On the Average Complexity of Strong Star Normal Form

Sabine Broda, António Machiavelo, Nelma Moreira, and Rogério Reis(✉)

CMUP & DM-DCC, Faculdade de Ciências da Universidade do Porto,
Rua do Campo Alegre, 4169-007 Porto, Portugal
{sbb,nam,rvr}@dcc.fc.up.pt, ajmachia@fc.up.pt

Abstract. For regular expressions in (strong) star normal form a large set of efficient algorithms is known, from conversions into finite automata to characterisations of unambiguity. In this paper we study the average complexity of this class of expressions using analytic combinatorics. As it is not always feasible to obtain explicit expressions for the generating functions involved, here we show how to get the required information for the asymptotic estimates with an indirect use of the existence of Puiseux expansions at singularities. We study, asymptotically and on average, the alphabetic size, the size of the ε-follow automaton and the ratio of these expressions to standard regular expressions.

1 Introduction

A regular expression α is in strong star normal form (ssnf) if for any subexpression of the form $\beta^\star$ or $\beta + \varepsilon$ the language represented by β does not include the empty word, ε. The star normal form was introduced by Brüggemann-Klein [5] as a step to improve the construction of the position automaton from a regular expression from cubic to quadratic time. Transforming a regular expression into this normal form can be achieved in linear time, and moreover the position automaton resulting from that normal form coincides with the one of the original expression. In the same paper, the star normal form was also used to characterize certain types of unambiguous expressions. The position automaton construction [9] is a basic conversion between regular expressions and ε-free nondeterministic finite automata (NFA), and several other constructions are known to be its quotients. This is the case for the partial derivative automaton [1,7] and the follow automaton [14]. Champarnaud et al. [6] showed that if a regular expression is in star normal form and is normalised modulo some regular expression equivalences, the partial derivative automaton is a quotient of the follow automaton. Many conversions from regular expressions to equivalent

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.

G. Pighizzini and C. Câmpeanu (Eds.): DCFS 2017, LNCS 10316, pp. 77–88, 2017.
DOI: 10.1007/978-3-319-60252-3_6

NFAs consider automata with transitions labelled by the empty word (ε-NFA). Although the most used of these conversions is the Thompson construction (implemented in many UNIX-like string search commands) [18], an older and more thrifty construction in the use of ε-transitions was presented by Ott and Feinstein in 1961 [16]. An improved version of this construction was redefined by Ilie and Yu, and called the ε-follow automaton. Gulan, Fernau and Gruber [10–12] studied the optimal (worst-case) size for all known constructions from regular expressions to ε-NFAs. It turns out that the optimal construction corresponds to the conversion of a regular expression in strong star normal form into an ε-follow automaton.

All this motivated us to study the average-case complexity of regular expressions in strong star normal form, as well as their conversions to NFAs. In previous work, we studied the asymptotic average complexity for some of the above mentioned conversions from regular expressions using the framework of analytic combinatorics [2–4], which relates the enumeration of combinatorial objects to the algebraic and complex analytic properties of generating functions. In particular, generating functions can be seen as complex analytic functions, and the study of their behaviour around their dominant singularities gives access to the asymptotic form of their coefficients. Starting with an unambiguous grammar for the set of regular expressions over a given alphabet, and a non-negative measure, the symbolic method allows to obtain a generating function associated with the sequence of the (finite) number of expressions of measure n. Multivariate generating functions can be used to analyse different measures apart from the size of combinatorial objects, e.g. the number of states of the automaton resulting from a given conversion method applied to a regular expression of given size, and thus allow to obtain estimates for the average values of those measures.

While in previous work we were able to get explicit expressions for the generating functions involved, here that would be unmanageable. Using the existence of a Puiseux expansion at a singularity, we show how to get the required information for the asymptotic estimates from an algebraic equation satisfied by the generating function, without actually computing that expansion. We note that the technique here presented allows to find, for the combinatorial classes considered, the form of the function without knowing beforehand the explicit value of the singularity. This provides a very useful method, at least for some combinatorial classes, that circumvents some of the more cumbersome steps of the *Algebraic Coefficient Asymptotics* algorithm presented by Flajolet and Sedgewick [8, pp. 504–505], as well as the need to know *a priori* the type of the singularity.

We use this method to derive the asymptotic estimates for the number of regular expressions in ssnf of a given size, as well as a parametric function of several related measures, which can give us, in particular, the alphabetic size or the size of the ε-follow automaton, on average. In the next section, we review some basics on regular expressions and ε-NFAs. In Sect. 3, we consider the transformation into strong star normal form and give some characterisations of expressions in this form. Section 4 describes a shortcut to obtain asymptotic estimates of the coefficients of generating functions. This is used in Sect. 5 to obtain the estimates mentioned before. Some experiments corroborating those estimates are presented in Sect. 6. Conclusions are drawn in Sect. 7.

