1\rho_2) \neq \emptyset$ then we have to consider two cases. Let $\partial_{x\mathrm{I}}(\rho_1\rho_2) = \partial_{x\mathrm{I}}(\rho_1) \odot \rho_2$ and $x\mathrm{I} = (x\mathrm{I})\big|_{\rho_1}$. Then, $\partial_{x\mathrm{I}}(\rho_1) \neq \emptyset$ and $\partial_{x\mathrm{I}}(\rho_1) = \{\mathsf{c}_{\mathrm{I}}(\rho_1)\}$. We conclude that $\mathsf{c}_{\mathrm{I}}(\rho_1) \neq \emptyset$ and $\mathsf{c}_{\mathrm{I}}(\rho_1\rho_2) = \mathsf{c}_{\mathrm{I}}(\rho_1)$. In the second case, $\partial_{x\mathrm{I}}(\rho_1\rho_2) = \partial_{z\mathrm{I}}(\rho_2) \neq \emptyset$, $x = yz$, $\varepsilon(\partial_y(\rho_1)) = \varepsilon$ and $z\mathrm{I} = (z\mathrm{I})\big|_{\rho_2}$. We conclude that $y = y\big|_{\rho_1}$ and $\mathrm{I} = \mathrm{I}\big|_{\rho_2}$. Then, $\mathsf{c}_{\mathrm{I}}(\rho_1) = \emptyset$ and $\mathsf{c}_{\mathrm{I}}(\rho_1\rho_2) = \mathsf{c}_{\mathrm{I}}(\rho_2)$. By the induction hypothesis, $\partial_{z\mathrm{I}}(\rho_2) = \{\mathsf{c}_{\mathrm{I}}(\rho_2)\}$ and the result follows. Let $\rho = \rho_1^\star$. If $\partial_{x\mathrm{I}}(\rho_1^\star) \neq \emptyset$, we can write $\partial_{x\mathrm{I}}(\rho_1^\star) = \partial_{v_1\mathrm{I}}(\rho_1) \odot \rho_1^\star \cup \cdots \cup \partial_{v_n\mathrm{I}}(\rho_1) \odot \rho_1^\star$, with $n \geq 1$, such that for all $1 \leq i \leq n$, $x = u_iv_i$ and $\partial_{v_i\mathrm{I}}(\rho_1) \odot \rho_1^\star \neq \emptyset$. By the induction hypothesis, each nonempty set of partial index-derivatives $\partial_{v_i\mathrm{I}}(\rho_1)$ is equal to $\{\mathsf{c}_{\mathrm{I}}(\rho_1)\} \neq \{\emptyset\}$. Thus, $\partial_{x\mathrm{I}}(\rho_1^\star) = \{\mathsf{c}_{\mathrm{I}}(\rho_1)\rho_1^\star\}$. Finally, let $\rho = \rho_1 \cap \rho_2$ be such that $\partial_{x\mathrm{I}}(\rho_1 \cap \rho_2) \neq \emptyset$. Then $\partial_{x\mathrm{I}}(\rho_1 \cap \rho_2) = \partial_{(x\mathrm{I})|_{\rho_1}}(\rho_1) \mathbin{\dot{\cap}} \partial_{(x\mathrm{I})|_{\rho_2}}(\rho_2)$, $x\mathrm{I} = (x\mathrm{I})\big|_{\rho_1} \cap_{\mathcal{I}} (x\mathrm{I})\big|_{\rho_2}$ and $\partial_{(x\mathrm{I})|_{\rho_i}}(\rho_i) \neq \emptyset$, for $i = 1, 2$. Moreover, $\partial_{(x\mathrm{I})|_{\rho_i}}(\rho_i) = \{\mathsf{c}_{\mathrm{I}|_{\rho_i}}(\rho_i)\}$. The result follows by the induction hypothesis and from the definition of $\mathsf{c}_{\mathrm{I}}(\rho_1 \cap \rho_2)$. □

This result guarantees that, given a linear indexed expression ρ and $\mathrm{I} \in \mathcal{I}_\rho$, all sets of partial index-derivatives $\partial_{x\mathrm{I}}(\rho)$ different from $\emptyset$ are singletons with an unique c-continuation $\mathsf{c}_{\mathrm{I}}(\rho)$ of ρ w.r.t. I.

