Regular Expressions Avoiding Absorbing Patterns and the Significance of Uniform Distribution

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

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

Although regular expressions do not correspond univocally to regular languages, it is still worthwhile to study their properties and algorithms. For the average case analysis one often relies on the uniform random generation using a specific grammar for regular expressions, that can represent regular languages with more or less redundancy. Generators that are uniform on the set of expressions are not necessarily uniform on the set of regular languages. Nevertheless, it is not straightforward that asymptotic estimates obtained by considering the whole set of regular expressions are different from those obtained using a more refined set that avoids some large class of equivalent expressions. In this paper we study a set of expressions that avoid a given absorbing pattern. It is shown that, although this set is significantly smaller than the standard one, the asymptotic average estimates for the size of the Glushkov automaton for these expressions does not differ from the standard case. Furthermore, for this set the asymptotic density of ε-accepting expressions is also the same as for the set of all standard regular expressions.

Keywords. Regular Expressions, Uniform Distribution, Average-case Complexity, Analytic Combinatorics

1. Introduction

Average-case studies often rely on uniform random generation of inputs. In general, those inputs correspond to trees, and generators are uniform on the set of these trees, but not on the set that those inputs represent (such as languages or boolean functions). Koechlin et al. [7, 8] studied expressions that have subexpressions which are (semantically) absorbing for a given operator, calling them *absorbing patterns*. For instance, $(a + b)^\star$ is absorbing for the union of regular expressions over the alphabet $\{a, b\}$, since $\alpha + (a + b)^\star$, or $(a + b)^\star + \alpha$, is equivalent to $(a + b)^\star$ for any expression α. After repeatedly applying the induced simplification, in the example above replacing $\alpha + (a + b)^\star$ by $(a + b)^\star$, the resulting expression can be significantly smaller. For uniformly random generated expressions of a given size, Koechlin et al. showed that the expression resulting from this simplification has constant expected size. That result led the authors to the conclusion that uniform random

*This work was partially supported by CMUP, member of LASI, which is financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the projects with reference UIDB/00144/2020 and UIDP/00144/2020.

generated regular expressions lack expressiveness, and in particular that uniform distribution should not be used to study the average case complexity in the context of regular languages. This conclusion is misleading in at least two aspects. First, as pointed out above, one is considering regular expressions and not regular languages themselves. For instance, if one wants to estimate the size of automata obtained from regular expressions, one disregards whether they represent the same language or not. What is implied by the results of Koechlin et al. is that, if one uniformly random generates regular expressions, one cannot expect to obtain, with a reasonable probability, regular languages outside a constant set of languages. This means that a core set of regular languages have so many regular expression representatives that the remaining languages very scarcely appear. While neither regular expressions (RE) nor nondeterministic finite automata (NFA) behave uniformly when representing regular languages, it is known that deterministic automata (DFA) are a better choice, in the uniform model, as they are asymptotically minimal [11]. In this sense, minimal DFAs are a perfect model for regular languages. However, in practice, regular expressions are usually preferred as a representation of regular languages, and are used in a not necessarily simplified form. Moreover, all of these objects (REs, NFAs, and DFAs) are combinatorial objects *per se* that can have their behaviour, as well as of the algorithms having them as input, studied on average and asymptotically. One should not confuse regular expressions by themselves with the languages that they represent. Second, the results of Koechlin et al. do not imply that asymptotic estimates obtained by considering the whole set of regular expressions are different from those obtained by using a more refined set with less equivalent expressions. For instance, some results obtained for expressions in strong star normal form coincide with the ones for standard regular expressions [2]. In order to further sustain the above claim, in this paper we consider the set R of regular expressions avoiding an absorbing pattern which extends the pattern in the example above and was the one considered by Koechlin et al. It is shown that, although the set R is significantly smaller than the set RE, the asymptotic estimates for the size of the Glushkov automaton on these sets is the same. We also show that for the set R the ratio of ε-accepting expressions is asymptotically and on average the same as for the set RE.

Given the complexity of the grammars expressing the classes here studied, we had to deal with algebraic curves and polynomials of degree depending on the size of the alphabet, k, which brought up challenges that are new, as far as we know. Not only we had to use the techniques developed in our previous work [3], but also some non-trivial estimates using Stirling approximation, and some asymptotic equivalence reductions in order to obtain the asymptotic estimates, and their limits with k.

2. The Analytic Tools

Given some measure of the objects of a combinatorial class, $\mathcal{A}$, for each non-negative integer $n \in \mathbb{N}_0$, let a_n be the sum of the values of this measure for all objects of size n. Now, let $A(z) = \sum_n a_n z^n$ be the corresponding generating function (*cf.* [5]). 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. When this function has a unique dominant singularity ρ, the study of the behaviour of $A(z)$ around it 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 [5] :

Theorem 1. *The coefficients of the series expansion of the complex function*

$$f(z) \underset{z \to \rho}{\sim} \lambda \left(1 - \frac{z}{\rho}\right)^{\nu},$$

where $\nu \in \mathbb{C} \setminus \mathbb{N}_0$, $\lambda \in \mathbb{C}$, *have the asymptotic approximation*

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

Here Γ *is, as usual, the Euler's gamma function and the notation* $f(z) \underset{z \to z_0}{\sim} g(z)$ *means that* $\lim_{z \to z_0} \frac{f(z)}{g(z)} = 1$.

2.1. *Regular Expressions*

Given an alphabet $\Sigma = \{\sigma_1, \ldots, \sigma_k\}$, the set RE of (standard) *regular expressions*, β, over Σ contains $\emptyset$ and the expressions defined by the following grammar:

$$\beta := \varepsilon \mid \sigma \in \Sigma \mid (\beta + \beta) \mid (\beta \cdot \beta) \mid (\beta^\star). \tag{1}$$

The *language* associated with β is denoted by $\mathcal{L}(\beta)$ and defined as usual (with ε representing the empty word). Two expressions β_1 and β_2 are *equivalent*, $\beta_1 = \beta_2$, if $\mathcal{L}(\beta_1) = \mathcal{L}(\beta_2)$. The *(tree-)size* $|\beta|$ of $\beta \in$ RE is the number of symbols in β (disregarding parentheses). The *alphabetic size* $|\beta|_\Sigma$ is the number of letters occurring in β. The generating function of RE is $B_k(z) = \sum_{\beta \in \mathrm{RE}} z^{|\beta|} = \sum_{n>0} b_n z^n$, where b_n is the number of expressions of size n,cf. [10, 1]. From grammar (1) one gets $B_k(z) = (k+1)z + 2zB_k(z)^2 + zB_k(z)$. Considering the quadratic equation this yields

$$B_k(z) = \frac{1 - z - \sqrt{1 - 2z - (7 + 8k)z^2}}{4z}.$$

To use Theorem 1 one needs to obtain the singularity, ρ, as well as the constants ν and λ. Following Broda et al [1, 3], we have

$$B_k(z) \underset{z \to \rho_k}{\sim} -\frac{\sqrt{2 - 2\rho_k}}{4\rho_k} \left(1 - \frac{z}{\rho_k}\right)^{\frac{1}{2}},$$

where the singularity $\rho_k = \frac{1}{1+\sqrt{8+8k}}$ is the positive root of $p_k(z) = 1-2z-(7+8k)z^2$. Thus, applying Theorem 1 and noting that $\Gamma(-\frac{1}{2}) = \sqrt{\pi}$, the number of expressions of size n is asymptotically given by

$$[z^n]B_k(z) \underset{n}{\sim} \frac{\sqrt{2-2\rho_k}}{8\rho_k\sqrt{\pi}}\, n^{-\frac{3}{2}}\, \rho_k^{-n}, \tag{2}$$

where we use the notation $\underset{n}{\sim}$ instead of $\underset{n\to\infty}{\sim}$.

