

## The Average State Complexity of Rational Operations on Finite Languages\*

Frédérique Bassino, LIPN UMR CNRS 7030, Université Paris 13, 93430 Villetaneuse, France.

Laura Giambruno, Dipartimento di Matematica e Applicazioni, Università di Palermo, Italy.

Cyril Nicaud, LIGM UMR CNRS 8049, Université Paris-Est, 77454 Marne-la-Vallée, France.

Considering the uniform distribution on sets of  $m$  non-empty words whose sum of lengths is  $n$ , we establish that the average state complexities of the rational operations are asymptotically linear.

### 1. Introduction

This paper addresses first and foremost the following issue: given a finite alphabet  $A$ , a set of words  $X \subset A^*$ , and a word  $u \in A^*$ , we want to efficiently determine whether  $u \in X^*$ .

With a nondeterministic automaton, one can determine whether a word  $u$  is in  $X^*$  in time proportional to the product of the lengths of  $u$  and  $X$ , where the length of  $X$  is defined as the sum of the lengths of its elements.

With a deterministic automaton recognizing  $X^*$ , one can check whether a word  $u$  is in  $X^*$  in time proportional to the length of  $u$ , once the automaton has been built. The problem here is that the number of states of the automaton can be very large: in [6], Ellul, Krawetz, Shallit and Wang found an example where the state complexity (*i.e.*, the number of states of the minimal automaton) of  $X^*$  is exponential. More precisely, for every integer  $h \geq 3$ , they gave a language  $X_h$  of length  $\Theta(h^2)$ , containing  $\Theta(h)$  words, whose state complexity is  $\Theta(h2^h)$ . Using another measure on finite sets of words, Campeanu, Culik, Salomaa and Yu proved in [3, 4] that if the set  $X$  is a finite language over an alphabet of at least three letters having state complexity  $n \geq 4$ , the state complexity of  $X^*$  is  $2^{n-3} + 2^{n-4}$  in the worst case. In addition when  $X$  is not necessarily finite, the state complexity of  $X^*$  is  $2^{n-1} + 2^{n-2}$  in the worst case [16, 17].

An efficient alternative solution based on the Aho-Corasick automaton was proposed in [5] by Clément, Duval, Guaiana, Perrin and Rindone. In their paper, they describe an algorithm to compute all the decompositions of a word as a concatenation of elements in a finite set of non-empty words.

Our paper is a contribution to the general problem, called the non-commutative Frobenius problem by Kao, Shallit and Xu [12]. This name is originated by its relation with the classical problem [10, 11] of which it is a generalization. We present

\*The authors acknowledge partial support from the ESF program AUTOMATHA. The first and third authors were supported by the ANR GAMMA (project BLAN07-2\_195422).

an average case analysis of this problem. More precisely we analyze the average state complexity of  $X^*$ , when all sets of  $m$  non-empty words and total length  $n$  are equally likely to occur. In this setting, we let  $n$  tend to infinity. The analysis relies on the general framework of analytic combinatorics [7] applied to sets of words, and on classical automata constructions.

Our main result is that, in average, the state complexity of the star  $X^*$  of a set  $X$  of  $m$  non-empty words is linear with respect to the length of  $X$ . For an alphabet with at least three letters, we slightly improve this result, and show that the average state complexity of  $X^*$  is asymptotically equivalent to  $n$ .

As a natural extension of this result, we also propose an average-case analysis of the union and the concatenation, the two other rational operations, for finite languages. In both cases we establish the linearity, in average, of the state complexity.

The probability distribution chosen in this article is quite natural for the problem. Indeed, the sum of the lengths of the words can be seen as the size of a finite language since it corresponds to the space needed for its direct representation (*i.e.* by listing its elements). If we remove the condition that the number of words is fixed, and consider the uniform distribution on finite languages of length  $n$ , the probability that a random language contains small words is very high. More precisely, all the words of length one, are contained in a random set with a non-negligible probability. Since our main focus is the star operation, this latter distribution is not very interesting, because the probability that the star of a random set is  $A^*$  (with state complexity one) is then too high.

It is important to remark that an interesting and rather different distribution has been introduced in [8]. For a given  $n$ , Gruber and Holzer consider the uniform distribution over finite languages whose longest word is of length at most  $n$ . The probability distribution is quite different from ours. For instance, there are  $2^{(|A|^{n+1}-1)/(|A|-1)}$  distinct sets of size  $n$ , where we have around  $\binom{n-1}{m-1}|A|^n$  sets. For the probability distribution considered in [8], it is likely to have a large amount of words of large size, and the authors proved that almost all languages have a state complexity in  $\Theta(|A|^n/n)$ . On the other side, for the distribution considered in this article, the average state complexity of a language of length  $n$  is asymptotically equivalent to  $n$ , as we shall see in Proposition 9 (p.8).