Lemma 25. *Consider a linear indexed expression ρ. Then,* $\mathrm{I} \in \mathsf{Lst}(\rho)$ *if and only if* $\varepsilon(\mathsf{c}_{\mathrm{I}}(\rho)) = \varepsilon$.

Lemma 26. *Consider a linear indexed expression ρ and sets of indexes* $\mathrm{I}, \mathrm{J} \in \mathcal{I}_\rho^\star$*. Then,* $(\mathrm{I}, \mathrm{J}) \in \mathsf{Fol}(\rho)$ *if and only if* $\mathrm{J} \in \mathsf{Fst}(\mathsf{c}_{\mathrm{I}}(\rho))$.

Definition 27. *The c-continuation automaton of an expression* $\alpha \in \mathsf{RE}_\cap$ *is*

$$\mathcal{A}_{\mathsf{c}}(\alpha) = \langle S_{\mathsf{c}}, \Sigma, \{(\{0\}, \mathsf{c}_{\{0\}}(\overline{\alpha}))\}, \delta_{\mathsf{c}}, F_{\mathsf{c}} \rangle,$$

where $S_{\mathsf{c}} = \{\ (\mathrm{I}, \mathsf{c}_{\mathrm{I}}(\overline{\alpha})) \mid \mathrm{I} \in S_{\mathsf{posi}}\ \}$, $F_{\mathsf{c}} = \{\ (\mathrm{I}, \mathsf{c}_{\mathrm{I}}(\overline{\alpha})) \mid \varepsilon(\mathsf{c}_{\mathrm{I}}(\overline{\alpha})) = \varepsilon\ \}$, $\mathsf{c}_{\{0\}}(\overline{\alpha}) = \overline{\alpha}$, $\delta_{\mathsf{c}} = \{\ ((\mathrm{I}, \mathsf{c}_{\mathrm{I}}(\overline{\alpha})), \ell(\mathrm{J}), (\mathrm{J}, \mathsf{c}_{\mathrm{J}}(\overline{\alpha}))) \mid \mathrm{J} \in \mathsf{Fst}(\mathsf{c}_{\mathrm{I}}(\overline{\alpha}))\ \}$.

By Lemma 25, Lemma 26, and considering $\varphi : S_{\mathsf{c}} \to S_{\mathsf{posi}}$ such that $\varphi((\mathrm{I}, \mathsf{c}_{\mathrm{I}}(\overline{\alpha}))) = \mathrm{I}$, the following holds.

Theorem 28. *For* $\alpha \in \mathsf{RE}_\cap$*, we have* $\mathcal{A}_{\mathsf{posi}}(\alpha) \simeq \mathcal{A}_{\mathsf{c}}(\alpha)$.

Example 29. Consider the expression $\overline{\alpha} = (b_1 a_2^\star b_3 + a_4) \cap (a_5 a_6 + b_7)^\star$, from Example 16, and let $\rho_2 = (a_5 a_6 + b_7)^\star$. We have the following c-continuations: $\mathsf{c}_{\{1,7\}}(\overline{\alpha}) = a_2^\star b_3 \cap \rho_2$, $\mathsf{c}_{\{4,5\}}(\overline{\alpha}) = \varepsilon \cap a_6 \rho_2$, $\mathsf{c}_{\{4,6\}}(\overline{\alpha}) = \varepsilon \cap \rho_2$, $\mathsf{c}_{\{2,5\}}(\overline{\alpha}) = a_2^\star b_3 \cap a_6 \rho_2$, $\mathsf{c}_{\{2,6\}}(\overline{\alpha}) = a_2^\star b_3 \cap \rho_2$, and $\mathsf{c}_{\{3,7\}}(\overline{\alpha}) = \varepsilon \cap \rho_2$.