3. Regular Expressions without an Absorbing Pattern

Let Θ denote any expression of the form $(\sigma_{i_1} + \cdots + \sigma_{i_k})^\star$ where $\sigma_{i_1}, \ldots, \sigma_{i_k}$ is a permutation of Σ. In this paper we consider the set R of all regular expressions α such that Θ does not occur in a union. Note that Θ represents an absorbing pattern in the sense of [7], i.e., $(\alpha+\Theta) \equiv (\Theta+\alpha) \equiv \Theta$, and that R still generates all regular languages over Σ.

For illustrating purposes, we first consider $\Sigma = \{a, b\}$, for which we have the following grammar $\mathcal{G}_2$ for R,

$$\begin{aligned}
\alpha &:= \varepsilon \mid a \mid b \mid (\alpha\cdot\alpha) \mid (\alpha^\star) \mid (\alpha_P + \alpha_P) \\
\alpha_P &:= \varepsilon \mid a \mid b \mid (\alpha\cdot\alpha) \mid (\alpha_\Sigma^\star) \mid (\alpha_P + \alpha_P) \\
\alpha_\Sigma &:= \varepsilon \mid a \mid b \mid (\alpha\cdot\alpha) \mid (\alpha^\star) \mid \gamma \\
\gamma &:= (\alpha_{ab} + \alpha_{ab}) \mid (\alpha_{ab} + a) \mid (\alpha_{ab} + b) \mid (a + \alpha_{ab}) \mid (b + \alpha_{ab}) \mid (a + a) \mid (b + b) \\
\alpha_{ab} &:= \varepsilon \mid (\alpha\cdot\alpha) \mid (\alpha_\Sigma^\star) \mid (\alpha_P + \alpha_P).
\end{aligned} \tag{3}$$

The set of expressions generated by the nonterminals of $\mathcal{G}_2$, are, respectively, the following:

$$\begin{aligned}
[\![\, \alpha \,]\!] &= \mathsf{R}, \\
[\![\, \alpha_P \,]\!] &= \{\, \alpha \in \mathsf{R} \mid \alpha \neq (a+b)^\star \wedge \alpha \neq (b+a)^\star \,\}, \\
[\![\, \alpha_\Sigma \,]\!] &= \{\, \alpha \in \mathsf{R} \mid \alpha \neq (a+b) \wedge \alpha \neq (b+a) \,\}, \\
[\![\, \gamma \,]\!] &= \{\, (\alpha_1 + \alpha_2) \in \mathsf{R} \mid \{\alpha_1, \alpha_2\} \neq \{a, b\} \,\}, \\
[\![\, \alpha_{ab} \,]\!] &= \{\, \alpha \in [\![\, \alpha_P \,]\!] \mid \alpha \neq a \wedge \alpha \neq b \,\}.
\end{aligned}$$

In particular, we obtain the correctness of $\mathcal{G}_2$.

Lemma 2. *An expression $\alpha \in$ RE is generated by $\mathcal{G}_2$ if and only the absorbing pattern $(a+b)^\star$ or $(b+a)^\star$ does not occur in a union.*

Let $R_2(z)$ denote the generating function for the class R when $|\Sigma| = 2$. It follows from (3) that

$$R_2(z) = 3z + zR_2(z)^2 + zR_2(z) + zR_P(z)^2,$$

where $R_P(z)$ is the generating function for the class of expressions generated by α_P. Comparing $[\![\, \alpha \,]\!]$ and $[\![\, \alpha_P \,]\!]$, one observes that the only expressions not generated

by α_P are $(a+b)^\star$ and $(b+a)^\star$, which are both of size 4. Thus,

$$R_P(z) = R_2(z) - 2z^4.$$

In general, for an arbitrary alphabet $\Sigma = \{\sigma_1, \ldots, \sigma_k\}$, the expressions $\alpha \in \mathsf{R}$ satisfy the following grammar $\mathcal{G}_k$

$$\alpha := \varepsilon \mid \sigma_1 \mid \cdots \mid \sigma_k \mid (\alpha \cdot \alpha) \mid (\alpha^\star) \mid (\alpha_P + \alpha_P), \tag{4}$$

where

$$[\![\, \alpha_P \,]\!] = \{\, \alpha \in \mathsf{R} \mid \alpha \neq (\sigma_{i_1} + \cdots + \sigma_{i_k})^\star \wedge \{\sigma_{i_1}, \ldots, \sigma_{i_k}\} = \Sigma \,\}.$$

As before, we obtain the following two equations for the corresponding generating functions, where $(k-1)!\binom{2k-2}{k-1}$ denotes the number of expression $(\sigma_{i_1}+\cdots+\sigma_{i_k})^\star$ with $\{\sigma_{i_1}, \ldots, \sigma_{i_k}\} = \Sigma$, each of which has size $2k$,

$$R_k(z) = (k+1)z + zR_k(z)^2 + zR_k(z) + zR_{P,k}(z)^2, \tag{5}$$

$$R_{P,k}(z) = R_k(z) - (k-1)!\binom{2k-2}{k-1}z^{2k}. \tag{6}$$

In the next section, the asymptotic estimates of $[z^n]R_k(z)$ are computed.

3.1. *Asymptotic Estimates for the Number of Expressions in* R

The generating function $R_k = R_k(z)$ satisfies the following equation:

$$2zR_k^2 - r_k R_k + z s_k = 0, \tag{7}$$

where

$$\begin{aligned} r_k = r_k(z) &= 1 - z + 2z^{2k+1}C_k, \\ s_k = s_k(z) &= 1 + k + z^{4k}C_k^2, \\ C_k &= \binom{2k-2}{k-1}(k-1)! = \frac{(2k-2)!}{(k-1)!}. \end{aligned}$$

The discriminant of equation (7) is $\Delta_k = \Delta_k(z) = p_k(z) + 4z^{2k+1}C_k h_k(z)$, where

$$\begin{aligned} p_k = p_k(z) &= 1 - 2z - (7+8k)z^2, \\ h_k = h_k(z) &= 1 - z - C_k\, z^{2k+1}. \end{aligned}$$

Thus,

$$R_k = R_k(z) = \frac{r_k - \sqrt{\Delta_k}}{4z}, \tag{8}$$

where the choice of the sign is determined by noticing that $r_k(0) = \Delta_k(0) = 1$.

Let us now show that $R_k(z)$ has a unique determinant singularity in the interval $]0,1[$, for all k. The ideia is to use the fact that the polynomial $p_k(z)$ has only one positive zero, namely ρ_k, use Rouché's Theorem to show that, in the disk

$|z| < \frac{1}{\sqrt{8+8k}}$, the polynomial $\Delta_k(z)$ has exactly one root in that disk, and finally show that that unique root is real. We recall that Rouché's Theorem states that, in particular, for polynomials $f(z)$ and $g(z)$ such that $|f(z) - g(z)| < |f(z)| + |g(z)|$ holds for all $|z| = R$, in the complex plane, then $f(z)$ and $g(z)$ have the same number of roots, taking into account multiplicities, in the disk $|z| < R$ [13] .