The paper is organized as follows. In Section 2 we recall some definitions, usual automata constructions and establish some technical combinatorial properties about words. In Section 3, we prove lower bounds for rational operations on finite languages, in the average case. The average state complexities are established in Section 4 for the union and the concatenation, and in Section 5 for the star operation. Finally, some algorithmic perspectives are discussed in Section 6.

A preliminary version of this work has been presented in [1].

## 2. Preliminary

### 2.1. Automata and Words

We recall some definitions about automata and combinatorics on words. We refer the readers to [9, 15, 2] for elements of theory of finite automata and to [13, 14, 15] for combinatorics on words.

A *finite automaton*  $\mathcal{A}$  over a finite alphabet  $A$  is a quintuple  $\mathcal{A} = (A, Q, T, I, F)$  where  $Q$  is a finite set of *states*,  $T \subset Q \times A \times Q$  is the set of *transitions*,  $I \subset Q$  is the set of *initial states* and  $F \subset Q$  is the set of final states. The automaton  $\mathcal{A}$  is *deterministic* if it has only one initial state and for any  $(p, a) \in Q \times A$  there exists at most one state  $q \in Q$  such that  $(p, a, q) \in T$ . It is *complete* if for each  $(p, a) \in Q \times A$ , there exists at least one state  $q \in Q$  such that  $(p, a, q) \in T$ . A deterministic finite automaton  $\mathcal{A}$  is *accessible* when for each state  $q$  of  $\mathcal{A}$ , there exists a path from the initial state to the state  $q$ . The *size*  $\#\mathcal{A}$  of an automaton  $\mathcal{A}$  is its number of states. Any finite automaton  $\mathcal{A} = (A, Q, T, I, F)$  can be transformed into a deterministic automaton  $\mathcal{B} = (A, \mathcal{P}(Q), T', \{I\}, F')$  recognizing the same language and in which  $F' = \{P \in \mathcal{P}(Q) \mid P \cap F \neq \emptyset\}$  and  $T' = \{(P, a, R) \text{ with } P \in \mathcal{P}(Q), a \in A \text{ and } R = \{q \mid \exists p \in P, (p, a, q) \in T\}\}$ . In practice only the accessible part of the automaton  $\mathcal{B}$  is built in this *subset construction*.

We say that the word  $v$  is a *proper prefix* (resp. *suffix*) of a word  $u$  if  $v$  is a prefix (resp. suffix) of  $u$  such that  $v \neq \varepsilon$  and  $v \neq u$ . The word  $v$  is called a *border* of  $u$  if  $v$  is both a proper prefix and a proper suffix of  $u$ . We denote by  $\text{Pr}(u)$  (resp.  $\text{Sf}(u)$ ) the set of all prefixes (resp. suffixes) of  $u$ , by  $\text{Pref}(u)$  (resp.  $\text{Suff}(u)$ ) the set of proper prefixes (resp. suffixes) and by  $\text{Bord}(u)$  the set of borders of  $u$ . A word is *primitive* when it is not the power of another word. Let  $u, v$  and  $w$  be three non-empty words such that  $w$  is a proper suffix of  $v$ ,  $v$  is a proper suffix of  $u$  and define the following sets:  $Q_u = \{\{u\} \cup P \mid P \subset \text{Suff}(u)\}$ ,  $Q_{u,v} = \{\{u\} \cup P \mid P \in Q_v\}$  and  $Q_{u,v,w} = \{\{u\} \cup P \mid P \in Q_{v,w}\}$ . The cardinalities of  $Q_u$ ,  $Q_{u,v}$  and  $Q_{u,v,w}$  are respectively equal to  $2^{|u|-1}$ ,  $2^{|v|-1}$  and  $2^{|w|-1}$ .

The *minimal automaton* of a regular language is the unique (up to isomorphism) smallest accessible and deterministic automaton recognizing this language. The *state complexity* of a regular language is the size of its minimal automaton. Moreover the state complexity of a regular language  $L$  is equal to its number of distinct left quotients, *i.e.* the languages of the form  $u^{-1}L = \{w \in A^* \mid uw \in L\}$ . Let  $L \subset A^*$  be a finite set of words. The automaton  $\mathcal{T}_L = (A, \text{Pr}(L), T_L, \{\varepsilon\}, L)$ , where  $T_L = \{(u, a, ua) \mid u \in \text{Pr}(L), a \in A, ua \in \text{Pr}(L)\}$ , recognizes the set  $L$  (see Figure 1 p.14 for an example). Therefore the state complexity of a finite language, whose sum of the lengths of its elements is  $n$ , is less than or equal to  $n + 1$ .

