On the Average State Complexity of Partial Derivative Transducers

Stavros Konstantinidis[1], António Machiavelo[2],
Nelma Moreira[2], and Rogério Reis[2] [*]

[1] Saint Mary's University, Halifax, Nova Scotia, Canada,
`s.konstantinidis@smu.ca`
[2] CMUP & DM-DCC, Faculdade de Ciências da Universidade do Porto, Portugal
`ajmachia@fc.up.pt,{nam,rvr}@dcc.fc.up.pt`

Abstract. $2D$ regular expressions represent rational relations over two alphabets. In this paper we study the average state complexity of partial derivative standard transducers ($\mathcal{T}_{\mathrm{PD}}$) that can be defined for (general) $2D$ expressions where basic terms are pairs of ordinary regular expressions ($1D$). While in the worst case the number of states of $\mathcal{T}_{\mathrm{PD}}$ can be $O(n^2)$, where n is the size of the expression, asymptotically and on average that value is bounded from above by $O(n^{\frac{3}{2}})$. Moreover, asymptotically and on average the alphabetic size of a $2D$ expression is half of the size of that expression. All results are obtained in the framework of analytic combinatorics considering generating functions of parametrised combinatorial classes defined implicitly by algebraic curves. In particular, we generalise the methods developed in previous work to a broad class of analytic functions.

1 Introduction

We consider $2D$ expressions that represent rational relations over two alphabets. Expressions and transducers with labels over finitely generated monoids were studied by Konstantinidis et al. [12, 11], and also by Demaille [7]. Partial derivative methods have become a standard method to manipulate several kinds of expressions [1, 14, 7, 2, 6, 5], not only because they are in general more succinct than other equivalent constructions, but for some operators they are easier to define (e.g. for intersection [2]). For regular languages, the average complexity of partial derivative automata ($\mathcal{A}_{\mathrm{PD}}$), considering different sets of operations, has been studied [3, 2, 5]. Using the framework of analytic combinatorics, for ordinary ($1D$) regular expressions of (tree-)size n (with concatenation, union and Kleene star) it was shown that, asymptotically and on average, the number of states of $\mathcal{A}_{\mathrm{PD}}$ is $\frac{1}{4}n$, (being the worst-case $O(n^2)$) while for expressions with intersection of (tree-)size n that number is upper bounded by $(1.056 + o(1))^n$ (being the worst-case $O(2^n)$) [3, 4, 2]. In this paper we consider general $2D$ expressions where

basic terms are pairs of $1D$ regular expressions. We define a partial derivative standard transducer construction ($\mathcal{T}_{\text{PD}}$) from these expressions, and study its average state complexity. The analytic combinatorial methods used for ordinary $1D$ regular expressions could not be applied for $2D$ expressions. In particular, to get explicit expressions for the generating functions involved would be unmanageable. So, generating functions implicitly defined by algebraic curves must be used, and in previous work it was shown how to get the required information for the asymptotic estimates with an indirect use of the existence of Puiseux expansions at singularities [6]. In this paper, as the involved algebraic curves are more intricate, we needed to refine the methods described in the literature, and use Puiseux expansions together with the Newton's polygon technique to find the estimates for the asymptotic behaviours of parametrised families of combinatorial classes. This new, more refined, method is introduced in Section 4. Section 2 reviews the partial derivative construction for ordinary $1D$ regular expressions. In Section 3 we define $2D$ expressions, and present the corresponding construction of partial derivative transducers ($\mathcal{T}_{\text{PD}}$). Section 5 presents the average complexity results obtained using the framework of Section 4. We show that for general $2D$ expressions, while in the worst case the number of states of $\mathcal{T}_{\text{PD}}$ can be $O(n^2)$, where n is the size of the expression, asymptotically and on average, that value is bounded from above by $O(n^{\frac{3}{2}})$. Restricting to pairs of $1D$ expressions, the previous bound is already reached, showing that these kind of expressions are responsible for the increasing of complexity. Furthermore, the same bounds apply to sums or concatenations of pairs of $1D$ expressions, i.e., regular relations.

2   Preliminares

A *nondeterministic finite automaton* (NFA) is a five-tuple $A = \langle Q, \Sigma, \delta, I, F \rangle$ 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 size of an NFA is its number of states. The transition function can be extended to words and to sets of states in the natural way. When $I = \{q_0\}$, we use $I = q_0$. The *language accepted* by A is $\mathcal{L}(A) = \{w \in \Sigma^\star \mid \delta(I, w) \cap F \neq \emptyset\}$. Given an alphabet Σ, the set RE of $(1D)$ *regular expressions*, $\mathbf{r}$, over Σ contains $\emptyset$ and the expressions defined by the following grammar:

$$\mathbf{r} := \varepsilon \mid \sigma \in \Sigma \mid (\mathbf{r} + \mathbf{r}) \mid (\mathbf{r} \cdot \mathbf{r}) \mid (\mathbf{r}^\star), \tag{1}$$

where the operator $\cdot$ (concatenation) and the outermost parentheses are often omitted. The *language* associated to $\mathbf{r}$ is denoted by $\mathcal{L}(\mathbf{r})$ and defined as usual (with ε representing the empty word). Two expressions $\mathbf{r}_1$ and $\mathbf{r}_2$ are *equivalent*, $\mathbf{r}_1 \sim \mathbf{r}_2$ if $\mathcal{L}(\mathbf{r}_1) = \mathcal{L}(\mathbf{r}_2)$. If $S \subseteq \text{RE}$, $\mathcal{L}(S) = \cup_{\mathbf{r} \in S} \mathcal{L}(\mathbf{r})$. The *(tree-)size* $|\mathbf{r}|$ of $\mathbf{r} \in \text{RE}$ is the number of symbols in $\mathbf{r}$ (disregarding parentheses). The *alphabetic size* $|\mathbf{r}|_\Sigma$ is the number of letters occurring in $\mathbf{r}$. We define the *constant part* of $\mathbf{r}$, $\mathsf{c}(\mathbf{r})$, by $\mathsf{c}(\mathbf{r}) = \varepsilon$ if $\varepsilon \in \mathcal{L}(\mathbf{r})$, and $\mathsf{c}(\mathbf{r}) = \emptyset$ otherwise. This function is extended to sets of expressions by $\mathsf{c}(S) = \varepsilon$ if and only if exists $\mathbf{r} \in S$ such