In order to estimate $|\Delta_k(z) - p_k(z)|$, we start by noticing that from Stirling approximation, $\sqrt{2\pi}\, n^{n+\frac{1}{2}} e^{-n} \leq n! \leq n^{n+\frac{1}{2}} e^{1-n}$, valid for all $n \in \mathbb{N}$, one gets that, for all $k \geq 2$,

$$\frac{\sqrt{2\pi}\,(2k-2)^{2k-\frac{3}{2}}\, e^{2-2k}}{(k-1)^{k-\frac{1}{2}}\, e^{2-k}} \leq C_k = \frac{(2k-2)!}{(k-1)!} \leq \frac{(2k-2)^{2k-\frac{3}{2}}\, e^{3-2k}}{\sqrt{2\pi}\,(k-1)^{k-\frac{1}{2}}\, e^{1-k}},$$

i.e.,

$$\frac{\sqrt{2\pi}\, 2^{2k-\frac{3}{2}} (k-1)^{k-1}}{e^k} \leq C_k \leq \frac{2^{2k-\frac{3}{2}} (k-1)^{k-1}}{\sqrt{2\pi}\, e^{k-2}}. \tag{9}$$

Therefore, for $|z| = \frac{1}{\sqrt{8+8k}}$,

$$\begin{aligned}
|\Delta_k(z) - p_k(z)| &\leq 4C_k \frac{1}{(8+8k)^{k+\frac{1}{2}}} |h_k(z)| \\
&\leq \frac{(k-1)^{k-1}}{\sqrt{2\pi}\, e^{k-2}\, 2^{k+1} (k+1)^{k+\frac{1}{2}}} \left(1 - \frac{1}{\sqrt{8+8k}} - \frac{C_k}{(8+8k)^{k+\frac{1}{2}}}\right) \\
&\leq \frac{1.48}{(2e)^k (k-1)\sqrt{k+1}} \left(1 - \frac{1}{\sqrt{8+8k}} - \frac{C_k}{(8+8k)^{k+\frac{1}{2}}}\right).
\end{aligned}$$

Noticing that, from (9), one has

$$\frac{\sqrt{2\pi}\,(k-1)^{k-1}}{e^k 2^{k+3} (k+1)^{k+\frac{1}{2}}} \leq \frac{C_k}{(8+8k)^{k+\frac{1}{2}}} \leq \frac{(k-1)^{k-1}}{\sqrt{2\pi}\, e^{k-2} 2^{k+3} (k+1)^{k+\frac{1}{2}}},$$

one concludes that

$$\lim_{k\to\infty} |\Delta_k(z) - p_k(z)| = 0.$$

Let us now find the minimum of $|p_k(z)|$ on the circumference $|z| = \frac{1}{\sqrt{8+8k}} = R$. Put $z = Re^{i\theta}$. One has

$$\begin{aligned}
|p_k(z)|^2 &= |1 - 2Re^{i\theta} - (7+8k)R^2 e^{2i\theta}|^2 \\
&= (1 - 2R\cos\theta - (7+8k)R^2\cos 2\theta)^2 + (1 - 2R\sin\theta - (7+8k)R^2 \sin 2\theta)^2 \\
&= 2 + 4R^2 + (7+8k)^2 R^4 - 4R(\cos\theta + \sin\theta) - 2(7+8k)R^2(\cos 2\theta + \sin 2\theta) \\
&\quad + 4R^3(7+8k)(\cos\theta\cos 2\theta + \sin\theta \sin 2\theta) \\
&= 2 + \frac{1}{2+2k} + \left(\frac{7+8k}{8+8k}\right)^2 - \frac{2(\cos\theta + \sin\theta)}{\sqrt{2+2k}} - \frac{(7+8k)(\cos 2\theta + \sin 2\theta)}{4+4k} \\
&\quad + \frac{(7+8k)(\cos\theta\cos 2\theta + \sin\theta\sin 2\theta)}{4(k+1)\sqrt{2+2k}}.
\end{aligned}$$

It follows that $\lim_{k\to\infty} |p_k(z)|^2 = 3 - 2(\cos 2\theta + \sin 2\theta)$. Since $\max_\theta(\cos\theta + \sin\theta) = \sqrt{2}$, one concludes that $\lim_{k\to\infty} |p_k(z)|^2 \geq 3 - 2\sqrt{2} > 0$. From all this, one concludes that $|\Delta_k(z) - p_k(z)| < |p_k(z)|$ for large enough values of k, and so Rouché's Theorem applies to show that the polynomial $\Delta_k(z)$ has exactly one root in the open disk $|z| <_{\text{a}} \frac{1}{\sqrt{8+8k}}$.

Since $\Delta_k(0) = 1$, in order to show that that root must be real it suffices to show that one has $\Delta_k\left(\frac{1}{\sqrt{8+8k}}\right) < 0$. This can be shown as follows. Since

$$\Delta_k\left(\frac{1}{\sqrt{8+8k}}\right) = 2^{-6k-\frac{7}{2}}(k+1)^{-2k-1}\left(2^{3k+2}\left(4\sqrt{k+1}-\sqrt{2}\right)(k+1)^k C_k \right.$$
$$\left. -4\sqrt{2}\,C_k^2 - 64^k\left(8\sqrt{k+1}-\sqrt{2}\right)(k+1)^{2k}\right),$$

we want to show that

$$2^{3k}\left(\sqrt{8k+8}-1\right)(k+1)^k C_k < C_k^2 + 2^{6k-2}\left(2\sqrt{8k+8}-1\right)(k+1)^{2k}.$$

Using (9), it is enough to show that

$$\frac{2^k e^{k+2}\left(\sqrt{8k+8}-1\right)}{\sqrt{\pi}} < 2^{2k}\left(2\sqrt{8k+8}-1\right)\frac{(k+1)^k}{(k-1)^{k-1}}e^{2k} + \pi\frac{(k-1)^{k-1}}{(k+1)^k},$$

that follows from this trivially true inequality

$$\frac{\sqrt{8k+8}-1}{\sqrt{\pi}} < 2^k\left(2\sqrt{8k+8}-1\right)\frac{(k+1)^k}{(k-1)^{k-1}}e^{k-2}.$$

The singularity of $R_k(z)$ is therefore given by the unique root of $\Delta_k(z)$ in the interval $\left]0, \frac{1}{\sqrt{8k+8}}\right[$, which will henceforth denote by η_k. It also follows from Rouché's Theorem that this root has multiplicity one. Now, $\Delta_k(z) = \left(1 - \frac{z}{\eta_k}\right)\psi_k(z)$, for some $\psi_k(z) \in \mathbb{R}[z]$. Using L'Hôpital's Rule, one has

$$\psi_k(\eta_k) = -\eta_k \Delta_k'(\eta_k). \tag{10}$$

Then, one has

$$R_k(z) \underset{z\to\eta_k}{\sim} \frac{-r_k(\eta_k) - \sqrt{\psi_k(\eta_k)}\left(1-\frac{z}{\eta_k}\right)^{\frac{1}{2}}}{4\eta_k}.$$

By Theorem 1, one gets the following asymptotic approximation for the number of regular expressions in appreciation

Theorem 3. *With the notation above, one has*

$$[z^n]R_k(z) \underset{n}{\sim} \frac{\sqrt{\psi_k(\eta_k)}}{8\eta_k\sqrt{\pi}} n^{-\frac{3}{2}} \eta_k^{-n}.$$

[a]It is actually true that $|\Delta_k(z) - p_k(z)| < |p_k(z)|$ for all $|z| = \frac{1}{\sqrt{8+8k}}$ and $k \geq 2$.