2 Regular Expressions and ε-NFAs

We consider the grammar for regular expressions proposed by Gruber and Gulan in [10,11], which has the major advantage of avoiding many redundant expressions built with the symbols ε and $\emptyset$. Given an alphabet $\Sigma = \{\sigma_1, \ldots, \sigma_k\}$ of size k, the set $\mathcal{R}_k$ of *regular expressions*, α, over Σ is defined by the following grammar,

$$\alpha := \emptyset \mid \varepsilon \mid \beta,$$

$$\beta := \sigma_1 \mid \cdots \mid \sigma_k \mid (\beta + \beta) \mid (\beta \cdot \beta) \mid \beta^{?} \mid \beta^{\star},$$

where the operator $\cdot$ (concatenation) is often omitted. The language associated with α is denoted by $\mathcal{L}(\alpha)$ and is defined as usual, with $\mathcal{L}(\beta^{?}) = \mathcal{L}(\beta) \cup \{\varepsilon\}$. It is clear that $\alpha^{?}$ is equivalent to the standard regular expression $\alpha + \varepsilon$.

For the *size* of a regular expression α, denoted by $|\alpha|$, we will consider reverse polish notation length, i.e., the number of symbols in α, not counting parentheses. The number of letters in α is denoted by $|\alpha|_{\Sigma}$, and usually called *alphabetic size*. The number of occurrences of each operator $c \in \{+, \cdot, \star, ?\}$ is denoted by $|\alpha|_c$.

A *nondeterministic finite automaton* is a tuple $\mathcal{N} = \langle Q, \Sigma, \delta, q_0, F\rangle$, where Q is a finite set of states, Σ is the alphabet, $\delta \subseteq Q \times (\Sigma \cup \{\varepsilon\}) \times Q$ is the transition relation, $q_0 \in Q$ is the initial state, and $F \subseteq Q$ is the set of final states. The *size* of an NFA $\mathcal{N}$ is $|\mathcal{N}| = |Q| + |\delta|$, the number of states $|\mathcal{N}|_Q = |Q|$, and the number of transitions $|\mathcal{N}|_{\delta} = |\delta|$. An NFA that has transitions labelled with ε is an ε-NFA. The *language* accepted by an automaton $\mathcal{N}$ is $\mathcal{L}(\mathcal{N}) = \{\ w \in \Sigma^{\star} \mid \delta(q_0, w) \cap F \neq \emptyset\ \}$, where δ is naturally extended to sets of states and words.

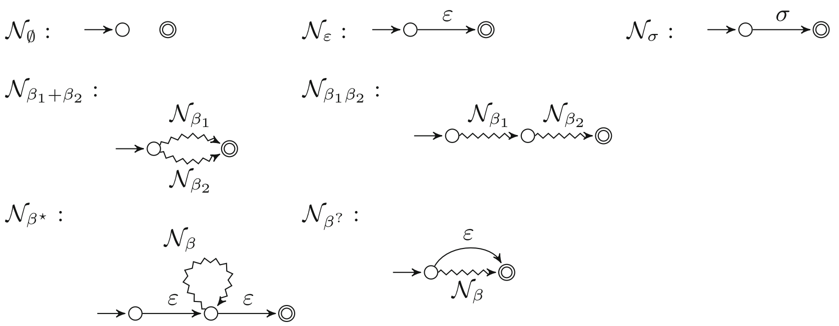

Fig. 1. The ε-follow construction, $\mathcal{A}_{\varepsilon\mathsf{f}}$.

Conversion of a regular expression into an equivalent NFA can be defined by induction on the structure of the regular expression. Let $\mathcal{N}_{\alpha}$ denote the automaton corresponding to a regular expression α. In Fig. 1 we present the construction of the ε-follow automaton, $\mathcal{A}_{\varepsilon\mathsf{f}}(\alpha)$ [14]. The size of the $\mathcal{A}_{\varepsilon\mathsf{f}}(\alpha)$ for the atomic expressions $\emptyset$, ε, and $\sigma \in \Sigma$ is 2, 3 and 3, respectively. For the remaining constructions, the size of the resulting automaton equals the sum of the sizes of its constituents plus some constant. For instance, for the operator $+$ one has $|\mathcal{N}_{\beta_1+\beta_2}|_Q = |\mathcal{N}_{\beta_1}|_Q + |\mathcal{N}_{\beta_2}|_Q - 2$, $|\mathcal{N}_{\beta_1+\beta_2}|_\delta = |\mathcal{N}_{\beta_1}|_\delta + |\mathcal{N}_{\beta_2}|_\delta$, and thus $|\mathcal{N}_{\beta_1+\beta_2}| = |\mathcal{N}_{\beta_1}| + |\mathcal{N}_{\beta_2}| - 2$. This can be generalised by considering constants $(c_\emptyset, c_\varepsilon, c_\sigma, c_+, c_\bullet, c_\star, c_?)$ that define functions that can be used to compute several interesting measures. For example, using $(2, 2, 2, -2, -1, 1, 0)$ one gets the number of states; the number of transitions are computed using $(0, 1, 1, 0, 0, 2, 1)$, and the combined size corresponds to $(2, 3, 3, -2, -1, 3, 1)$.