that $c(\mathbf{r}) = \varepsilon$. In the case of a singleton $\{s\}$ we write it simply as s. Given $L \subseteq \Sigma^\star$ and $\sigma \in \Sigma$, let $\sigma^{-1}L = \{w \mid \sigma w \in L\}$. This notion can be extended to words and languages. The partial derivative automaton of a regular expression was introduced independently by Mirkin [15] and Antimirov [1]. For a regular expression $\mathbf{r} \in \mathrm{RE}$, let the *linear form* of $\mathbf{r}$, $n : \mathrm{RE} \to 2^{\Sigma \times \mathrm{RE}}$, be inductively defined by

$$\begin{aligned} n(\emptyset) = n(\varepsilon) = \emptyset, && n(\mathbf{r} + \mathbf{r}') = n(\mathbf{r}) \cup n(\mathbf{r}'), \\ n(\sigma) = \{(\sigma, \varepsilon)\}, && n(\mathbf{r}\mathbf{r}') = n(\mathbf{r})\mathbf{r}' \cup c(\mathbf{r})\, n(\mathbf{r}'), \\ && n(\mathbf{r}^\star) = n(\mathbf{r})\mathbf{r}^\star, \end{aligned} \qquad (2)$$

where for any $S \subseteq \Sigma \times \mathrm{RE}$, we define $S\emptyset = \emptyset S = \emptyset$, $S\varepsilon = \varepsilon S = S$, and $S\mathbf{r}' = \{(\sigma, \mathbf{r}\mathbf{r}') \mid (\sigma, \mathbf{r}) \in S \wedge \mathbf{r} \neq \varepsilon\} \cup \{(\sigma, \mathbf{r}') \mid \exists (\sigma, \varepsilon) \in S\}$ if $\mathbf{r}' \neq \emptyset, \varepsilon$ (and analogously for $\mathbf{r}'S$).

Proposition 1 ([1]). *For all $\boldsymbol{r} \in \mathrm{RE}$, $\boldsymbol{r} \sim \bigcup_{(\sigma, r') \in n(r)} \sigma r' \cup c(\boldsymbol{r})$.*

For a regular expression $\mathbf{r} \in \mathrm{RE}$ and a symbol $\sigma \in \Sigma$, *the set of partial derivatives of $\boldsymbol{r}$ w.r.t. σ* is defined by $\partial_\sigma(\mathbf{r}) = \{\mathbf{r}' \mid (\sigma, \mathbf{r}') \in n(\mathbf{r})\}$. We have $\mathcal{L}(\partial_\sigma(\mathbf{r})) = \sigma^{-1}\mathcal{L}(\mathbf{r})$. Partial derivatives can be extended w.r.t words and set of partial derivatives of an expression $\mathbf{r}$ can be defined by iterating the linear form. Let $\pi_0(\mathbf{r}) = \downarrow_2(n(\mathbf{r}))$, where $\downarrow_2(s, t) = t$ is the standard second projection on pairs of objects and naturally extended to sets of pairs. Iteratively applying the operator π_0 we have, $\pi_i(\mathbf{r}) = \pi_0(\pi_{i-1}(\mathbf{r}))$, for $i \in \mathbb{N}$, and $\pi(\mathbf{r}) = \bigcup_{i \in \mathbb{N}_0} \pi_i(\mathbf{r})$. The set $\mathrm{PD}(\mathbf{r}) = \pi(\mathbf{r}) \cup \{\mathbf{r}\}$ is the *set of partial derivatives* of $\mathbf{r}$ and $\pi(\mathbf{r})$ is the *support* [3].

Proposition 2 ([15]). *The support $\pi(\boldsymbol{r})$ is inductively defined by*

$$\begin{aligned} \pi(\emptyset) &= \emptyset, & \pi(\boldsymbol{r} + \boldsymbol{r}') &= \pi(\boldsymbol{r}) \cup \pi(\boldsymbol{r}'), \\ \pi(\varepsilon) &= \emptyset, & \pi(\boldsymbol{r}\boldsymbol{r}') &= \pi(\boldsymbol{r})\boldsymbol{r}' \cup \pi(\boldsymbol{r}'), \\ \pi(\sigma) &= \{\varepsilon\}, & \pi(\boldsymbol{r}^\star) &= \pi(\boldsymbol{r})\boldsymbol{r}^\star, \end{aligned}$$

where, for any $S \subseteq \mathrm{RE}$, we define $S\emptyset = \emptyset S = \emptyset$, $S\varepsilon = \varepsilon S = S$, and $S\boldsymbol{r}' = \{\boldsymbol{r}\boldsymbol{r}' \mid \boldsymbol{r} \in S \wedge \boldsymbol{r} \neq \varepsilon\} \cup \{\boldsymbol{r}' \mid \exists \varepsilon \in S\}$ if $\boldsymbol{r}' \neq \emptyset, \varepsilon$ (and analogously for $\boldsymbol{r}'S$).

Proposition 3 ([1, 15]). $|\pi(\boldsymbol{r})| \leq |\boldsymbol{r}|_\Sigma$ *and* $|\mathrm{PD}(\boldsymbol{r})| \leq |\boldsymbol{r}|_\Sigma + 1$.

The *partial derivative automaton* of $\mathbf{r}$ is $\mathcal{A}_{\mathrm{PD}}(\mathbf{r}) = \langle \mathrm{PD}(\mathbf{r}), \Sigma, \delta_{\mathrm{PD}}, \mathbf{r}, F \rangle$, where $F = \{\mathbf{r}_1 \in \mathrm{PD}(\mathbf{r}) \mid c(\mathbf{r}_1) = \varepsilon\}$, and $\delta_{\mathrm{PD}} = \{(\mathbf{r}_1, \sigma, \mathbf{r}') \mid \mathbf{r}_1 \in \mathrm{PD}(\mathbf{r}) \wedge (\sigma, \mathbf{r}') \in n(\mathbf{r}_1)\}$.