Using (2), we have

The asymptotic ratio of the number of expressions in R *and the number of expressions in* RE *is given by,*

$$\frac{[z^n]R_k(z)}{[z^n]B_k(z)} \underset{n}{\sim} \frac{\frac{\sqrt{\psi_k(\eta_k)}}{8\eta_k\sqrt{\pi}} n^{-\frac{3}{2}} \eta_k^{-n}}{\frac{\sqrt{2-2\rho_k}}{8\rho_k\sqrt{\pi}}\, n^{-\frac{3}{2}}\, \rho_k^{-n}} = \frac{\sqrt{\psi_k(\eta_k)}}{\sqrt{2-2\rho_k}} \left(\frac{\rho_k}{\eta_k}\right)^{n+1}.$$

Since, as seen before, $\eta_k > \rho_k$, for all k, this yields that, for every k, this ratio tends to 0 as $n \to \infty$. As such, considering R instead of RE, actually avoids a significant set of redundant expressions. Such an improvement, in the sense of [7], might influence the results obtained by asymptotic studies.

In Section 5 we show that is not the case for the average asymptotic size of the Glushkov automaton in terms of states and transitions [10, 1]. In the next section, we estimate the number of expressions from R that accept the word ε and show that their density is asymptotically and on average the same as for standard regular expressions, RE.

4. Density of ε-accepting Regular Expressions

In this section, we estimate the ratio of ε-accepting regular expressions to regular expressions avoiding the absorbing pattern Θ. Formally, let $\alpha_\varepsilon \in \mathsf{R}$ be the set of expressions such that $\varepsilon \in \mathcal{L}(\alpha_\varepsilon)$ and let $\alpha_{\overline{\varepsilon}}$ represent the set of expressions such that $\varepsilon \notin \mathcal{L}(\alpha_{\overline{\varepsilon}})$. We have that those sets satisfy the following grammars:

$$\begin{aligned}
\alpha_\varepsilon &:= \varepsilon \mid (\alpha_\varepsilon \cdot \alpha_\varepsilon) \mid (\alpha^\star) \mid (\alpha_{P,\varepsilon} + \alpha_P) \mid (\alpha_{P,\overline{\varepsilon}} + \alpha_{P,\varepsilon}),\\
\alpha_{\overline{\varepsilon}} &:= \sigma \in \Sigma \mid (\alpha_{\overline{\varepsilon}} \cdot \alpha) \mid (\alpha_\varepsilon \cdot \alpha_{\overline{\varepsilon}}) \mid (\alpha_{P,\overline{\varepsilon}} + \alpha_{P,\overline{\varepsilon}}),
\end{aligned}$$

where $\alpha_{P,\varepsilon}$ and $\alpha_{P,\overline{\varepsilon}}$ represent the expressions α_P such that $\varepsilon \in \mathcal{L}(\alpha_{P,\varepsilon})$ and $\varepsilon \notin \mathcal{L}(\alpha_{P,\overline{\varepsilon}})$, respectively. The correspondent generating functions satisfy

$$\begin{aligned}
R_{\varepsilon,k}(z) &= z + zR_{\varepsilon,k}(z)^2 + zR_k(z) + 2zR_{P,\varepsilon,k}(z)R_{P,k}(z) - zR_{P,\varepsilon,k}(z)^2,\\
R_{\overline{\varepsilon},k}(z) &= R_k(z) - R_{\varepsilon,k}(z),\\
R_{P,\overline{\varepsilon},k}(z) &= R_{\overline{\varepsilon},k}(z),\\
R_{P,\varepsilon,k}(z) &= R_{P,k}(z) - R_{\overline{\varepsilon},k}(z) = R_{\varepsilon,k}(z) - C_k z^{2k}.
\end{aligned}$$

From that we conclude that $R_{\varepsilon,k} = R_{\varepsilon,k}(z)$ satisfies

$$R_{\varepsilon,k} = z + zR_k + 2zR_{\varepsilon,k}R_k + zC_k^2 z^{4k} - 2zR_kC_k z^{2k}, \tag{11}$$

Solving equation (6) in order to $R_{P,k}$, substituting the obtained value in equation (5), one gets the polynomial having R_k as root:

$$g(X) = 2z\,X^2 - (1 - z + 2C_k z^{2k+1})\,X + (k+1)z + C_k^2 z^{4k+1}. \tag{12}$$

Solving equation (11),

in order to $R_{\varepsilon,k}$, one gets

$$R_{\varepsilon,k} = z\,\frac{1 + C_k^2 z^{4k} + R_k - 2C_k z^{2k} R_k}{1 - 2zR_k}. \tag{13}$$

Working in the field $\mathbb{K} = \mathbb{Q}(k,z)[X]/\langle g(X)\rangle$ ($g(X)$ is irreducible over the field $\mathbb{Q}(k,z)$), similarly to what was done in Section 6 of [9], one finds that

$$R_{\varepsilon,k} = a_{11} + a_{12}\,R_k, \tag{14}$$

where

$$\begin{aligned} a_{11} &= \frac{kz(-1 + 2C_k z^{2k})}{1 + 2(k+1)z - 2C_k z^{2k} + 2C_k^2 z^{4k+1}} \\ a_{12} &= 1 - \frac{2kz}{1 + 2(k+1)z - 2C_k z^{2k} + 2C_k^2 z^{4k+1}}. \end{aligned}$$

Also, in $\mathbb{K}$,

$$R_{\varepsilon,k} R_k = a_{21} + a_{22}\,R_k, \tag{15}$$

for some $a_{21}, a_{22} \in \mathbb{Q}(k,z)$.

From equations (14) and (15) one obtains

$$\begin{vmatrix} R_{\varepsilon,k} - a_{11} & -a_{12} \\ -a_{21} & R_{\varepsilon,k} - a_{22} \end{vmatrix} = 0,$$

yielding a second degree polynomial, $2zd_2X^2 - d_1X + zd_0$ having $R_{\varepsilon,k}$ as a root, where

$$\begin{aligned} d_0 =& 1 + (k+2)z - 2C_k z^{2k} + C_k^2 z^{4k} - 2C_k^3 z^{6k} - 4C_k k z^{2k+1} \\ &+ 4(k+1)C_k^2 z^{4k+1} + 2C_k^4 z^{8k+1} \\ d_1 =& 1 + z - 2(2k+1)z^2 - 2C_k z^{2k} + 4C_k z^{2k+1} + 4C_k(2k+1)z^{2k+2} \\ &- 2C_k^2 z^{4k+1} - 2C_k^2 z^{4k+2} + 4C_k^3 z^{6k+2} \\ d_2 =& 1 + 2(k+1)z - 2C_k z^{2k} + 2C_k^2 z^{4k+1}. \end{aligned}$$

Therefore,

$$R_{\varepsilon,k} = \frac{d_1 - \sqrt{d_1^2 - 8z^2 d_0 d_2}}{4z\,d_2} \tag{16}$$

(the sign was chosen so that $R_{\varepsilon,k}(0) = 0$).