### 2.2. Enumeration

Recall that  $f(n) = \mathcal{O}(g(n))$  if there exist a positive real  $c$  and a nonnegative integer  $n_0$  such that, for all  $n \geq n_0$ ,  $|f(n)| \leq c|g(n)|$ . We write  $f(n) = \Omega(g(n))$  if there exist

a positive real  $c$  and a nonnegative integer  $n_0$  such that, for all  $n \geq n_0$ ,  $|f(n)| \geq c|g(n)|$ . And we say that  $f(n) = \Theta(g(n))$  if  $f(n) = \mathcal{O}(g(n))$  and  $f(n) = \Omega(g(n))$ .

Let  $X \subset A^*$  be a finite set of words. We denote by  $|X|$  the cardinality of  $X$  and by  $\|X\|$  the *length* of  $X$  defined as the sum of the lengths of its elements:  $\|X\| = \sum_{u \in X} |u|$ . Let  $\mathcal{S}et_{n,m}$  be the set of sets of  $m$  non-empty words whose sum of lengths is  $n$ :

$$\mathcal{S}et_{n,m} = \{X = \{u_1, \dots, u_m\} \mid \|X\| = n, \forall i \in \{1, \dots, m\} u_i \in A^+\}$$

and  $\mathcal{S}_{n,m}$  be the set of sequences of  $m$  non-empty words whose sum of lengths is  $n$ :

$$\mathcal{S}_{n,m} = \{S = (u_1, \dots, u_m) \mid \|S\| = n, \forall i \in \{1, \dots, m\} u_i \in A^+\}.$$

We denote by  $\mathcal{S}_{n,m}^\neq \subset \mathcal{S}_{n,m}$  the set of sequences of pairwise distinct words.

**Proposition 1.** *For any fixed integer  $m \geq 2$ , the number  $|\mathcal{S}_{n,m}|$  of sequences of  $m$  non-empty words whose sum of lengths is  $n$  satisfies the following equality:*

$$|\mathcal{S}_{n,m}| = \binom{n-1}{m-1} |A|^n, \quad (1)$$

and the number  $|\mathcal{S}et_{n,m}|$  of sets of  $m$  non-empty words whose sum of lengths is  $n$  satisfies the following one:

$$|\mathcal{S}et_{n,m}| = \frac{1}{m!} |\mathcal{S}_{n,m}| \left(1 + \mathcal{O}\left(\frac{1}{n^2}\right)\right). \quad (2)$$

**Proof.** Any sequence  $S$  of  $\mathcal{S}_{n,m}$  can be uniquely defined by a word  $v$  of length  $n$ , which is the concatenation of the elements of  $S$ , and a composition of  $n$  into  $m$  parts, that indicates how to cut the word of length  $n$  into  $m$  parts. Therefore  $|\mathcal{S}_{n,m}| = \binom{n-1}{m-1} |A|^n$ . Moreover, since  $m$  is fixed,

$$\binom{n-1}{m-1} |A|^n \sim \frac{n^{m-1}}{(m-1)!} |A|^n. \quad (3)$$

Let  $\mathcal{F}_{n,m}$  be the set of the elements  $S = (u_1, \dots, u_m)$  of  $\mathcal{S}_{n,m}$  such that  $u_1 = u_2$ , then we will prove that

$$|\mathcal{F}_{n,m}| = |\mathcal{S}_{n,m}| \mathcal{O}\left(\frac{1}{n^2}\right). \quad (4)$$

Indeed, if  $m = 2$  then  $|\mathcal{F}_{n,2}| = \begin{cases} 0 & \text{if } n \text{ is odd} \\ |A|^{n/2} & \text{if } n \text{ is even} \end{cases}$  which proves the result.

If  $m \geq 3$ , the generating function for the number of pairs of non-empty words  $(u, v)$  such that  $u = v$  is  $z \mapsto \frac{|A|z^2}{1-|A|z^2}$ , and

$$F_m(z) = \sum_{n \geq 0} \mathcal{F}_{n,m} z^n = \frac{|A|z^2}{1-|A|z^2} S_{m-2}(z) = \frac{|A|z^2}{1-|A|z^2} \left(\frac{|A|z}{1-|A|z}\right)^{m-2},$$

where  $S_{m-2}(z) = \sum_{n \geq 0} \mathcal{S}_{n,m-2} z^n$ . Therefore  $F_m(z)$  is a rational function with a simple pole at  $\frac{1}{\sqrt{|A|}}$  and a pole of order  $m-2$  at  $\frac{1}{|A|}$ . Hence there exist a polynomial  $P$  of degree  $m-3$  and a constant  $c$ , such that

$$F_{n,m} = P(n)|A|^n + c|A|^{n/2} = P(n)|A|^n \left(1 + \mathcal{O}(|A|^{-n/2})\right).$$

Equation (4) is then obtained using Equation (3) and the degree of  $P$ .