Proposition 4 ([1, 15]). *For all $\boldsymbol{r} \in \mathrm{RE}$, $\mathcal{L}(\mathcal{A}_{\mathrm{PD}}(\boldsymbol{r})) = \mathcal{L}(\boldsymbol{r})$.*

---

[3] Extending partial derivatives w.r.t. words, one could also define $\mathrm{PD}(r) = \bigcup_{w \in \Sigma^\star} \partial_w(\mathbf{r})$.

3 $2D$ Expressions

Let Σ and Δ be two alphabets. A relation R is any subset of $\Sigma^\star \times \Delta^\star$. The concatenation of two relations R and S is the relation $RS = \{(u_1u_2, v_1v_2) \mid (u_1, v_1) \in R \wedge (u_2, v_2) \in S\}$. The Kleene closure of the relation R is the relation $R^\star = \bigcup_{n \geq 0} R^n$. The monoid $\Sigma^\star \times \Delta^\star$ has the identity $(\varepsilon, \varepsilon)$, and the following set of generators $\{(\sigma, \varepsilon), (\varepsilon, \tau) \mid \sigma \in \Sigma \wedge \tau \in \Delta\}$ with the set of equations

$$\{ (\sigma, \varepsilon)(\varepsilon, \tau) \doteq (\sigma, \tau), (\varepsilon, \tau)(\sigma, \varepsilon) \doteq (\sigma, \tau) \mid \sigma \in \Sigma \wedge \tau \in \Delta \}. \qquad (3)$$

For a relation $R \subseteq \Sigma^\star \times \Delta^\star$, the quotient of R by a symbol is defined as before, but one needs to take into account the above equations. For instance, for $\sigma \in \Sigma$ and $\tau \in \Delta$, $(\sigma, \varepsilon)^{-1}R = \{ (\varepsilon, \tau)w \mid (\sigma, \tau)w \in R\}$ and $(\varepsilon, \tau)^{-1}R = \{ (\sigma, \varepsilon)w \mid (\sigma, \tau)w \in R\}$. The set of *rational relations* is the smallest set of relations that contains the finite relations and is closed under union, concatenation and Kleene closure. Rational relations are accepted by transducers. A *finite transducer in standard-form* (SFT) over two alphabets Σ and Δ is defined as an NFA, except that the transition function is $\delta : Q \times (\Sigma_\varepsilon \times \Delta_\varepsilon) \to 2^Q$, where for a set X, $X_\varepsilon = X \cup \{\varepsilon\}$. The relation realised by an SFT t is denoted by $\mathcal{R}(t)$. In this section we consider $2D$ expressions that represent rational relations. The notions of linear form, of partial derivative and of partial derivative transducers are extend to $2D$ expressions. In Section 5 we study the average state complexity of these transducers. Recently, Demaille [7] defined derivative automata for multitape weighted regular expressions. The expressions and transducers studied in this paper are restrictions of those models to two tapes and the Boolean semiring.

To represent rational relations one could just consider $1D$ expressions where basic terms are the generators of $\Sigma^\star \times \Delta^\star$. Those expressions are called *standard $2D$ regular expressions* (S2D-RE) and are a particular case of the ones considered in [11]. For *standard $2D$ regular expressions*, and using the same methods, it can be shown that the asymptotic bounds for partial derivative transducers are the same as for partial derivative automata (for $1D$ expressions) [3].

A *(general) $2D$ regular expression* (2D-RE) over Σ and Δ , where Σ is the *input alphabet* and Δ the *output alphabet*, is an expression that is either $\emptyset$, or can be defined by the following grammar

$$\mathbf{g} := \mathbf{r}/\mathbf{r}' \mid (\mathbf{g} + \mathbf{g}) \mid (\mathbf{g} \cdot \mathbf{g}) \mid (\mathbf{g}^\star), \qquad (4)$$

where $\mathbf{r} \in \mathrm{RE}$ over Σ and $\mathbf{r}' \in \mathrm{RE}$ over Δ. The *relation* $\mathcal{R}(\mathbf{g}) \subseteq \Sigma^\star \times \Delta^\star$ realised by a 2D-RE $\mathbf{g}$ is defined inductively as follows $\mathcal{R}(\mathbf{r}/\mathbf{r}') = \mathcal{L}(\mathbf{r}) \times \mathcal{L}(\mathbf{r}')$, $\mathcal{R}(\mathbf{g} \cdot \mathbf{g}') = \mathcal{R}(\mathbf{g})\mathcal{R}(\mathbf{g}')$, and $\mathcal{R}(\mathbf{g}^\star) = (\mathcal{R}(\mathbf{g}))^\star$. Two expressions $\mathbf{g}, \mathbf{g}'$ are *equivalent*, $\mathbf{g} \sim \mathbf{g}'$, if $\mathcal{R}(\mathbf{g}) = \mathcal{R}(\mathbf{g}')$. A relation is *rational* if and only if it is represented by a 2D-RE[4]. The *constant part* of a 2D-RE expression $\mathbf{g}$ is given by $\mathsf{c} : \text{2D-RE} \longrightarrow \{\emptyset, \varepsilon/\varepsilon\}$ such that $\mathsf{c}(\mathbf{g}) = \varepsilon/\varepsilon$ if $(\varepsilon, \varepsilon) \in \mathcal{R}(\mathbf{g})$, and $\mathsf{c}(\mathbf{g}) = \emptyset$, otherwise. For $S \subseteq \text{2D-RE}$ or $S \subseteq (\Sigma_\varepsilon \times \Delta_\varepsilon) \times \text{2D-RE}$ and an expression $\mathbf{g}$, we adopt the same conventions as for $1D$ expressions regarding $\mathbf{g}S$ and $S\mathbf{g}$. In particular, we let $(\varepsilon/\varepsilon)S = S(\varepsilon/\varepsilon) = S$ (and also $S\varepsilon = \varepsilon S = S$).

---

[4] This follows from the definition above.