We now show that $d_2(z)$ is positive for $0 < z < 1$, which implies, using Pringsheim Theorem, that the singularity comes from the smallest positive zero of the polynomial inside the square root. To show that, one first makes the change of variable $x = \frac{1}{z} > 1$, so that the inequality $d_2(z) > 0$ is then equivalent to

$$f(x) = \frac{k+1}{C_k} x^{2k-1} + \frac{C_k}{x^{2k+1}} > 1.$$

Since $f(x) \to +\infty$ both as $x \to 0$ and $x \to +\infty$, and as $f'(x)$ has only one zero, namely

$$x_0 = \left(\frac{2k+1}{(k+1)(2k-1)}\right)^{\frac{1}{4k}} C_k^{\frac{1}{2k}},$$

this has to be the absolute minimum of f. Since

$$f(x_0) = (k+1)\,\frac{\left(\frac{2k+1}{(k+1)(2k-1)}\right)^{\frac{1}{2}-\frac{1}{4k}}}{C_k^{\frac{1}{2k}}} + \frac{1}{\left(\frac{2k+1}{(k+1)(2k-1)}\right)^{\frac{1}{2}+\frac{1}{4k}} C_k^{\frac{1}{2k}}},$$

it is enough to ensure that

$$(k+1)\left(\frac{2k+1}{(k+1)(2k-1)}\right)^{\frac{1}{2}-\frac{1}{4k}} + \frac{1}{\left(\frac{2k+1}{(k+1)(2k-1)}\right)^{\frac{1}{2}+\frac{1}{4k}}} > C_k^{\frac{1}{2k}}.$$

This can easily be done by noticing that the first summand is always greater that the second, and then using the right inequality in (9).

Next we found out that:

$$d_1^2 - 8z^2 d_0 d_2 = t_k(z)^2 \Delta_k(z), \tag{17}$$

where $\Delta_k(z)$ is as above, and $t_k(z) = 1 + 2z - 2C_k z^{2k} + 2C_k^2 z^{4k+1}$. This implies that η_k, defined in p. 7, is the dominant singularity of $R_{\varepsilon,k}$, and that

$$R_{\varepsilon,k} \underset{z\,\to\,\eta_k}{\sim} \frac{-t_k(\eta_k)\psi_k(\eta_k)}{4\eta_k d_2(\eta_k)}\left(1 - \frac{z}{\eta_k}\right)^{\frac{1}{2}} \tag{18}$$

from which one gets

$$[z^n]R_{\varepsilon,k} \underset{n}{\sim} \frac{-t_k(\eta_k)\psi_k(\eta_k)}{4\eta_k d_2(\eta_k)} n^{-\frac{3}{2}} \eta_k^n. \tag{19}$$

Using Theorem 3, one obtains

$$\frac{[z^n]R_{\varepsilon,k}}{[z^n]R_k} \underset{n}{\sim} \frac{2t_k(\eta_k)\eta_k}{d_2(\eta_k)}. \tag{20}$$

Let us now see that

$$\lim_{k\to\infty} k\,\eta_k^2 = \frac{1}{8}. \tag{21}$$

Since we know that $\Delta_k(0) = 1$, and $\Delta_k(x)$ has exactly one real root in the interval $\left[0, \frac{1}{\sqrt{8+8k}}\right]$, in order to show that $\eta_k > \rho_k$ for all k, it is enough to show that:

$$\Delta_k(\rho_k) = p_k(\rho_k) + 4\rho_k^{2k+1} C_k h_k(\rho_k) > 0, \text{ i.e., } h_k(\rho_k) > 0.$$

Now, $h_k(\rho_k) > 0 \iff 1 > \rho_k + C_k \rho_k^{2k+1} \iff \sqrt{8+8k} > \frac{C_k}{(1+\sqrt{8+8k})^{2k}}$. From (9) it follows that

$$\frac{C_k}{(1+\sqrt{8+8k})^{2k}} \le \frac{2^{2k-\frac{3}{2}}(k-1)^{k-1}}{\sqrt{2\pi}\,e^{k-2}(1+\sqrt{8+8k})^{2k}}.$$

It is therefore enough to show:

$$\frac{2^{2k-\frac{3}{2}}(k-1)^{k-1}}{\sqrt{2\pi}\,e^{k-2}(1+\sqrt{8+8k})^{2k}} < \sqrt{8+8k},$$

which is equivalent to

$$2^{2k-\frac{3}{2}}(k-1)^{k-1} < \sqrt{2\pi}\,e^{k-2}(1+\sqrt{8+8k})^{2k}\sqrt{8+8k}.$$

This is the same as

$$\left(\frac{4}{e}\right)^k (k-1)^{k-1} < \frac{2^{\frac{3}{2}}\sqrt{2\pi}}{e^2}\,(1+\sqrt{8+8k})^{2k}\sqrt{8+8k},$$

which follows from:

$$\left(\frac{4}{e}\right)^k (k-1)^{k} < \frac{2^{\frac{3}{2}}\sqrt{2\pi}}{e^2}\,2^{2k+1}(2+2k)^{k+\frac{1}{2}}.$$

That is obvious when rewritten as

$$\left(\frac{4}{e}\right)^k (k-1)^{k} < \left(\frac{2^{\frac{3}{2}}\sqrt{2\pi}}{e^2}\,2\right)\,4^k(2+2k)^{k+\frac{1}{2}}.$$

Thus, we conclude that

$$\rho_k = \frac{1}{1+\sqrt{8+8k}} < \eta_k < \frac{1}{\sqrt{8+8k}}. \tag{22}$$

From this it immediately follows that $\lim\limits_{k\to\infty} k\,\eta_k^2 = \frac{1}{8}$, and then $\lim\limits_{k\to\infty} p_k(\eta_k) = 0$. Using the right hand inequality in (9) together with (22), it is not hard to show the following result, which will be useful latter to deduce equation (31).

Lemma 5. *For all $t, s \in \mathbb{R}$, one has*

$$\lim_{k\to\infty} C_k k^t \eta_k^{2k+s} = 0. \tag{23}$$

Using (21) one can easily show that

$$\frac{2t_k(\eta_k)\eta_k}{d_2(\eta_k)} \underset{k}{\sim} \sqrt{\frac{2}{k}}. \tag{24}$$

We note that it follows from Lemma 2 and Equation (4) in [10] (and also [12], Thm. 12) that for the usual regular expressions one has the exact same asymptotic density of ε-accepting regular expressions for the set of all regular expressions as the regular expressions avoiding the absorbing pattern Θ.

5. Asymptotic Average Size of the Glushkov Automaton

The Glushkov automaton [6] is constructed from an equivalent regular expression β using the set $\mathsf{Pos}(\beta)$ of positions of the letters in β, as the set of states (plus one initial state). Let $\mathsf{Pos}(\beta) = \{1, 2, \ldots, |\beta|_\Sigma\}$, $\mathsf{Pos}_0(\beta) = \mathsf{Pos}(\beta) \cup \{0\}$ and $\overline{\beta}$ denote the expression obtained from β by marking each letter with its position in β. The construction is based on the position sets

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

The *Glushkov automaton* for β is

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

where the set of final states is $F = \mathsf{Last}(\overline{\beta}) \cup \{0\}$ if $\varepsilon \in \mathcal{L}(\beta)$, and $F = \mathsf{Last}(\overline{\beta})$, otherwise; and the set of transitions is

$$\delta_{\mathrm{POS}} = \{\, (0, \overline{\sigma_j}, j) \mid j \in \mathsf{First}(\overline{\beta}) \,\} \cup \{\, (i, \overline{\sigma_j}, j) \mid\ (i,j) \in \mathsf{Follow}(\overline{\beta}) \,\}.$$

In this section, we estimate the average size of $\mathcal{A}_{\mathrm{POS}}$ for expressions in R. In the next subsection, we estimate the average number of letters in $\alpha \in \mathsf{R}$, i.e., the number of states of $\mathcal{A}_{\mathrm{POS}}(\alpha)$. In the last subsection, we consider the number of transitions.