We note that the worst-case complexity for this conversion can be reached for expressions with only one letter and $n-1$ stars. For such an expression of size n, the corresponding $\mathcal{A}_{\varepsilon\mathsf{f}}$ automaton has size $3n$.

3 Strong Star Normal Form

A regular expression α is in *star normal form* if for any subexpression of the form $\beta^\star$, $\varepsilon \notin \mathcal{L}(\beta)$ [5]. The original notion of star normal form makes use of two operators on regular expressions. Gulan and Gruber simplified that definition and adapted it to forbid that subexpressions of the form $\beta^?$ could have $\varepsilon \in \mathcal{L}(\beta)$. The resulting form was called *strong star normal form.*

Definition 1. *The operators* $\circ$ *and* $\bullet$ *are inductively defined as follows. Let* $\varepsilon^\circ = \emptyset^\circ = \emptyset$, $\sigma^\circ = \sigma$ *for* $\sigma \in \Sigma$, $(\beta_1 + \beta_2)^\circ = \beta_1^\circ + \beta_2^\circ$, $\beta^{?\circ} = \beta^\circ$, $\beta^{\star\circ} = \beta^\circ$*; finally* $(\beta_1\beta_2)^\circ = \beta_1^\circ + \beta_2^\circ$ *if* $\varepsilon \in \mathcal{L}(\beta_1\beta_2)$ *and* $(\beta_1\beta_2)^\circ = \beta_1\beta_2$*, otherwise.*

Let $\emptyset^\bullet = \emptyset$, $\varepsilon^\bullet = \varepsilon$, $\sigma^\bullet = \sigma$ *for* $\sigma \in \Sigma$, $(\beta_1 + \beta_2)^\bullet = \beta_1^\bullet + \beta_2^\bullet$, $(\beta_1\beta_2)^\bullet = \beta_1^\bullet\beta_2^\bullet$, $\beta^{\star\bullet} = \beta^{\circ\bullet\star}$*; finally* $\beta^{?\bullet} = \beta^\bullet$ *if* $\varepsilon \in \mathcal{L}(\beta)$*, and* $\beta^{?\bullet} = (\beta^\bullet)^?$*, otherwise.*

The expression $\alpha^\bullet$ *is the* strong star normal form *(*ssnf*) of* α*.*

For a regular expression α, $\mathcal{L}(\alpha^\bullet) = \mathcal{L}(\alpha)$ and $|\alpha^\bullet| \leq |\alpha|$. The following theorem characterizes the regular expressions in strong star normal form.

Theorem 1 [11, Theorem 3.2.8]**.** *A regular expression* α *is in strong star normal form, i.e.* $\alpha = \alpha^\bullet$*, if and only if for every subexpression* $\beta^\star$ *or* $\beta^?$ *of* α*, one has* $\varepsilon \notin \mathcal{L}(\beta)$*.*

Using this theorem it is possible to write a context-free grammar for regular expressions in ssnf, i.e., in which every subexpression of the form $\alpha^\star$ or $\alpha^?$, satisfies $\varepsilon \notin \mathcal{L}(\alpha)$. The set $\mathcal{S}_k$ of *regular expressions in ssnf* over Σ is defined by:

$$\begin{aligned} \alpha &:= \varepsilon \mid \emptyset \mid \alpha_\varepsilon \mid \alpha_{\overline{\varepsilon}} \\ \alpha_\varepsilon &:= \alpha_\varepsilon\alpha_\varepsilon \mid \alpha_\varepsilon + \alpha_{\overline{\varepsilon}} \mid \alpha_{\overline{\varepsilon}} + \alpha_\varepsilon \mid \alpha_\varepsilon + \alpha_\varepsilon \mid \alpha_{\overline{\varepsilon}}^\star \mid \alpha_{\overline{\varepsilon}}^? \\ \alpha_{\overline{\varepsilon}} &:= \sigma_1 \mid \cdots \mid \sigma_k \mid \alpha_{\overline{\varepsilon}}\alpha_{\overline{\varepsilon}} \mid \alpha_{\overline{\varepsilon}}\alpha_\varepsilon \mid \alpha_\varepsilon\alpha_{\overline{\varepsilon}} \mid \alpha_{\overline{\varepsilon}} + \alpha_{\overline{\varepsilon}}, \end{aligned} \tag{1}$$

where α_ε are regular expressions whose language includes ε, while for $\alpha_{\overline{\varepsilon}}$, $\varepsilon \notin \mathcal{L}(\alpha_{\overline{\varepsilon}})$. The following theorem summarizes the results by Gruber and Gulan [10, Theorems 4 and 6] (see also Gulan [11]).