For the *linear form* of an expression $\mathbf{g} \in \text{2D-RE}$, $\mathsf{n} : \text{2D-RE} \to 2^{(\Sigma_\varepsilon \times \Delta_\varepsilon) \times \text{2D-RE}}$, one only needs to extend the definition for expressions of the form $\mathbf{r}_1/\mathbf{r}_2$, being the remaining cases as in Equation (2), considering expressions $\mathbf{g} \in \text{2D-RE}$. We note that one possibility was to consider $\mathsf{n}(\mathbf{r}_1/\mathbf{r}_2) = \{(\mathbf{r}_1/\mathbf{r}_2, \varepsilon/\varepsilon)\}$ (see [11]), but then one could not construct directly an SFT. Here we define

$$\mathsf{n}(\mathbf{r}_1/\mathbf{r}_2) = (\mathsf{n}(\mathbf{r}_1) \,\|\, \mathsf{n}(\mathbf{r}_2)) \,\cup\, \mathsf{c}(\mathbf{r}_2)(\mathsf{n}(\mathbf{r}_1) \,\|\, \{(\varepsilon, \varepsilon)\}) \qquad (5)$$
$$\cup\, \mathsf{c}(\mathbf{r}_1)(\{(\varepsilon, \varepsilon)\} \,\|\, \mathsf{n}(\mathbf{r}_2)),$$

where for $N \subseteq \Sigma_\varepsilon \times \text{RE}$ and $M \subseteq \Delta_\varepsilon \times \text{RE}$,

$$N \,\|\, M = \{\, ((\gamma, \gamma'), \mathbf{r}/\mathbf{r}') \mid (\gamma, \mathbf{r}) \in N \,\wedge\, (\gamma', \mathbf{r}') \in M \,\}.$$

The correctness of the previous definition is given by the following proposition.

Proposition 5. *For all $\mathbf{r}_1, \mathbf{r}_2 \in \text{RE}$, $\mathbf{r}_1/\mathbf{r}_2 \sim \displaystyle\bigcup_{((\gamma, \gamma'), \mathbf{g}') \in \mathsf{n}(\mathbf{r}_1/\mathbf{r}_2)} (\gamma/\gamma')\mathbf{g}' \cup \mathsf{c}(\mathbf{r}_1/\mathbf{r}_2).$*

Then, we have

Proposition 6. *For all $\mathbf{g} \in \text{2D-RE}$, $\mathbf{g} \sim \displaystyle\bigcup_{((\gamma, \gamma'), \mathbf{g}') \in \mathsf{n}(\mathbf{g})} (\gamma/\gamma')\mathbf{g}' \cup \mathsf{c}(\mathbf{g}).$*

As before, one can obtain the *support* of an expression $\mathbf{g}$, $\pi(\mathbf{g})$, by iterating the linear form. Only the base case differs from the ones in Proposition 2.

Proposition 7. *For all $\mathbf{r}_1, \mathbf{r}_2 \in \text{RE}$,*

$$\pi(\mathbf{r}_1/\mathbf{r}_2) \subseteq \pi(\mathbf{r}_1) \,\|\, \pi(\mathbf{r}_2) \,\cup\, \pi(\mathbf{r}_1) \,\|\, \{\varepsilon\} \,\cup\, \{\varepsilon\} \,\|\, \pi(\mathbf{r}_2),$$

where for $S, T \subseteq \text{RE}$, $S \,\|\, T = \{\, \mathbf{r}/\mathbf{r}' \mid \mathbf{r} \in S \,\wedge\, \mathbf{r}' \in T \,\}.$

Note that the inclusion in Proposition 7 can be strict, as $\pi(ab/abc) = \{b/bc, \varepsilon/c, \varepsilon/\varepsilon\}$, $\pi(ab) = \{\varepsilon, b\}$, $\pi(abc) = \{\varepsilon, bc, c\}$ and $\mathsf{c}(ab) = \mathsf{c}(abc) = \emptyset$. Proposition 7 and Proposition 2 ensure that for every $\mathbf{g} \in \text{2D-RE}$, the support $\pi(\mathbf{g})$ is finite and in the worst-case of size $O(n^2)$, where n is the size of $\mathbf{g}$. The quadratic blow-up is achieved if one considers $\mathbf{r}_n = (a^\star)^n$, $n \geq 1$, and the 2D-RE $\mathbf{r}_n/\mathbf{r}_n$.

Corollary 8. *For all $\mathbf{g} \in \text{2D-RE}$, $|\pi(\mathbf{g})| \leq (|\mathbf{g}|_{\Sigma \cup \Delta})^2$, where $|\mathbf{g}|_{\Sigma \cup \Delta}$ is the alphabetic size of $\mathbf{g}$.*

The *partial derivative transducer* of $\mathbf{g}$ is $\mathcal{T}_{\text{PD}}(\mathbf{g}) = \langle \pi(\mathbf{g}) \cup \{\mathbf{g}\}, \Sigma, \Delta, \delta_{\text{PD}}, \mathbf{g}, F \rangle$, where $F = \{\, \mathbf{g}_1 \in \pi(\mathbf{g}) \cup \{\mathbf{g}\} \mid \mathsf{c}(\mathbf{g}_1) = \varepsilon/\varepsilon \,\}$, and $\delta_{\text{PD}} = \{\, (\mathbf{g}_1, (\gamma, \gamma'), \mathbf{g}') \mid \mathbf{g}_1 \in \pi(\mathbf{g}) \cup \{\mathbf{g}\} \wedge ((\gamma, \gamma'), \mathbf{g}') \in \mathsf{n}(\mathbf{g}_1) \,\}$.

Proposition 9. *For all $\mathbf{g} \in \text{2D-RE}$, $\mathcal{R}(\mathcal{T}_{\text{PD}}(\mathbf{g})) = \mathcal{R}(\mathbf{g})$.*

An upper bound of the number of states of $\mathcal{T}_{\text{PD}}(\mathbf{g})$ is obtained if one assumes that

$$\pi(\mathbf{r}_1/\mathbf{r}_2) = \pi(\mathbf{r}_1) \,\|\, \pi(\mathbf{r}_2) \,\cup\, \pi(\mathbf{r}_1) \,\|\, \{\varepsilon\} \,\cup\, \{\varepsilon\} \,\|\, \pi(\mathbf{r}_2)$$

always holds, and as usual $\pi(\mathbf{g} + \mathbf{g}') = \pi(\mathbf{g}) \cup \pi(\mathbf{g}')$, $\pi(\mathbf{g}\mathbf{g}') = \pi(\mathbf{g})\mathbf{g}' \cup \pi(\mathbf{g}')$, and $\pi(\mathbf{g}^\star) = \pi(\mathbf{g})\mathbf{g}^\star$. These equalities are used in Section 5 to obtain an upper bound for the average case size of partial derivative transducers. In the next section we set up the analitic combinatorics framework that allows to obtain those estimates.