5.1. *Estimates for the Number of Letters*

The average number of letters in uniform random generated regular expressions of a given size have been estimated for different kinds of expressions [10, 3]. For standard regular expressions that value is half the size of the expressions as the size of the alphabet goes to ∞. In the following we obtain the same value for expressions in R. To count the number of letters in all expressions of a given size we use the bivariate generating function $\mathcal{L}_k(u,z) = \sum_{n,i\geq 1} c_{n,i} u^i z^n$, where $c_{n,i}$ is the number of regular expressions of size n with i letters. Therefore, the total number of letters in all the regular expressions of size n is given by the coefficients of the sum of the two series

$$L_k(z) = \frac{\partial \mathcal{L}_k(u,z)}{\partial u}\bigg|_{u=1} = \sum_{n,i\geq 1} i\, c_{n,i}\, z^n.$$

From grammar (4) the generating function $L_k(z)$ satisfies the following.

$$L_k(z) = kz + 2zL_k(z)R_k(z) + zL_k(z) + 2zP_k(z)R_P(z), \tag{25}$$

$$P_k(z) = L_k(z) - k!\binom{2k-2}{k-1} z^{2k}. \tag{26}$$

Using equations (5), (6), (25), (26) and Buchberger's algorithm [4] one obtains the following equation, which is satisfied by the generating function $L_k = L_k(z)$:

$$\Delta_k L_k^2 + \bar{r}_k L_k - \bar{s}_k = 0, \tag{27}$$

where

$$\bar{r}_k = kz^{2k}C_k\,\Delta_k,$$
$$\bar{s}_k = kz^2 + k^2 z^{2k+1}\,C_k\,\left((z-1)(1+2z^{4k+1}C_k^2) + 2C_k(2+k) + 2z^{6k+1}C_k^3\right).$$

The discriminant of equation (27) can be shown to be

$$\bar{\Delta}_k(z) = z^2k^2\Delta_k(z)g_k(z)^2, \tag{28}$$

where

$$g_k(z) = 2 - C_k z^{2k-1}\left(h_k(z) - C_k z^{2k-1}\right). \tag{29}$$

Therefore,

$$L_k(z) = \frac{kz^{2k}C_k\Delta_k(z) \pm \sqrt{\bar{\Delta}_k(z)}}{2\Delta_k(z)} = \frac{kz^{2k}C_k}{2} \pm \frac{kzg_k(z)}{2\sqrt{\Delta_k(z)}}.$$

Using the fact that we know $L_k'(0) = k$, one deduces that

$$L_k(z) = \frac{kz^{2k}C_k}{2} + \frac{kzg_k(z)}{2\sqrt{\Delta_k(z)}}. \tag{30}$$

Now, applying the procedure described in Broda et al. [3] one obtains

Theorem 6. *With the same notation as above, where η_k is as defined in page 7,*

$$[z^n]L_k(z) \underset{n}{\sim} \frac{k\,\eta_k\,g_k(\eta_k)}{2\sqrt{\pi}\sqrt{\psi_k(\eta_k)}} n^{-\frac{1}{2}}\eta_k^{-n}.$$

Therefore, from Theorems 3 and 6, one deduces

Theorem 7. *The asymptotic ratio of letters in the expressions in* R *is given by*

$$\frac{[z^n]L_k(z)}{n[z^n]R_k(z)} \underset{n}{\sim} \frac{4k\,\eta_k^2\,g_k(\eta_k)}{\psi_k(\eta_k)}.$$

From Lemma 5, and from (29) and (10), one easily gets $\lim_{k\to\infty} g_k(\eta_k) = \lim_{k\to\infty} \psi_k(\eta_k) = 2$, and thus:

$$\lim_{k\to\infty} \frac{4k\,\eta_k^2\,g_k(\eta_k)}{\psi_k(\eta_k)} = \frac{1}{2}. \tag{31}$$

This means that the following result holds.

Theorem 8. *In regular expressions without Θ in unions, the asymptotic ratio of letters and the size of the expression goes to $\frac{1}{2}$ as k goes to ∞.*

5.2. *Estimates for the Number of Transitions*

The transitions of the Glushkov automaton are defined using the sets First, Last and Follow. These sets can be inductively define for $\alpha \in \mathsf{R}$, as it is usually done [1]. Let $\alpha_\varepsilon \in \mathsf{R}$ be the set of expressions such that $\varepsilon \in \mathcal{L}(\alpha_\varepsilon)$ and let $\alpha_{\overline{\varepsilon}}$ represent the set of expressions such that $\varepsilon \notin \mathcal{L}(\alpha_{\overline{\varepsilon}})$. We have

$$\begin{aligned}
\mathsf{First}(\varepsilon) &= \emptyset, & \mathsf{First}(\alpha_P + \alpha'_P) &= \mathsf{First}(\alpha_P) \cup \mathsf{First}(\alpha'_P),\\
\mathsf{First}(\sigma_i) &= \{i\}, & \mathsf{First}(\alpha_\varepsilon \cdot \alpha) &= \mathsf{First}(\alpha_\varepsilon) \cup \mathsf{First}(\alpha),\\
\mathsf{First}(\alpha^\star) &= \mathsf{First}(\alpha), & \mathsf{First}(\alpha_{\overline{\varepsilon}} \cdot \alpha) &= \mathsf{First}(\alpha_{\overline{\varepsilon}}).
\end{aligned}$$

The definition of Last is almost identical and differs only for the case of concatenation, which is $\mathsf{Last}(\alpha\cdot\alpha_\varepsilon) = \mathsf{Last}(\alpha)\cup\mathsf{Last}(\alpha_\varepsilon)$ and $\mathsf{Last}(\alpha\cdot\alpha_{\overline{\varepsilon}}) = \mathsf{Last}(\alpha_{\overline{\varepsilon}})$. Following Broda et al. [1] the set Follow satisfies

$$\begin{aligned}
\mathsf{Follow}(\varepsilon) = \mathsf{Follow}(\sigma_i) &= \emptyset,\\
\mathsf{Follow}(\alpha_P + \alpha'_P) &= \mathsf{Follow}(\alpha_P) \cup \mathsf{Follow}(\alpha'_P),\\
\mathsf{Follow}(\alpha \cdot \alpha') &= \mathsf{Follow}(\alpha) \cup \mathsf{Follow}(\alpha') \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\alpha'),\\
\mathsf{Follow}(\alpha^\star) &= \mathsf{E}^\star(\alpha),
\end{aligned}$$

where

$$\begin{aligned}
\mathsf{E}^\star(\varepsilon) = \emptyset,\ \ \mathsf{E}^\star(\sigma_i) &= \{(i,i)\},\ \ \mathsf{E}^\star(\alpha^\star) = \mathsf{E}^\star(\alpha),\\
\mathsf{E}^\star(\alpha_P + \alpha'_P) &= \mathsf{E}^\star(\alpha_P) \cup \mathsf{E}^\star(\alpha'_P) \cup \mathsf{Cross}(\alpha_P, \alpha'_P),\\
\mathsf{E}^\star(\alpha_\varepsilon \cdot \alpha'_\varepsilon) &= \mathsf{E}^\star(\alpha_\varepsilon) \ \cup\ \mathsf{E}^\star(\alpha'_\varepsilon) \ \cup\ \mathsf{Cross}(\alpha_\varepsilon, \alpha'_\varepsilon),\\
\mathsf{E}^\star(\alpha_\varepsilon \cdot \alpha'_{\overline{\varepsilon}}) &= \mathsf{Follow}(\alpha_\varepsilon) \ \cup\ \mathsf{Follow}^\star(\alpha'_{\overline{\varepsilon}}) \ \cup\ \mathsf{Cross}(\alpha_\varepsilon, \alpha'_{\overline{\varepsilon}}),\\
\mathsf{E}^\star(\alpha_{\overline{\varepsilon}} \cdot \alpha'_\varepsilon) &= \mathsf{Follow}^\star(\alpha_{\overline{\varepsilon}}) \ \cup\ \mathsf{Follow}(\alpha'_\varepsilon) \ \cup\ \mathsf{Cross}(\alpha_{\overline{\varepsilon}}, \alpha'_\varepsilon),\\
\mathsf{E}^\star(\alpha_{\overline{\varepsilon}} \cdot \alpha'_{\overline{\varepsilon}}) &= \mathsf{Follow}(\alpha_{\overline{\varepsilon}}) \ \cup\ \mathsf{Follow}(\alpha'_{\overline{\varepsilon}}) \ \cup\ \mathsf{Cross}(\alpha_{\overline{\varepsilon}}, \alpha'_{\overline{\varepsilon}}),
\end{aligned}$$