Theorem 2. *Let α be in* ssnf *of size n and alphabetic size m. Then, $\mathcal{A}_{\varepsilon\mathsf{f}}(\alpha)$ has size at most* $\min(\frac{22}{15}(n+1)+1, \frac{22}{5}m+1)$.

4 Asymptotic Average Complexity

Let $A(z) = \sum_n a_n z^n$ be the generating function associated with some combinatorial class $\mathcal{A}$ (*cf.* [8]). Given some measure of the objects of the class, the coefficient a_n represents the sum of the values of this measure for all objects of size n. We will use the notation $[z^n]A(z)$ 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 [3,8]:

1. if $A(z) = a - b\sqrt{1-z/\rho} + o\left(\sqrt{1-z/\rho}\right)$, with $a, b \in \mathbb{R}$, $b \neq 0$, then
$$[z^n]A(z) \sim \frac{b}{2\sqrt{\pi}}\,\rho^{-n}n^{-3/2}; \tag{2}$$

2. if $A(z) = \frac{c}{\sqrt{1-z/\rho}} + o\left(\frac{1}{\sqrt{1-z/\rho}}\right)$, with $c \in \mathbb{R}^*$, then
$$[z^n]A(z) \sim \frac{c}{\sqrt{\pi}}\,\rho^{-n}n^{-1/2}. \tag{3}$$

Applying this result for the generating function $R_k(z)$, corresponding to the number of expressions in $\mathcal{R}_k$ of size n, the following asymptotic values were obtained in Broda *et al.* [3]:

$$[z^n]R_k(z) \sim \frac{\sqrt[4]{2k}\sqrt{\rho_k}}{4\sqrt{\pi}}\rho_k^{-(n+1)}(n+1)^{-3/2}, \text{ with } \rho_k = \frac{1}{2(\sqrt{2k}+1)}. \tag{4}$$

In the same paper, the average size of the ε-follow automata construction was studied, and it was shown that, as the alphabet grows, the size of $\mathcal{A}_{\varepsilon\mathsf{f}}$ approaches $0.75n$, asymptotically and on average.

Let us now give a generic description of the method used for the combinatorial classes that show up within the present paper. From a grammar one obtains, by the symbolic method expounded in [8], a set of polynomial equations involving the generating function of whose coefficients we want to have an asymptotic 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 [8, Theorem IV.6], a real positive singularity, ρ, smaller than 1. At this singularity two cases may occur: either $\lim_{z\to\rho} w(z) = a$, a positive real number, or $\lim_{z\to\rho} w(z) = +\infty$.

In the first case, 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 [13, Chap. 12]. In particular, w must have the form

$$w(s) = a - g(s)s^{\alpha}, \tag{5}$$

for some $a \in \mathbb{R}$, $\alpha \in \mathbb{Q}^+$, the first positive exponent of that expansion, and $g(s)$ such that $g(s) = b + h(s)s^{\beta}$, $h(0) \neq 0$, $\beta \in \mathbb{Q}^+$, and $b \in \mathbb{R}^*$. We will show that, under some generic conditions that happen to be satisfied in all the cases treated below, one has $\alpha = \frac{1}{2}$ or $\alpha = -\frac{1}{2}$. One then needs to find the values of ρ and of b or c, depending on the case, to use either (2) or (3) to obtain the sought-after asymptotic estimates of the coefficients of $w(z)$.

Using Taylor expansion of $G(z, w)$ at (ρ, a),

$$\begin{aligned} G(z,w) =& G(\rho,a) + \frac{\partial G}{\partial z}(\rho,a)(z-\rho) + \frac{\partial G}{\partial w}(\rho,a)(w-a) + \\ &+ \frac{1}{2}\frac{\partial^2 G}{\partial z^2}(\rho,a)(z-\rho)^2 + \frac{1}{2}\frac{\partial^2 G}{\partial w^2}(\rho,a)(w-a)^2 + \\ &+ \frac{\partial^2 G}{\partial z\,\partial w}(\rho,a)(z-\rho)(w-a) + \cdots, \end{aligned}$$

and noticing that $G(z, w(z)) = 0$, that $G(\rho, a) = 0$, and using Eq. (5), one has,

$$\begin{aligned} 0 =& -\frac{\partial G}{\partial z}(\rho,a)\rho s - \frac{\partial G}{\partial w}(\rho,a)g(s)s^{\alpha} + \frac{1}{2}\frac{\partial^2 G}{\partial z^2}(\rho,a)\rho^2 s^2 + \\ &+ \frac{1}{2}\frac{\partial^2 G}{\partial w^2}(\rho,a)g(s)^2 s^{2\alpha} - \frac{\partial^2 G}{\partial z\,\partial w}(\rho,a)\rho g(s)s^{1+\alpha} + Q(s)s^{3\alpha}, \end{aligned} \tag{6}$$

for some function $Q(s)$, a Puiseux series with non-negative exponents.