4   The Analytic Combinatorics Framework

Given some measure of the objects of a combinatorial class, $\mathcal{A}$, for each $n \in \mathbb{N}_0$ let a_n be the sum of the values of this measure for all objects of size n. Let $A(z) = \sum_n a_n z^n$ be the corresponding generating function (*cf.* [8]). We will use the notation $[z^n]A(z)$ for a_n. The generating function $A(z)$ can be seen as a complex analytic function, and the study of its behaviour around its dominant singularity ρ, when unique, gives us access to the asymptotic form of its coefficients. In particular, if $A(z)$ is analytic in some indented disc neighbourhood of ρ, then one has the following [8, 4]:

Theorem 10. *The coefficients of the series expansion of the complex function*

$$f(z) = (1 - z)^\alpha,$$

where $\alpha \in \mathbb{C} \setminus \mathbb{N}_0$, have the following asymptotic approximation:

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

Here Γ is Euler's gamma function.

The combinatorial classes that we deal with in the present paper give rise to generating functions implicitly defined by algebraic curves that are quite a bit more convoluted than those previously described in the literature. We, therefore, needed to refine the method to pursue these calculations, and we will expound that, in some detail, here. Generically, from an unambiguous generating grammar, one obtains a set of polynomial equations involving the generating functions for the objects corresponding to the variables of the grammar, in particular the one whose coefficients we want to asymptotically estimate. Computing a Gröbner basis for the ideal generated by those polynomials, one gets an algebraic equation for that generating function $w = w(z)$, i.e., an equation of the form

$$G(z, w) = 0,$$

where $G(z, w)$ is a polynomial in $\mathbb{Z}[z][w]$ of which $w(z)$ is a root.

Since $w(z)$ is the generating function of a combinatorial class, thus a series with non-negative integer coefficients which is not a polynomial, it must have, by Pringsheim's Theorem (*cf.* [8], Thm IV.6), a real positive singularity, ρ, smaller than or equal to 1. In all that follows we will assume that there is no other singularity with that norm, which is the case of all generating functions dealt with in this paper, as we will see. At this singularity, ρ, two cases may occur:

Case I: $\lim_{z \to \rho} w(z) = a$, where a is a positive real number.
Case II: $\lim_{z \to \rho} w(z) = +\infty$.

In the first case the curve defined by G has a shape similar to the one depicted in Fig. 1, on the left, and

$$\frac{\partial G}{\partial w}(\rho, a) = 0. \tag{6}$$
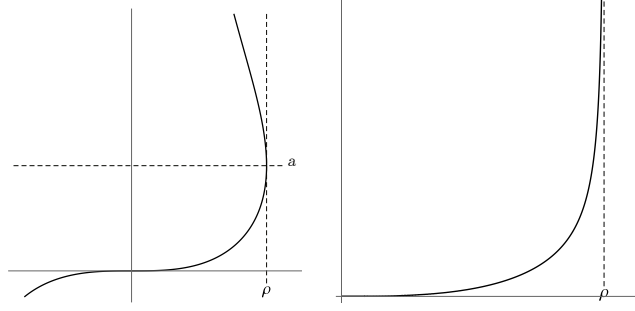
Fig. 1. Generic shape of $G(z, w)$ near its dominant singularity (cases I and II).

This, together with the fact that $G(\rho, a) = 0$, shows that ρ is a root of the resultant, $\mathrm{res}_w(G(z, w), \frac{\partial G}{\partial w}(z, w))$, of $G(z, w)$ and $\frac{\partial G}{\partial w}(z, w)$ with respect to w (*cf.* [13, p. 204]). With the help of a numerical solver and drawing the relevant part of the algebraic curve $G(z, w) = 0$, one can, by an elimination process, find out the minimum polynomial, in $\mathbb{Q}[z]$, of ρ. We will denote this polynomial by $m(z)$. Using now the $\mathrm{res}_z(G(z, w), \frac{\partial}{\partial w} G(z, w))$ one can get, in a similar fashion, an irreducible polynomial that has a as a root.

In Case II, the irreducible polynomial for ρ is a factor of the leading coefficient of $G(z, w)$, seen as a polynomial in w (*cf* [10], Th. 12.2.1).

In Case I, after making the change of variable $s = 1 - z/\rho$, one knows that $w = w(s)$ has a Puiseux series expansion at the singularity $s = 0$, i.e., there exists a slit neighbourhood of that point in which $w(s)$ has a representation as a power series with fractional powers (*cf.* [10], Chap. 12). In particular, w must have the form

$$w(s) = a - g(s)s^\alpha, \tag{7}$$

for some $\alpha \in \mathbb{Q}^+$, the first positive exponent of that expansion, and where $g(s)$ is such that $g(s) = b + h(s)s^\beta$, $h(0) \neq 0$, $\beta \in \mathbb{Q}^+$, and $b \in \mathbb{R}^\star$.

The value of α can be obtained by looking at the Taylor expansion of $G(z, w)$ at (ρ, a),

$$G(z, w) = \sum_{i,j \geq 0} \frac{1}{i!j!} \frac{\partial^{i+j} G}{\partial z^i w^j} \bigg|_{\substack{z=\rho \\ w=a}} (z - \rho)^i (w - a)^j.$$

Noticing that $z = \rho - \rho s$, and using Equation (7), one has

$$G(\rho - \rho s, a - g(s)s^\alpha) = \sum_{i,j \geq 0} \frac{(-1)^{i+j}}{i!j!} \frac{\partial^{i+j} G}{\partial z^i w^j} \bigg|_{\substack{z=\rho \\ w=a}} \rho^i g(s)^j s^{i+j\alpha}. \tag{8}$$

Using that $G(z, w(z)) = 0$, $G(\rho, a) = 0$, and (6), and dividing it through by s^α, one gets

$$0 = \sum_{\substack{i,j \geq 0 \\ (i,j) \notin \{(0,0),(0,1)\}}} \frac{(-1)^{i+j}}{i!j!} \frac{\partial^{i+j} G}{\partial z^i w^j} \bigg|_{\substack{z=\rho \\ w=a}} \rho^i g(s)^j s^{i+(j-1)\alpha}. \tag{9}$$