6 The $\mathcal{A}_{\mathsf{pd}}$ as a Quotient of $\mathcal{A}_{\mathsf{pos}}$

Using $\mathcal{A}_{\mathsf{c}}$ we show that the partial derivative automaton $\mathcal{A}_{\mathsf{pd}}$ is a quotient of $\mathcal{A}_{\mathsf{pos}}$. This extends the corresponding result for simple regular expressions, although the proof cannot use the same technique. Recall that, for a simple regular expression α, one builds $\mathcal{A}_{\mathsf{pd}}(\overline{\alpha})$, and then shows that when its transitions are unmarked, the result $\overline{\mathcal{A}_{\mathsf{pd}}(\overline{\alpha})}$ is isomorphic to a quotient of $\mathcal{A}_{\mathsf{c}}(\alpha)$. However, with $\alpha \in \mathsf{RE}_\cap$, this method cannot be used because, as mentioned in the introduction, intersection does not commute with marking. For $\alpha \in \mathsf{RE}_\cap$, we will present a direct isomorphism between $\mathcal{A}_{\mathsf{pd}}(\alpha)$ and a quotient of $\mathcal{A}_{\mathsf{c}}(\alpha)$. The next lemmas will be needed to build that isomorphism.

Lemma 30. *Consider a linear indexed expression ρ and* $\mathrm{I} \in \mathcal{I}_\rho$*. If* $\mathrm{I} \in \mathsf{Fst}(\rho)$*, then* $\mathsf{c}_{\mathrm{I}}(\rho) \neq \emptyset$ *and* $\mathsf{c}_{\mathrm{I}}(\rho) \in \partial_{\mathrm{I}}(\rho)$.

Lemma 31. *Consider a linear indexed expression ρ and* $\mathrm{I}, \mathrm{J} \in \mathcal{I}_\rho$*, such that* $\mathrm{J} \in \mathsf{Fst}(\mathsf{c}_{\mathrm{I}}(\rho))$*. Then,* $\mathsf{c}_{\mathrm{J}}(\rho) \in \partial_{\mathrm{J}}(\mathsf{c}_{\mathrm{I}}(\rho))$.

Lemma 32. *Consider well-indexed expressions* ρ', ρ *and* $\mathrm{I} \in \mathcal{I}_\rho$*, such that* $\rho' \in \partial_{\mathrm{I}}(\rho)$*. Then,* $\overline{\rho'} \in \partial_{\ell(\mathrm{I})}(\overline{\rho})$.

Lemma 33. *Consider a well-indexed expression ρ,* $a \in \Sigma$ *and* $\beta \in \partial_a(\overline{\rho})$*. Then, there exist* $\mathrm{I} \in \mathcal{I}_\rho$ *and* $\rho' \in \partial_{\mathrm{I}}(\rho)$ *with* $\ell(\mathrm{I}) = a$ *and* $\overline{\rho'} = \beta$*. Furthermore, for* $x = a_1 \cdots a_n \in \Sigma^\star$*, if* $\beta \in \partial_x(\overline{\rho})$*, there exist* $\mathrm{I}_1 \cdots \mathrm{I}_n \in \mathcal{I}_\rho^\star$ *and* $\rho' \in \partial_{\mathrm{I}_1 \cdots \mathrm{I}_n}(\rho)$ *with* $\ell(\mathrm{I}_1 \cdots \mathrm{I}_n) = x$ *and* $\overline{\rho'} = \beta$.

Given $\alpha \in \mathsf{RE}_\cap$, consider $\mathcal{A}_\mathsf{c}(\alpha)$ and the equivalence relation $\equiv_\ell$ on S_c given by $(\mathrm{I}, \mathsf{c}_\mathrm{I}(\overline{\alpha})) \equiv_\ell (\mathrm{J}, \mathsf{c}_\mathrm{J}(\overline{\alpha}))$ if and only if $\overline{\mathsf{c}_\mathrm{I}(\overline{\alpha})} = \overline{\mathsf{c}_\mathrm{J}(\overline{\alpha})}$, for $\mathrm{I}, \mathrm{J} \in \mathcal{I}_{\overline{\alpha}} \cup \{\{0\}\}$.