In the case under study, the curve defined by G has a shape similar to the one depicted in Fig. 2, and therefore

$$\frac{\partial G}{\partial w}(\rho, a) = 0. \tag{7}$$

This, together with the fact that $G(\rho, a) = 0$, shows that ρ is a root of the discriminant polynomial of G with respect to variable w, which is a polynomial in z (*cf.* [15, p. 204]). In all the cases studied here, this polynomial has only one root in $]0, 1[$, a fact that allows to numerically get an approximation for the value of ρ. The minimum polynomial in $\mathbb{Q}[z]$ of ρ can be obtained by analysing the greatest common divisor of the polynomials $G(z, w)$ and $\frac{\partial}{\partial w}G(z, w)$ with respect to w: $\gcd_w(G(z, w), \frac{\partial}{\partial w}G(z, w))$. We will denote this polynomial by $m_\rho(z)$. Using now the $\gcd_z(G(z, w), \frac{\partial}{\partial w}G(z, w))$ one can get a polynomial that has a as a root. One can then numerically compute all the real roots of that polynomial, and then check which one is an approximation for the value of a by means of a numerical study of the curve $G(z, w)$.

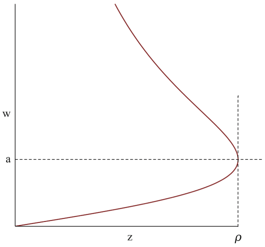

Fig. 2. Generic shape of $G(z,w)$ near its dominant singularity.

Using (7) in (6), and dividing it through by s^α, one gets

$$\begin{aligned}0 = &-\frac{\partial G}{\partial z}(\rho,a)\rho s^{1-\alpha} + \frac{1}{2}\frac{\partial^2 G}{\partial z^2}(\rho,a)\rho^2 s^{2-\alpha}\\ &+\frac{1}{2}\frac{\partial^2 G}{\partial w^2}(\rho,a)g(s)^2 s^\alpha +\\ &+\frac{\partial^2 G}{\partial z\,\partial w}(\rho,a)\rho g(s)s + Q(s)s^{2\alpha}.\end{aligned} \tag{8}$$

Now, in all cases studied in this paper, one has

$$\frac{\partial G}{\partial z}(\rho,a) \neq 0, \text{ and } \frac{\partial^2 G}{\partial w^2}(\rho,a) \neq 0. \tag{9}$$

This was checked by computing

$$p_1(z) = \gcd_w(G(z,w), \frac{\partial}{\partial z}G(z,w)), \quad p_2(z) = \gcd_w(G(z,w), \frac{\partial^2}{\partial w^2}G(z,w)),$$

$\gcd(p_1(z), m_\rho(z))$ and $\gcd(p_2(z), m_\rho(z))$, obtaining a constant depending only on k, that is non-zero for all $k \neq 54$ in all cases dealt with in this paper. The case $k = 54$ was dealt separately. Using the explicit value for ρ, the validity of (9) for this value of k was verified.

It now follows from (8), by noticing that the first and third summands have the smallest degrees in s, that they must have the same degree and cancel each other. Dividing, then, by s^α and letting $s \to 0$, one obtains

$$\alpha = \frac{1}{2}, \text{ and } b = g(0) = \sqrt{\frac{2\rho\,\frac{\partial G}{\partial z}(\rho,a)}{\frac{\partial^2 G}{\partial w^2}(\rho,a)}}.$$

In conclusion, for the case where $\lim_{z\to\rho} w(z) = a$, using (2), one has

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

For the case where $\lim_{z\to\rho} w(z) = +\infty$, making $v = 1/w$ 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. The polynomial satisfied by v is then

$$H(z,v) = v^n G\left(z, \frac{1}{v}\right), \tag{10}$$

which is the reciprocal polynomial of $G(z,w)$ with respect to the variable w. Using the same procedure as above, one computes ρ, and checking that the corresponding derivatives are non-zero, i.e.

$$\frac{\partial H}{\partial z}(\rho,0) \neq 0, \text{ and } \frac{\partial^2 H}{\partial w^2}(\rho,0) \neq 0,$$

one gets in the same way that

$$\alpha = \frac{1}{2}, \text{ and } c = \sqrt{\frac{2\rho \frac{\partial H}{\partial z}(\rho, 0)}{\frac{\partial^2 H}{\partial w^2}(\rho, 0)}}. \tag{11}$$