One can now compute $p_{ij}(z) = \text{res}_w \left(G(z, w), \dfrac{\partial^{i+j} G}{\partial z^i w^j} \right)$, and $\gcd(p_{ij}(z), m(z))$ to see which derivatives are non-zero at ρ. Then, one can use the Newton's polygon technique to find α [17, 18, 9]. The points of Newton polygon that lead to the value of α correspond to the terms of (9) with the lowest exponent, that must cancel out together. This conduces, after setting $s = 0$, to a polynomial equation for the value b defined in the sentence containing (7). One then uses this value in Theorem 10 to get the desired asymptotic approximation. In conclusion, for the case where $\lim_{z \to \rho} w(z) = a$, one has

$$[z^n]w(z) \sim \frac{-b}{\Gamma(-\alpha)} \rho^{-n} n^{-\alpha-1}. \tag{10}$$

In Case II, the one where $\lim_{z \to \rho} w(z) = +\infty$, making $v = \frac{1}{w}$ one concludes as above that $v = cs^\alpha - g(s)s^{\alpha+\beta}$, for some $0 < \alpha < 1$, $\beta > 0$, and for some Puiseux series $g(s)$, with non-negative exponents. Denoting by m the degree of G relative to w, the polynomial satisfied by v is then

$$H(z, v) = v^m G\left(z, \frac{1}{v} \right), \tag{11}$$

which is the reciprocal polynomial of $G(z, w)$ with respect to the variable w. In this case the equation that corresponds to equation (8) is:

$$H(\rho - \rho s, cs^\alpha - g(s)s^{\alpha+\beta}) = \sum_{i,j \geq 0} \frac{(-1)^i}{i!j!} \frac{\partial^{i+j} H}{\partial z^i w^j} \bigg|_{\substack{z=\rho \\ w=0}} \rho^i (c - g(s)^\beta)^j s^{i+j\alpha}. \tag{12}$$

Using the same procedure as above, one computes ρ, and then the value of c. Since

$$w = \frac{1}{cs^\alpha - g(s)s^{\alpha+\beta}} = \frac{1}{c}s^{-\alpha} \frac{1}{1 - \frac{g(s)}{c}s^\beta}$$

$$= \frac{1}{c}s^{-\alpha}\left(1 + \frac{g(s)}{c}s^\beta + \frac{g(s)^2}{c^2}s^{2\beta} + \cdots \right),$$

one sees, using again Theorem 10, that

$$[z^n]w(z) \sim \frac{1}{c\,\Gamma(\alpha)} \rho^{-n} n^{\alpha-1}. \tag{13}$$

Summing up, we have the following.

Theorem 11. *With the notations and in the conditions above described, one has*

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

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

5   Average Descriptional Complexity Results

Using the framework just described, we obtain asymptotic estimates for an upper bound of the average state complexity of partial derivative transducer for $2D$ expressions of size $n \geq 0$. Those estimates depend on the size of the alphabets Σ and Δ, which we assume both to be equal to some integer $k > 0$. Moreover we denote by RE_k the set of $1D$ expressions over an alphabet of size k.

5.1   Average State Complexity of $\mathcal{T}_{\mathrm{PD}}$ for 2D-RE

The generating function $G_k(z)$ associated with $\mathbf{g} \in$ 2D-RE is the following[5], where $R_k(z)$ is the generating function of regular expressions $\mathbf{r} \in \mathrm{RE}_k$ [4].

$$G_k(z) = zR_k(z)^2 + zG_k(z) + 2zG_k(z)^2, \tag{14}$$
$$R_k(z) = (k+1)z + zR_k(z) + 2zR_k(z)^2. \tag{15}$$

Considering Proposition 2, let $p(\mathbf{r})$ be the size of the support of an expression $\mathbf{r} \in \mathrm{RE}_k$ which is defined by $p(\varepsilon) = 0$, $p(\sigma) = 1$, $p(\mathbf{s}+\mathbf{s}') = p(\mathbf{s}\cdot\mathbf{s}') = p(\mathbf{s})+p(\mathbf{s}')$, and $p(\mathbf{s}^\star) = p(\mathbf{s})$. An upper bound for the size of the support $\pi(\mathbf{g})$, $q(\mathbf{g})$, is defined by $q(\mathbf{r}/\mathbf{r}') = p(\mathbf{r})p(\mathbf{r}') + p(\mathbf{r}) + p(\mathbf{r}')$, $q(\mathbf{g}+\mathbf{g}') = q(\mathbf{g}\cdot\mathbf{g}') = q(\mathbf{g}) + q(\mathbf{g}')$, and $q(\mathbf{g}^\star) = q(\mathbf{g})$. Thus, the generating function $Q_k(z) = \sum_{\mathbf{g}} q(\mathbf{g})z^{|\mathbf{g}|}$ for $\pi(\mathbf{g})$ satisfies the following equation,

$$Q_k(z) = zQ_k(z) + 4zQ_k(z)G_k(z) + 2zP_k(z)R_k(z) + zP_k(z)^2, \tag{16}$$

where $P_k(z)$ is the generating function for the support of regular expressions in RE_k, which satisfy

$$P_k(z) = kz + zP_k(z) + 4zR_k(z)P_k(z). \tag{17}$$

From equations (15), (17), (14) and (16), using Gröbner basis, one obtains algebraic equations for $G_k(z)$ and $Q_k(z)$:

$$\mathcal{C}_G(z,w) = 16z^3w^4 + 16(z^3 - z^2)w^3 - g_2(z)w^2 + g_1(z)w + (1+k)^2z^3 = 0, \tag{18}$$

where $g_2(z) = 2z((1+4k)z^2 + 6z - 3)$ and $g_1(z) = (1-z)((3+4k)z^2 + 2z - 1)$ and