Lemma 34. *The relation $\equiv_\ell$ is right invariant w.r.t. $\mathcal{A}_\mathsf{c}$.*

Theorem 35. *For $\alpha \in \mathsf{RE}_\cap$, $\mathcal{A}_\mathsf{pd}(\alpha) \simeq \mathcal{A}_\mathsf{c}(\alpha)^\mathsf{ac}/{\equiv_\ell}$.*

Proof. Let $\mathcal{A}_\mathsf{c}(\alpha)^\mathsf{ac}/{\equiv_\ell} = (S_\ell, \Sigma, \delta_\ell, [(\{0\}, \overline{\alpha})], F_\ell)$. We define the map $\varphi : S_\ell \to \partial(\alpha)$, by $\varphi([(\mathrm{I}, \mathsf{c}_\mathrm{I}(\overline{\alpha}))]) = \overline{\mathsf{c}_\mathrm{I}(\overline{\alpha})}$. We have to show that: 1) φ is well-defined; 2) φ is bijective; 3) $\varphi(\delta_\ell(s, a)) = \delta_\mathsf{pd}(\varphi(s), a)$ for every $s \in S_\ell, a \in \Sigma$; 4) $\varphi(F_\ell) = F_\mathsf{pd}$; 5) $\varphi([(\{0\}, \mathsf{c}_{\{0\}}(\overline{\alpha}))]) = \alpha$.