Since

$$\begin{aligned} w = \frac{1}{cs^{\alpha} - g(s)s^{\alpha+\beta}} &= \frac{1}{c} s^{-\alpha} \frac{1}{1 - \frac{g(s)}{c} s^{\beta}} \\ &= \frac{1}{c} s^{-\alpha} \left(1 + \frac{g(s)}{c} s^{\beta} + \frac{g(s)^2}{c^2} s^{2\beta} + \cdots \right), \end{aligned}$$

one sees, using (3), that

$$[z^n]w(z) \sim \frac{1}{c\sqrt{\pi}} \rho^{-n} n^{-1/2}. \tag{12}$$

5 Average Sizes: Concrete Results

Let $A_k(z)$ and $B_k(z)$ be the generating functions for α_{ε} and $\alpha_{\overline{\varepsilon}}$, respectively. They satisfy the following equations

$$A_k(z) = 2zA_k(z)^2 + 2zA_k(z)B_k(z) + 2zB_k(z) \tag{13}$$

$$B_k(z) = kz + 2zA_k(z)B_k(z) + 2zB_k(z)^2. \tag{14}$$

From (13) one gets

$$B_k(z) = \frac{A_k(z)(1 - 2zA_k(z))}{2z(A_k(z) + 1)},$$

and then substituting $B_k(z)$ in (14) one obtains, after clearing up denominators,

$$4z^2 A_k(z)^3 - (2kz^2 + 4z)A_k(z)^2 - (4kz^2 - 1)A_k(z) - 2kz^2 = 0,$$

i.e., $A_k(z)$ is an algebraic function that is a root of

$$G(z, w) = 4z^2 w^3 - (2kz^2 + 4z)w^2 - (4kz^2 - 1)w - 2kz^2.$$

Using now (14) to get $A_k(z)$ as a function of $B_k(z)$, and then substituting that into (13), one easily sees that $B_k(z)$ is a root of

$$H(z, w) = 4zw^3 + 2kzw^2 - kw + k^2 z.$$

Using the technique described in the previous section, one sees that $A_k(z)$ and $B_k(z)$ have the same singularity, namely the only root in the interval $]0, 1[$ of the polynomial

$$m_{\rho}(z) = z^3 + \frac{9z^2}{2k + 27} - \frac{z}{8k + 108} - \frac{1}{k(2k + 27)}. \tag{15}$$

Also one gets that $\alpha = \frac{1}{2}$, and that the values of $a_A = A_k(\rho)$ and of $a_B = B_k(\rho)$ are roots of the polynomials $8z^3 - kz^2 + 2kz - k$, and $8z^3 + 2kz^2 - k^2$, respectively. With all this, and writing $S_k(z) = A_k(z) + B_k(z)$ one then gets that

$$[z^n]S_k(z) \sim \frac{b_k}{2\sqrt{\pi}} \rho_k^{-n} n^{-3/2}, \tag{16}$$

where, for example,

$$\begin{aligned} b_2 &= 1.089338906, \quad \rho_2 = 0.1915181504 \\ b_{10} &= 2.313181803, \rho_{10} = 0.09581011247 \\ b_{50} &= 5.054983041, \rho_{50} = 0.4606805763. \end{aligned}$$

Using these results and the one mentioned in (4), the ratio of regular expressions in ssnf, $r_{(k,n)} = \frac{[z^n]S_k(n)}{[z^n]R_k(n)}$, can now be computed for any k and n. In particular, one finds that, for example, $r_{(2,1000)} = 4.427117336 \times 10^{-59}$, $r_{(10,1000)} = 2.562752010 \times 10^{-19}$, $r_{(50,1000)} = 1.517513555 \times 10^{-4}$.

5.1 Counting Letters

To obtain the asymptotic average value of several measures for regular expressions of a given size, we consider bivariate generating functions parametrized by weights of the form c_o, with $o \in \{\emptyset, \varepsilon, \sigma, +, \cdot, \star, ?\}$, associated to each regular expression element. Considering the grammar (1), let $A_k(u,z)$ and $B_k(u,z)$ be the bivariate generating functions associated to α_ε and $\alpha_{\overline{\varepsilon}}$, respectively. Then

$$\begin{aligned} A_k(u,z) &= (u^{c_\bullet} + u^{c_+}) z A_k(u,z)^2 + 2u^{c_+} z A_k(u,z) B_k(u,z) + (u^{c_?} + u^{c_\star}) z B_k(u,z), \\ B_k(u,z) &= k u^{c_\sigma} z + (u^{c_\bullet} + u^{c_+}) z B_k(u,z)^2 + 2u^{c_\bullet} z A_k(u,z) B_k(u,z). \end{aligned}$$

Note that A and B depend on the parameters $(c_\emptyset, c_\varepsilon, c_\sigma, c_+, c_\bullet, c_\star, c_?)$, but for sake of simplicity we choose to omit them. For computing the average number of letters those parameters are $(0,0,1,0,0,0,0)$, and analogously for each operator.