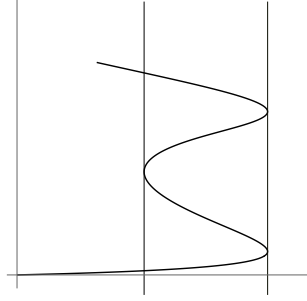$$\mathcal{C}_Q(z,w) = p(z)^4 q_4(z)w^4 - k^2z^2p(z)^2q_2(z)w^2 + k^4z^8q_0(z)^2 = 0, \tag{19}$$

where

$$p(z) = (8k+7)z^2 + 2z - 1$$
$$q_4(z) = (16k^2 + 40k + 23)z^4 - 4(4k+3)z^3 + (8k+2)z^2 + 4z - 1$$
$$q_2(z) = (200k^3 + 544k^2 + 474k + 133)z^6 - (48k^2 + 24k - 10)z^5 +$$
$$\qquad (24k^2 - 44k - 41)z^4 + 28(2k+1)z^3 + (3 - 14k)z^2 - 6z + 1$$
$$q_0(z) = (25k^2 + 37k + 14)z^2 + (6k+4)z - (3k+2).$$

For $G_k(z)$, we conclude to be in Case I. The irreducible polynomial that implicitly defines the singularity ρ_k of $G_k(z)$ is, computed using the resultant $\mathrm{res}_w(\mathcal{C}_G(z,w), \frac{\partial \mathcal{C}_G}{\partial w}(z,w))$. In this case we obtain two candidates for the minimal

---

[5] I.e. $[z^n]G_k(z)$ gives the number of expressions $\mathbf{g}$ of size n.

polynomial $m_G(z)$ of the singularity ρ_k, each one having only one root in $]0, 1[$. Using a computer algebra system, one can show that those roots are only equal for $k = -1$. This implies, by continuity (in k), that they always keep their relative position, for all $k > -1$. Now, $\mathrm{res}_z(\mathcal{C}_G(z, w), \frac{\partial \mathcal{C}_G}{\partial w}(z, w))$ factors into three irreducible polynomials, one of which has a_k as a root. These three polynomials have, among them, four positive roots, which a computer algebra system can find, as a function of k. Then, one can check which pairs (ρ'_k, a'_k), where ρ'_k is a candidate for ρ_k, and a'_k a can-

Fig. 2. Possible values for (ρ_k, a_k).

didate for a_k, belong to the curve $\mathcal{C}_G$, and their relative location. By a simple topological argument, one then can conclude that $m_G(z) = (8k + 7)z^2 + 2z - 1$, $\rho_k = \frac{1}{1+\sqrt{8k+8}}$, and $a_k = \frac{\sqrt{2}-1}{2}\sqrt{k+1}$. One then checks that $\frac{\partial \mathcal{C}_G}{\partial z}(\rho_k, a_k)$ and $\frac{\partial^2 \mathcal{C}_G}{\partial w^2}(\rho_k, a_k)$ are both non-zero, for all k, which entails that $\alpha = \frac{1}{2}$. The value for b_k can then be computed, and $b_k \sim \sqrt{\frac{k}{2}}$. As for $Q_k(z)$, one sees that Case II applies, and that the minimal polynomial is either $p(z)$ or $q_4(z)$. It turns out that each of these polynomials has exactly one positive real root, ρ_k and ζ_k. One can then check that these roots coincide only for $k = -1$, and so that one of them is always bigger than the other for all positive values of k, namely ρ_k. One then can check that the curve $\mathcal{C}_Q$ crosses the vertical line $z = \zeta_k$ exactly once above the z-axis, which makes clear that the singularity for $Q_k(z)$ is ρ_k, thus the same as for $G_k(z)$. In this case, the Newton polygon analysis shows that $\alpha = 1$ and that the polynomial satisfied by c, as explained after (9), and noticing that here we make use of inversion explained in (11), is given by

$$\left.\frac{\partial^4 H}{\partial v^4}\right|_{\substack{z=\rho \\ v=0}} c^4 + 6\left.\frac{\partial^4 H}{\partial z^2 v^2}\right|_{\substack{z=\rho \\ v=0}} \rho^2\, c^2 + \left.\frac{\partial^4 H}{\partial z^4}\right|_{\substack{z=\rho \\ v=0}} \rho^4 = 0.$$

This is a quadratic equation in c^2, whose discriminant can be seen to be zero. One gets

$$c_k^2 = -3\rho_k^2 \left(\left.\frac{\partial^4 H}{\partial z^2 \partial v^2}\right|_{\substack{z=\rho \\ v=0}}\right) \Big/ \left(\left.\frac{\partial^4 H}{\partial v^4}\right|_{\substack{z=\rho \\ v=0}}\right). \tag{20}$$

From all this, it follows that

Theorem 12. *With the notations above introduced, the ratio of the total number of states in the partial derivative transducer $\mathcal{T}_{\mathrm{PD}}(\boldsymbol{g})$ of expressions of size n to the total number of expressions of the same size is given by*

$$\frac{[z^n]\, Q_k(z)}{[z^n]\, G_k(z)} \sim \frac{-\Gamma(-\frac{1}{2})}{b_k c_k}\, n^{\frac{3}{2}}, \text{ for all } k, \text{ and } \lim_{k \to \infty} \frac{-\Gamma(-\frac{1}{2})}{b_k c_k} = \frac{\sqrt{\pi}}{8\sqrt{2}}.$$

5.2   Average State Complexity of $\mathcal{T}_{\mathrm{PD}}$ for pairs of *REs*

If we consider only $2D$-expressions of the form $\mathbf{r}/\mathbf{r}'$, the generating function for these expressions is $G'_k(z) = 2zR_k(z)$ and for the support π is, following Proposition 7, $Q'_k(z) = 2zP_k(z)R_k(z) + zP_k^2(z)$. From these, one can deduce the following algebraic equations for $G'_k(z)$ and $Q'_k(z)$:

$$\mathcal{C}_{G'}(z, w) = w^2 + (z - 1)w + 2(k + 1)z^2 = 0, \tag{21}$$

and

$$\mathcal{C}_{Q'}(z, w) = p(z)^2 w^2 + kzg'_1(z)w + k^2 z^4 g'_0(z) = 0, \tag{22}$$

where $p(z)$ is as above, and

$$g'_1(z) = (80k^2 + 126k + 49)z^4 + 4(9k + 7)z^3 - 2(9k + 5)z^2 - 4z + 1,$$
$$g'_0(z) = (25k^2 + 37k + 14)z^2 + (6k + 4)z - 3k - 2.$$