Claim 1) follows from lemmas 30 and 31. The last two are obvious. That φ is injective follows from the definition of $\equiv_\ell$. Furthermore, if $\beta \in \partial(\alpha)$, then there are terms $\beta_0 = \alpha, \beta_1, \ldots, \beta_n = \beta$ and letters $a_1, \ldots, a_n \in \Sigma$, with $n \geq 0$, such that $\beta_{i+1} \in \partial_{a_{i+1}}(\beta_i)$ for $0 \leq i \leq n-1$. It follows from Lemma 33 that there exist $\mathrm{I}_1 \cdots \mathrm{I}_n \in \mathcal{I}_\rho^\star$ and $\rho' \in \partial_{\mathrm{I}_1 \cdots \mathrm{I}_n}(\overline{\alpha})$ with $\ell(\mathrm{I}_1 \cdots \mathrm{I}_n) = a_1 \cdots a_n$ and $\overline{\rho'} = \beta$. Furthermore, by Proposition 24, we know that $\partial_{\mathrm{I}_1 \cdots \mathrm{I}_n}(\overline{\alpha}) = \{\mathsf{c}_{\mathrm{I}_n}(\overline{\alpha})\}$, with $\mathsf{c}_{\mathrm{I}_n}(\overline{\alpha}) \neq \emptyset$. Thus, $[(\mathrm{I}_n, \mathsf{c}_{\mathrm{I}_n}(\overline{\alpha}))] \in S_\ell$ and we conclude that φ is surjective. For 3) we consider both inclusions. Consider $\beta \in \varphi(\delta_\ell(s, a))$, for $s \in S_\ell$ and $a \in \Sigma$. Then, there exist $\mathrm{I}, \mathrm{J} \in \mathcal{I}_{\overline{\alpha}}$ such that $[(\mathrm{I}, \mathsf{c}_\mathrm{I}(\overline{\alpha}))] = s$, $\overline{\mathsf{c}_\mathrm{J}(\overline{\alpha})} = \beta$, $(\mathrm{J}, \mathsf{c}_\mathrm{J}(\overline{\alpha})) \in \delta_\mathsf{c}((\mathrm{I}, \mathsf{c}_\mathrm{I}(\overline{\alpha})), \ell(\mathrm{J}))$ and $\ell(\mathrm{J}) = a$, i.e. $\mathrm{J} \in \mathsf{Fst}(\mathsf{c}_\mathrm{I}(\overline{\alpha}))$. By Lemma 31, we have $\mathsf{c}_\mathrm{J}(\overline{\alpha}) \in \partial_\mathrm{J}(\mathsf{c}_\mathrm{I}(\overline{\alpha}))$ and by Lemma 32, $\overline{\mathsf{c}_\mathrm{J}(\overline{\alpha})} \in \partial_a(\overline{\mathsf{c}_\mathrm{I}(\overline{\alpha})})$. Thus, $\overline{\mathsf{c}_\mathrm{J}(\overline{\alpha})} \in \delta_\mathsf{pd}(\overline{\mathsf{c}_\mathrm{I}(\overline{\alpha})}, a)$. Now, let $\beta \in \delta_\mathsf{pd}(\tau, a)$, where $\tau = \overline{\mathsf{c}_\mathrm{I}(\overline{\alpha})}$, for some $\mathrm{I} \in \mathcal{I}_{\overline{\alpha}}$ and $a \in \Sigma$. Then, there is a sequence of terms $\tau_0 = \alpha, \tau_1, \ldots, \tau_n = \tau$ and a sequence of letters $a_1, \ldots, a_n \in \Sigma$ such that $\tau_{i+1} \in \partial_{a_{i+1}}(\tau_i)$, for $0 \leq i \leq n-1$, and $\beta \in \partial_a(\tau)$, i.e. $\beta \in \partial_{a_1 \cdots a_n a}(\alpha)$. By Lemma 33, there exist $\mathrm{J}_1, \ldots, \mathrm{J}_n, \mathrm{J} \in \mathcal{I}_{\overline{\alpha}}$, with $\ell(\mathrm{J}_1 \cdots \mathrm{J}_n \mathrm{J}) = a_1 \cdots a_n a$, and $\rho' \in \partial_{\mathrm{J}_1 \cdots \mathrm{J}_n \mathrm{J}}(\overline{\alpha})$ such that $\overline{\rho'} = \beta$. By Proposition 24, $\rho' = \mathsf{c}_\mathrm{J}(\overline{\alpha})$. On the other hand, it is straightforward to show by structural induction on a well-indexed expression ρ, that $\partial_\mathrm{J}(\rho) \neq \emptyset$ implies $\mathrm{J} \in \mathsf{Fst}(\rho)$. Thus, $[(\mathrm{J}, \mathsf{c}_\mathrm{J}(\overline{\alpha}))] \in \delta_\ell([(\mathrm{I}, \mathsf{c}_\mathrm{I}(\overline{\alpha}))], \ell(\mathrm{J}))$ and consequently $\beta = \overline{\mathsf{c}_\mathrm{J}(\overline{\alpha})} \in \varphi(\delta_\ell([(\mathrm{I}, \mathsf{c}_\mathrm{I}(\overline{\alpha}))], a))$. □

Example 36. Consider $\alpha = (ba^\star b + a) \cap (aa + b)^\star$ from examples 16 and 29. Set $\beta = (aa + b)^\star$. For the positions present in $\mathcal{A}_\mathsf{c}(\alpha)^\mathsf{ac}$, we have $\overline{\mathsf{c}_{\{4,5\}}(\overline{\alpha})} = \varepsilon \cap a\beta$, $\overline{\mathsf{c}_{\{3,7\}}(\overline{\alpha})} = \varepsilon \cap \beta$, $\overline{\mathsf{c}_{\{2,5\}}(\overline{\alpha})} = a^\star b \cap a\beta$, and $\overline{\mathsf{c}_{\{1,7\}}(\overline{\alpha})} = \overline{\mathsf{c}_{\{2,6\}}(\overline{\alpha})} = a^\star b \cap \beta$. Merging states $(\{1, 7\}, \mathsf{c}_{\{1,7\}}(\overline{\alpha}))$ and $(\{2, 6\}, \mathsf{c}_{\{2,6\}}(\overline{\alpha}))$ in $\mathcal{A}_\mathsf{c}(\alpha)^\mathsf{ac}$, one obtains an NFA isomorphic to $\mathcal{A}_\mathsf{pd}(\alpha)$, which is represented in Figure 2.