The generating function $L_k(z)$ for the number of letters is given by

$$L_k(z) = \left.\frac{\partial}{\partial u}\right|_{u=1} (A_k(u,z) + B_k(u,z)).$$

Setting $A = A_k(1,z), B = B_k(1,z), A_1 = \left.\frac{\partial}{\partial u}\right|_{u=1} A_k(u,z), B_1 = \left.\frac{\partial}{\partial u}\right|_{u=1} B_k(u,z)$, so that $L_k = A_1 + B_1$, one has:

$$\begin{aligned} A &= 2A^2 z + 2ABz + 2Bz, \\ B &= 2ABz + 2B^2 z + kz, \\ A_1 &= 4AA_1 z + 2AB_1 z + 2BA_1 z + 2B_1 z, \\ B_1 &= 2AB_1 z + 2BA_1 z + 4BB_1 z + kz, \\ L_k &= A_1 + B_1. \end{aligned}$$

Using Gröbner basis, as mentioned above, one gets the following polynomial for $w = L_k$:

$$G(z,w) = \left(\left(8\,k^2 + 108\,k\right)\ z^3 + 36\,kz^2 - kz - 4\right) w^3 + \\ + \left(\left(k^3 + 12\,k^2\right) z^3 + 4\,k^2 z^2 + kz\right) w - 2\,k^2 z^3 - k^2 z^2.$$

It turns out that, from this, one can deduce that the singularity for this algebraic function w has the same minimal polynomial as in (15), and so it is the same as for the number of regular expressions there considered. One then finds that, in this case, $\alpha = -\frac{1}{2}$, and that

$$[z^n]L_k(z) \sim \frac{1}{c_k\sqrt{\pi}}\rho_k^{-n} n^{-1/2}, \tag{17}$$

where, for example,

$$\begin{aligned} c_2 &= 2.725255757, & \rho_2 &= 0.1915181504,\\ c_{10} &= 1.271387537, & \rho_{10} &= 0.09581011247,\\ c_{50} &= 0.5749569245, & \rho_{50} &= 0.04606805763. \end{aligned}$$

From this one gets, for any given k, the density of letters in expressions of size n, $\ell_k = \frac{[z^n]L_k(n)}{n[z^n]S_k(n)}$, which is independent of n since the singularities of L_k and S_k are the same. In particular, one finds that, for example, $\ell_2 = 0.4172563448$, $\ell_{10} = 0.4432524170$, $\ell_{50} = 0.4657465002$.

5.2 Size of ε-Follow Automata

Considering the parameters $(2, 3, 3, -2, -1, 3, 1)$, as defined in Sect. 2, the generating function $F_k(z)$ for the size of the $\mathcal{A}_{\varepsilon\mathsf{f}}$ automaton is given by

$$F_k(z) = \left.\frac{\partial}{\partial u}\right|_{u=1} \left(A_k(u,z) + B_k(u,z)\right).$$

Using the same abbreviations as above, one has:

$$\begin{aligned} A &= 2A^2z + 2ABz + 2Bz\\ B &= 2ABz + 2B^2z + kz\\ A_1 &= -3A^2z - 4ABz + 4AA_1z + 2AA_2z + 2BA_1z + 4Bz + 2A_2z\\ A_2 &= -2ABz + 2AA_2z - 3B^2z + 2BA_1z + 4BA_2z + 3kz\\ F_k &= A_1 + A_2. \end{aligned}$$

Proceeding as above, one can verify that the singularity for $F_k(z)$ still has the same minimal polynomial as in (15), that $\alpha = -\frac{1}{2}$, and that

$$[z^n]F_k(z) \sim \frac{1}{c_k\sqrt{\pi}}\rho_k^{-n} n^{-1/2}, \tag{18}$$

where, for example,

$$\begin{aligned} c_2 &= 1.159914873, & \rho_2 &= 0.1915181504,\\ c_{10} &= 0.6237795132, & \rho_{10} &= 0.09581011247,\\ c_{50} &= 0.3187807970, & \rho_{50} &= 0.4606805763. \end{aligned}$$

For the average ratio, $f_k = \frac{[z^n]F_k(n)}{n[z^n]S_k(n)}$, between the size of the $\mathcal{A}_{\varepsilon\mathsf{f}}$ and the size of the respective regular expression (also independent of n) one has, for example, $f_2 = 0.9803566472$, $f_{10} = 0.9034371711$, $f_{50} = 0.8400260553$.

6 Experimental Results

We ran some experiments, using the FAdo package [17], to obtain average sizes of the measures studied above for small values of k and n. For the results to be statistically significant, regular expressions were uniformly random generated using a version of the grammar for $\mathcal{S}_k$ in reverse polish notation. For each size $n \in \{200, 500, 1000\}$, and alphabet size $k \in \{2, 10, 50\}$, samples of 10000 expressions were generated. This is sufficient to ensure a 95% confidence level within a 1% error margin. The results are presented in Table 1, together with the values of ℓ_k and f_k calculated in the previous section. The last column, labeled wc, presents the worst case size of $\mathcal{A}_{\varepsilon\mathsf{f}}$ as given in Theorem 2, for expressions of size n.