Let us first deal with $G'_k(z)$. We easily conclude that we are in Case I. The irreducible polynomial that implicitly defines the singularity ρ_k of $G'_k(z)$ is computed using $\mathrm{res}_w(\mathcal{C}_{G'}(z, w), \frac{\partial \mathcal{C}_{G'}}{\partial w}(z, w))$. In this case we obtain a single candidate for the minimal polynomial, $m_{G'}(z)$, of the singularity, ρ_k, namely

$$m_{G'}(z) = (8k + 7)z^2 + 2z - 1,$$

and thus $\rho_k = \frac{1}{1 + \sqrt{8k + 8}}$. One has

$$\mathrm{res}_w(\mathcal{C}_{G'}(z, w), \frac{\partial \mathcal{C}_{G'}}{\partial w}(z, w)) = (7 + 8k)w^2 - 8(1 + k)w + 2(1 + k),$$

from which one gets $a_k = \frac{4(1+k) - \sqrt{2(1+k)}}{7 + 8k}$, where $a_k = G'_k(\rho_k)$.

Using now the Newton's polygon method, one gets that $\alpha = \frac{1}{2}$, and

$$b_k = \sqrt{\frac{2\rho_k \frac{\partial \mathcal{C}_{G'}}{\partial z}(\rho_k, a_k)}{\frac{\partial^2 \mathcal{C}_{G'}}{\partial w^2}(\rho_k, a_k)}} \sim \frac{1}{\sqrt{2}}.$$

As for Q'_k, one sees that one is in Case II, and that the dominant singularity is the same as for G'_k. Using the methods expounded above, one gets that $\alpha = 1$, and that c_k is a zero of the equation

$$\frac{\partial^2 H}{\partial v^2}(\rho_k, 0)c_k^2 - 2\rho_k \frac{\partial^2 H}{\partial z \partial v}(\rho_k, 0)c_k + \rho_k^2 \frac{\partial^2 H}{\partial z^2}(\rho_k, 0) = 0,$$

where $H(z, v) = v^2 G_{Q'}(z, \frac{1}{v})$. It turns out that this equation has a single solution, namely $c_k = \frac{4}{k^2}\left(8 + 8k + (9 + 8k)\sqrt{2 + 2k}\right) \sim 32\sqrt{\frac{2}{k}}$. Therefore, in this case an upper bound of the average state complexity of $\mathcal{T}_{\mathrm{PD}}(\mathbf{r}/\mathbf{r}')$ is,

Theorem 13. *With the notations above introduced, one has*

$$\frac{[z^n] Q'_k(z)}{[z^n] G'_k(z)} \sim \frac{-\Gamma(-\frac{1}{2})}{b_k c_k} n^{\frac{3}{2}} \sim \frac{\sqrt{\pi}}{16}\sqrt{k}\, n^{\frac{3}{2}}.$$

6   Conclusions

We defined partial derivative transducers for $2D$ regular expressions over pairs of $1D$ regular expressions. For studying the average state complexity, and given the intricacy of the resulting generating functions, we refine known methods In Section 5, we conclude that for $2D$ expressions of size n, both general and restricted, asymptotically and on average, the state complexity of the partial derivative transducers is bounded from above by $O(n^{\frac{3}{2}})$. For ordinary $1D$ regular expressions, the number of letters in an expression is, asymptotically and on average, $\frac{1}{2}n$ [16, 3]. The same holds for general $2D$ expressions.

References

1. Antimirov, V.M.: 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.: On the average complexity of partial derivative automata for semi-extended expressions. J. Autom. Lang. Comb. **22**(1–3), 5–28 (2017)
3. Broda, S., Machiavelo, A., Moreira, N., Reis, R.: On the average state complexity of partial derivative automata: an analytic combinatorics approach. Int. J. Found. Comput. Sci. **22**(7), 1593–1606 (2011)
4. Broda, S., Machiavelo, A., Moreira, N., Reis, R.: A hitchhiker's guide to descriptional complexity through analytic combinatorics. Theoret. Comput. Sci. **528**, 85–100 (2014)
5. Broda, S., Machiavelo, A., Moreira, N., Reis, R.: Automata for regular expressions with shuffle. Inf. Comput. **259**(2), 162–173 (2018)
6. Broda, S., Machiavelo, A., Moreira, N., Reis, R.: On average behaviour of regular expressions in strong star normal form. Int. J. Found. Comput. Sci. **30**(6–7), 899–920 (2019)
7. Demaille, A.: Derived-term automata of multitape expressions with composition. Scien. Annals of Comput. Science **27**(2), 137–176 (2017)
8. Flajolet, P., R.Sedgewick: Analytic Combinatorics. CUP (2008)
9. Ghys, É.: A singular mathematical promenade. ENS Éditions (2017)
10. Hille, E.: Analytic Function Theory, vol. 2. Blaisdell Publishing Company (1962)
11. Konstantinidis, S., Moreira, N., Pires, J., Reis, R.: Partial derivatives of regular expressions over alphabet-invariant and user-defined labels. In: Hospodár, M., Jirásková, G. (eds.) Proc. 24th CIAA. LNCS, vol. 11601, pp. 184–196 (2019)
12. Konstantinidis, S., Moreira, N., Reis, R., Young, J.: Regular expressions and transducers over alphabet-invariant and user-defined labels. In: Câmpeanu, C. (ed.) Proc. 23rd CIAA. LNCS, vol. 10977, pp. 1–27. Springer (2018)
13. Lang, S.: Algebra, Grad. Texts Math., vol. 211. Springer, 3rd edn. (2001)
14. Lombardy, S., Sakarovitch, J.: Derivatives of rational expressions with multiplicity. Theor. Comput. Sci. **332**(1-3), 141–177 (2005)
15. Mirkin, B.G.: An algorithm for constructing a base in a language of regular expressions. Eng. Cybernetics **5**, 51—57 (1966)
16. Nicaud, C.: On the average size of Glushkov's automata. In: Dediu, A., Ionescu, A.M., Vide, C.M. (eds.) Proc. 3rd LATA. LNCS, vol. 5457, pp. 626–637 (2009)
17. Walker, R.J.: Algebraic Curves. Princeton University Press (1950)
18. Wall, C.T.C.: Singular Points of Plane Curves. No. 63 in London Mathematical Society Student Texts, Cambridge University Press (2004)