with $\mathsf{Cross}(\alpha, \alpha') = \mathsf{Last}(\alpha)\times\mathsf{First}(\alpha')\cup\mathsf{Last}(\alpha')\times\mathsf{First}(\alpha)$. The function that counts the cardinality of $\mathsf{First}(\alpha)$ is $\mathsf{f}(\alpha)$ and is defined as follows:

$$\begin{aligned}
\mathsf{f}(\sigma_i) &= 1,\\
\mathsf{f}(\alpha_P + \alpha'_P) &= \mathsf{f}(\alpha_P) + \mathsf{f}(\alpha'_P),\\
\mathsf{f}(\alpha_\varepsilon \cdot \alpha') &= \mathsf{f}(\alpha_\varepsilon) + \mathsf{f}(\alpha'),\\
\mathsf{f}(\alpha_{\overline{\varepsilon}} \cdot \alpha') &= \mathsf{f}(\alpha_{\overline{\varepsilon}}),\\
\mathsf{f}(\alpha^\star) &= \mathsf{f}(\alpha).
\end{aligned}$$

Note that $\mathsf{f}((\sigma_{i_1} + \cdots + \sigma_{i_k})^\star) = k$ for any permutation $\sigma_{i_1}, \ldots, \sigma_{i_k}$ of $\Sigma = \{\sigma_1, \ldots, \sigma_k\}$. The correspondent generating function $F_k(z) = \sum_\alpha \mathsf{f}(\alpha) z^{|\alpha|} = F_k$ satisfies

$$\begin{aligned}
F_k &= kz + zF_k + 2zF_{P,k}R_{P,k} + zF_kR_{\varepsilon,k} + zF_kR_k,\\
F_{P,k} &= F_k - kC_kz^{2k},
\end{aligned}$$

where $R_{\varepsilon,k} = R_{\varepsilon,k}(z)$ is the generating function for expressions $\alpha_\varepsilon \in \mathsf{R}$, studied in Section 4. Let $\mathsf{s}(\alpha)$ be the function that counts the cardinality of $\mathsf{Last}(\alpha)$ and $S_k(z)$

the correspondent generating function. By symmetry we have that $S_k(z) = F_k(z)$. The functions counting the cardinalities of $\mathsf{Follow}(\alpha)$ and $\mathsf{E}^\star(\alpha)$ are $\mathsf{e}(\alpha)$ and $\mathsf{e}^\star(\alpha)$, respectively. Those functions are defined as follows:

$$\begin{aligned}
\mathsf{e}(\sigma) = \mathsf{e}(\varepsilon) &= 0,\\
\mathsf{e}(\alpha_P + \alpha'_P) &= \mathsf{e}(\alpha_P) + \mathsf{e}(\alpha'_P),\\
\mathsf{e}(\alpha \cdot \alpha') &= \mathsf{e}(\alpha) + \mathsf{e}(\alpha') + \mathsf{s}(\alpha)\,\mathsf{f}(\alpha'),\\
\mathsf{e}(\alpha^\star) &= \mathsf{e}^\star(\alpha),
\end{aligned}$$

where $\mathsf{e}^\star(\alpha)$ is given by

$$\begin{aligned}
\mathsf{e}^\star(\varepsilon) = 0, \qquad \mathsf{e}^\star(\sigma) &= 1,\\
\mathsf{e}^\star(\alpha_P + \alpha'_P) &= \mathsf{e}^\star(\alpha_P) + \mathsf{e}^\star(\alpha'_P) + \mathsf{c}(\alpha_P, \alpha'_P),\\
\mathsf{e}^\star(\alpha_\varepsilon \cdot \alpha'_\varepsilon) &= \mathsf{e}^\star(\alpha_\varepsilon) + \mathsf{e}^\star(\alpha'_\varepsilon) + \mathsf{c}(\alpha_\varepsilon, \alpha'_\varepsilon),\\
\mathsf{e}^\star(\alpha_{\overline{\varepsilon}} \cdot \alpha'_\varepsilon) &= \mathsf{e}^\star(\alpha_{\overline{\varepsilon}}) + \mathsf{e}(\alpha'_\varepsilon) + \mathsf{c}(\alpha_{\overline{\varepsilon}}, \alpha'_\varepsilon),\\
\mathsf{e}^\star(\alpha_\varepsilon \cdot \alpha'_{\overline{\varepsilon}}) &= \mathsf{e}(\alpha_\varepsilon) + \mathsf{e}^\star(\alpha'_{\overline{\varepsilon}}) + \mathsf{c}(\alpha_\varepsilon, \alpha'_{\overline{\varepsilon}}),\\
\mathsf{e}^\star(\alpha_{\overline{\varepsilon}} \cdot \alpha'_{\overline{\varepsilon}}) &= \mathsf{e}(\alpha_{\overline{\varepsilon}}) + \mathsf{e}(\alpha'_{\overline{\varepsilon}}) + \mathsf{c}(\alpha_{\overline{\varepsilon}}, \alpha'_{\overline{\varepsilon}}),\\
\mathsf{e}^\star(\alpha^\star) &= \mathsf{e}^\star(\alpha),
\end{aligned}$$

with $\mathsf{c}(\alpha, \alpha') = \mathsf{s}(\alpha)\,\mathsf{f}(\alpha') + \mathsf{s}(\alpha')\,\mathsf{f}(\alpha)$. From the above the corresponding generating functions $E_k(z) = \sum_\alpha \mathsf{e}(\alpha) z^{|\alpha|} = E_k$ and $E^\star_k(z) = \sum_\alpha \mathsf{e}^\star(\alpha) z^{|\alpha|} = E^\star_k$, respectively, satisfy

$$\begin{aligned}
E_k &= 2zE_{P,k}R_{P,k} + 2zE_kR_k + zF_k^2 + zE^\star_k,\\
E^\star_k &= kz + 2zE^\star_{P,k}R_{P,k} + 2zF^2_{P,k} + 2zE^\star_{\varepsilon,k}R_{\varepsilon,k} + 2zF_{\varepsilon,k}F_{\varepsilon,k}\\
&\quad + zE^\star_{\overline{\varepsilon},k}R_{\varepsilon,k} + zE_{\varepsilon,k}R_{\overline{\varepsilon},k} + 2zF_{\varepsilon,k}F_{\overline{\varepsilon},k} + zE_{\varepsilon,k}R_{\overline{\varepsilon},k}\\
&\quad + zE^\star_{\overline{\varepsilon},k}R_{\varepsilon,k} + 2zF_{\varepsilon,k}F_{\overline{\varepsilon},k} + 2zE_{\overline{\varepsilon},k}R_{\overline{\varepsilon},k} + 2zF_{\overline{\varepsilon},k}F_{\overline{\varepsilon},k} + zE^\star_k\\
&= kz + 2zE^\star_{P,k}R_{P,k} + 2zE^\star_kR_{\varepsilon,k} + 2zE_k(R_k - R_{\varepsilon,k})\\
&\quad + 2zF^2_{P,k} + 2zF^2_k + zE^\star_k,\\
E_{P,k} &= E_k - k^2C_kz^{2k},\\
E^\star_{P,k} &= E^\star_k - k^2C_kz^{2k}.
\end{aligned}$$