Table 1. Results for regular expressions in ssnf

k	$\lvert\alpha\rvert$	$\lvert\alpha\rvert_\Sigma$	$\lvert\delta_{\varepsilon\mathsf{f}}\rvert$	$\lvert Q_{\varepsilon\mathsf{f}}\rvert$	$\lvert\varepsilon\mathsf{f}\rvert$	$\frac{\lvert\alpha\rvert_\Sigma}{\lvert\alpha\rvert}$	ℓ_k	$\frac{\lvert\varepsilon\mathsf{f}\rvert}{\lvert\alpha\rvert}$	f_k	wc
2	200	83.86	112.20	52.86	165.06	0.42	**0.417**	0.83	**0.98**	1.479
	500	208.99	279.97	129.74	409.71	0.42		0.82		1.472
	1000	417.70	559.04	257.85	816.89	0.42		0.82		1.469
10	200	89.13	111.98	51.80	163.78	0.45	**0.443**	0.82	**0.90**	1.479
	500	222.09	279.11	126.91	406.02	0.44		0.81		1.472
	1000	443.77	557.72	252.30	810.02	0.44		0.81		1.469
50	200	93.63	108.53	51.29	159.82	0.47	**0.466**	0.80	**0.84**	1.479
	500	233.34	270.66	125.80	396.46	0.47		0.79		1.472
	1000	466.20	540.84	249.94	790.78	0.47		0.79		1.469

7 Conclusions

The average complexity results obtained for expressions in ssnf are only slightly smaller than the ones obtained for general regular expressions. Indeed, for the size of $\mathcal{A}_{\varepsilon\mathsf{f}}$, and the same values of k, the asymptotic values obtained in [3], were $f_2 = 1.2$, $f_{10} = 1$, and $f_{50} = 0.9$. In that study, we got an explicit expression, depending on k, for the asymptotic size of $\mathcal{A}_{\varepsilon\mathsf{f}}$, allowing us to compute its limit of 0.75 as k goes to ∞. Here we were not able to obtain such an expression, but we conjecture that the limit is the same. This would mean that the average size is half the worst-case one. This is corroborated by the experimental results. Furthermore, the ratio between the number of ssnf expressions and the number of general expressions, of a certain size, tends to zero.

References

1. Antimirov, V.M.: Partial derivatives of regular expressions and finite automaton constructions. Theor. Comput. Sci. **155**(2), 291–319 (1996)
2. 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)
3. Broda, S., Machiavelo, A., Moreira, N., Reis, R.: A hitchhiker's guide to descriptional complexity through analytic combinatorics. Theor. Comput. Sci. **528**, 85–100 (2014)
4. Broda, S., Machiavelo, A., Moreira, N., Reis, R.: Average size of automata constructions from regular expressions. BEATCS **116**, 167–192 (2015)
5. Brüggemann-Klein, A.: Regular expressions into finite automata. Theor. Comput. Sci. **48**, 197–213 (1993)
6. Champarnaud, J.M., Ouardi, F., Ziadi, D.: Normalized expressions and finite automata. Int. J. Algebra Comput. **17**(1), 141–154 (2007)
7. Champarnaud, J.M., Ziadi, D.: Canonical derivatives, partial derivatives and finite automaton constructions. Theor. Comput. Sci. **289**, 137–163 (2002)
8. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. CUP, Cambridge (2008)
9. Glushkov, V.M.: The abstract theory of automata. Russ. Math. Surv. **16**(5), 1–53 (1961)
10. Gruber, H., Gulan, S.: Simplifying regular expressions. In: Dediu, A.-H., Fernau, H., Martín-Vide, C. (eds.) LATA 2010. LNCS, vol. 6031, pp. 285–296. Springer, Heidelberg (2010). doi:10.1007/978-3-642-13089-2_24
11. Gulan, S.: On the relative descriptional complexity of regular expressions and finite automata. Ph.D. thesis, Universität Trier (2011)
12. Gulan, S., Fernau, H.: Local elimination-strategies in automata for shorter regular expressions. In: Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., Bieliková, M. (eds.) SOFSEM 2008, Vol. II, pp. 46–57 (2008)
13. Hille, E.: Analytic Function Theory, vol. 2. Blaisdell Publishing Company, New York (1962)
14. Ilie, L., Yu, S.: Follow automata. Inf. Comput. **186**(1), 140–162 (2003)
15. Lang, S.: Algebra. Graduate Texts in Mathematics, vol. 211, 3rd edn. Springer, New York (2001)
16. Ott, G., Feinstein, N.H.: Design of sequential machines from their regular expressions. J. ACM **8**(4), 585–600 (1961)
17. Project FAdo: tools for formal languages manipulation. http://fado.dcc.up.pt. Accessed Feb 2017
18. Thompson, K.: Regular expression search algorithm. Commun. ACM **11**(6), 410–422 (1968)