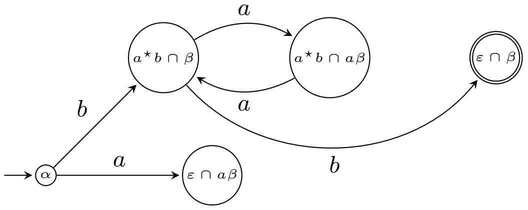

Fig. 2. $\mathcal{A}_\mathsf{pd}((ba^\star b + a) \cap (aa + b)^\star)$

7 Final Remarks

For simple regular expressions of size n, the size of $\mathcal{A}_{\mathsf{pos}}(\alpha)$ is $O(n^2)$, and using $\mathcal{A}_{\mathsf{c}}(\alpha)$ it is possible to efficiently compute $\mathcal{A}_{\mathsf{pd}}(\alpha)$ [9]. For regular expressions with intersection the conversion to NFA's has exponential computational complexity [11] and both the size of $\mathcal{A}_{\mathsf{pos}}$ and $\mathcal{A}_{\mathsf{pd}}$ can be exponential in the size of the regular expression. On the average case, however, the size of these automata seem to be much smaller [2], and thus feasible for practical applications. In this scenario, algorithms for building $\mathcal{A}_{\mathsf{pd}}$ using $\mathcal{A}_{\mathsf{pos}}$ seem worthwhile to develop.

References

1. Antimirov, V.: Partial derivatives of regular expressions and finite automaton constructions. Theoret. Comput. Sci. 155(2), 291–319 (1996)
2. Bastos, R., Broda, S., Machiavelo, A., Moreira, N., Reis, R.: Partial derivative automaton for regular expressions with intersection. In: 18th DCFS. LNCS, Springer (2016)
3. Berry, G., Sethi, R.: From regular expressions to deterministic automata. Theoret. Comput. Sci. 48, 117–126 (1986)
4. Brüggemann-Klein, A.: Regular expressions into finite automata. Theoret. Comput. Sci. 48, 197–213 (1993)
5. Brzozowski, J.: Derivatives of regular expressions. JACM 11(4), 481–494 (1964)
6. Caron, P., Champarnaud, J., Mignot, L.: Partial derivatives of an extended regular expression. In: Dediu, A.H., Inenaga, S., Martín-Vide, C. (eds.) 5th LATA. LNCS, vol. 6638, pp. 179–191. Springer (2011)
7. Caron, P., Champarnaud, J., Mignot, L.: A general framework for the derivation of regular expressions. RAIRO - Theor. Inf. and Applic. 48(3), 281–305 (2014)
8. Caron, P., Ziadi, D.: Characterization of Glushkov automata. Theoret. Comput. Sci. 233(1-2), 75–90 (2000)
9. Champarnaud, J.M., Ziadi, D.: Canonical derivatives, partial derivatives and finite automaton constructions. Theoret. Comput. Sci. 289, 137–163 (2002)
10. Chen, H., Yu, S.: Derivatives of regular expressions and an application. In: Dinneen, M.J., Khoussainov, B., Nies, A. (eds.) Computation, Physics and Beyond, WTCS 2012. LNCS, vol. 7160, pp. 343–356. Springer (2012)
11. Gelade, W.: Succinctness of regular expressions with interleaving, intersection and counting. Theor. Comput. Sci. 411(31-33), 2987–2998 (2010)
12. Glushkov, V.M.: The abstract theory of automata. Russian Math. Surveys 16, 1–53 (1961)
13. Hopcroft, J.E., Ullman, J.D.: Introduction to Automata Theory, Languages and Computation. Addison Wesley (1979)
14. Ilie, L., Yu, S.: Follow automata. Inf. Comput. 186(1), 140–162 (2003)
15. McNaughton, R., Yamada, H.: Regular expressions and state graphs for automata. IEEE Transactions on Electronic Computers 9, 39–47 (1960)
16. Sakarovitch, J.: Elements of Automata Theory. Cambridge University Press (2009)
17. Yu, S.: Regular languages. In: Rozenberg, G., Salomaa, A. (eds.) Handbook of Formal Languages, vol. 1, pp. 41–110. Springer (1997)