The last two equations follow from the fact that

$$\mathsf{e}((\sigma_{i_1} + \cdots + \sigma_{i_k})^\star) = \overset{\star}{\mathsf{e}}((\sigma_{i_1} + \cdots + \sigma_{i_k})^\star) = k^2,$$

for any permutation $\sigma_{i_1}, \ldots, \sigma_{i_k}$ of Σ. The cost function $\mathsf{t}(\alpha) = \mathsf{f}(\alpha) + \mathsf{e}(\alpha)$ computes the number of transitions in the Glushkov automaton of α. The generating function associated to t is given by $T_k(z) = F_k(z) + E_k(z)$. Setting $w = T_k(z)$, one has

$$c_2w^2 + c_1w + c_0 = 0,$$

where the $c_i = c_i(k, z)$. Therefore,

$$w = \frac{-c_1 \pm \sqrt{c_1^2 - 4c_0c_2}}{2c_2}.$$

Now, one can see that $c_1 = \Delta_k s_k$, $c_2 = \Delta_k a_k b_k^2$ and $c_1^2 - 4c_0c_2 = k^2\Delta_k q_k^2$, for some polynomials[b] $a_k, b_k, q_k \in \mathbb{Q}[z]$. From this it follows that

$$w = -\frac{s_k}{2a_k b_k^2} \pm \frac{kq_k}{2a_k b_k^2\sqrt{\Delta_k}}.$$

With η_k as defined in p.7, one can now deduce, as above, that

$$T_k(z) \underset{z\,\to\,\eta_k}{\sim} \frac{kq_k(\eta_k)}{2a_k(\eta_k)b_k(\eta_k)^2\sqrt{\psi_k(\eta_k)}}\left(1-\frac{z}{\eta_k}\right)^{-\frac{1}{2}},$$

and therefore

$$[z^n]T_k(z) \underset{n}{\sim} \frac{kq_k(\eta_k)}{2\sqrt{\pi}a_k(\eta_k)b_k(\eta_k)^2\sqrt{\psi_k(\eta_k)}}\,\eta_k^{-n}n^{-\frac{1}{2}}.$$

From all this, one gets:

$$\frac{[z^n]T_k(z)}{[z^n]R_k(z)} \underset{n}{\sim} \frac{4k\eta_k q_k(\eta_k)}{a_k(\eta_k)b_k(\eta_k)^2\psi_k(\eta_k)}\,n.$$

With the help of a symbolic and numeric computing system one can explicitly find out the polynomials a_k, b_k, q_k, and then reducing them modulo Δ_k (which has η_k as a root), and then using Lemma 5 and (21), one obtains:

$$a_k(\eta_k) \underset{k}{\sim} \frac{1}{2}k\eta_k \quad ; \quad b_k(\eta_k) \underset{k}{\sim} \frac{1}{8}k\eta_k \quad ; \quad q_k(\eta_k) \underset{k}{\sim} \frac{1}{2048}k.$$

This yields

$$\lim_{k\to\infty} \frac{4k\eta_k q_k(\eta_k)}{a_k(\eta_k)b_k(\eta_k)^2\psi_k(\eta_k)} = 1.$$

We have thus obtained the following result.

Theorem 9. *For expressions of size n over an alphabet of size k, the number of transitions in the Glushkov automaton for regular expressions, without Θ in unions, is asymptotically, with respect to n, given by $\lambda_k n$, where $\lim_{k\to\infty} \lambda_k = 1$.*

To grasp the progression of λ_k, observe that $\lambda_2 = 4.03$, $\lambda_5 = 2.91$, $\lambda_{10} = 2.30$, $\lambda_{10} = 1.89$, $\lambda_{50} = 1.54$, $\lambda_{100} = 1.38$, $\lambda_{10000} = 1.03$. Theorems 8 and 9 show that the size of the Glushkov automaton, both in states and transitions, is, on average and asymptotically, independent of whether we consider all regular expressions or the restricted set R mentioned by Koechlin et al. [7].

[b]These polynomials are quite large, *e.g.* q_k has 437 monomials and degree $10 + 28k$.

6. Conclusions

We consider a set of regular expressions R that avoids a given absorbing pattern Θ and that is significantly smaller than the set of standard regular expressions, RE. Nevertheless, the on average asymptotic estimates for several complexity measures remain the same. Some experiments also corroborate those results. Using samples of uniformly random generated expressions $\alpha \in \mathsf{R}$ for small values of the alphabet size and of the tree-size of the expressions the average values for the same complexity measures coincide with the ones for expressions in RE. We conclude that, despite the important conclusions on the expressivity of regular expressions obtained by Koechlin et al. [7], the usage of the analytic combinatorics framework remains an essential tool to study the descriptional complexity, on average, of algorithms on regular expressions.

References

[1] S. Broda, A. Machiavelo, N. Moreira and R. Reis, On the average size of Glushkov and partial derivative automata, *Int. J. Found. Comput. Sci.* **23**(5) (2012) 969–984.
[2] S. Broda, A. Machiavelo, N. Moreira and R. Reis, On average behaviour of regular expressions in strong star normal form, *Int. J. Found. Comput. Sci.* **30**(6-7) (2019) 899–920.
[3] S. Broda, A. Machiavelo, N. Moreira and R. Reis, Analytic combinatorics and descriptional complexity of regular languages on average, *ACM SIGACT News* **51** (March 2020) 38–56.
[4] B. Buchberger, Gröbner bases: A short introduction for systems theorists, *Computer Aided Systems Theory - EUROCAST*, (2001), pp. 1–19.
[5] P. Flajolet and R.Sedgewick, *Analytic Combinatorics* (CUP, 2008).
[6] V. M. Glushkov, The abstract theory of automata, *Russian Math. Surveys* **16**(5) (1961) 1–53.
[7] F. Koechlin, C. Nicaud and P. Rotondo, Uniform random expressions lack expressivity, *44th MFCS 2019*, eds. P. Rossmanith, P. Heggernes and J. Katoen *LIPIcs* **138** (2019), pp. 51:1–51:14.
[8] F. Koechlin, C. Nicaud and P. Rotondo, On the degeneracy of random expressions specified by systems of combinatorial equations, *24th DLT 2020*, eds. N. Jonoska and D. Savchuk *LNCS* **12086**, (Springer, 2020), pp. 164–177.
[9] S. Konstantinidis, A. Machiavelo, N. Moreira and R. Reis, On the size of partial derivatives and the word membership problem, *Acta Informatica* **58**(4) (2021) 357–375.
[10] C. Nicaud, On the average size of Glushkov's automata, *3rd LATA*, eds. A. Dediu, A.-M. Ionescu and C. M. Vide *LNCS* **5457**, (Springer, 2009), pp. 626–637.
[11] C. Nicaud, Random deterministic automata, *MFCS 2014*, eds. E. Csuhaj-Varjú, M. Dietzfelbinger and Z. Ésik *LNCS* **8634**, (Springer, 2014), pp. 5–23.
[12] C. Nicaud and P. Rotondo, Random regular expression over huge alphabets, *Int. J. Found. Comput. Sci.* **32**(5) (2021) 419–438.
[13] B. Simon, *Basic Complex Analysis* (American Mathematical Society, 2015).