For the general case let  $i, j \in \{1, \dots, m\}$  and denote by  $\mathcal{F}_{n,m}^{(i,j)}$  the subset of  $\mathcal{S}_{n,m}$  containing all sequences  $(u_1, \dots, u_m)$  such that  $u_i = u_j$ . Then the following equation holds

$$\mathcal{S}_{n,m} = \mathcal{S}_{n,m}^{\neq} \cup \bigcup_{1 \leq i < j \leq m} \mathcal{F}_{n,m}^{(i,j)}$$

where  $\mathcal{S}_{n,m}^{\neq} \subset \mathcal{S}_{n,m}$  is the subset of sequences whose elements are pairwise distinct. By the symmetry of the problem  $|\mathcal{F}_{n,m}^{(i,j)}| = |\mathcal{F}_{n,m}|$  and consequently  $|\mathcal{S}_{n,m}| - |\mathcal{S}_{n,m}^{\neq}| \leq \binom{m}{2} |\mathcal{F}_{n,m}|$ . Hence by Equation (4),  $|\mathcal{S}_{n,m}| - |\mathcal{S}_{n,m}^{\neq}| = |\mathcal{S}_{n,m}| \mathcal{O}\left(\frac{1}{n^2}\right)$ .

Finally, to complete the proof, since an element in  $\text{Set}_{n,m}$  is mapped on exactly  $m!$  sequences in  $\mathcal{S}_{n,m}^{\neq}$ , we obtain  $|\mathcal{S}_{n,m}^{\neq}| = m! |\text{Set}_{n,m}|$ .  $\square$

In the sequel we shall count the number of states of automata according to their labels. This enumeration is based on combinatorial properties of words.

**Lemma 2.** *Let  $u$  be a non-empty word of length  $\ell$ . The number of sequences  $S$  belonging to  $\mathcal{S}_{n,m}$  such that  $u$  is a prefix (resp. suffix) of a word of  $S$  is smaller than or equal to  $m \binom{n-\ell}{m-1} |A|^{n-\ell}$ .*

**Proof.** There are at most  $m \binom{n-\ell-1}{m-2} |A|^{n-\ell}$  elements in  $\mathcal{S}_{n,m}$  containing  $u$ . This can be seen by taking an element of  $\mathcal{S}_{n-\ell, m-1}$  and adding  $u$  at one of the  $m$  possible places. This construction covers all the possibilities (with some over-counting by repetitions). Moreover there are at most  $m \binom{n-\ell-1}{m-1} |A|^{n-\ell}$  elements in  $\mathcal{S}_{n,m}$  containing a word having  $u$  as a prefix (resp. suffix), since taking an element of  $\mathcal{S}_{n-\ell, m}$  and concatenating  $u$  at the beginning (resp. end) of one of the words covers all the possibilities (with some over-counting). We conclude the proof by noticing that  $m \binom{n-\ell-1}{m-2} |A|^{n-\ell} + m \binom{n-\ell-1}{m-1} |A|^{n-\ell} = m \binom{n-\ell}{m-1} |A|^{n-\ell}$ .  $\square$

**Lemma 3.** *Let  $u, v \in A^+$  such that  $v$  is not a prefix of  $u$ ,  $|u| = \ell$  and  $|v| = i$ . The number of sequences  $S \in \mathcal{S}_{n,m}$  such that both  $u$  and  $v$  are prefixes of words of  $S$  is smaller than or equal to  $m(m-1) |A|^{n-\ell-i} \binom{n-\ell-i+1}{m-1}$ .*

**Proof.** It is similar to the proof of Lemma 2. We have to distinguish four cases:

- $u$  and  $v$  are proper prefixes of words in  $S$ ,
- $u \in S$  and  $v$  is a proper prefix,
- $v \in S$  and  $u$  is a proper prefix,
- and both  $u$  and  $v$  are in  $S$ .

One can upper bound the number of sequences for these different cases by  $m(m-1)|S_{n-\ell-i,m}|$ ,  $m(m-1)|S_{n-\ell-i,m-1}|$ ,  $m(m-1)|S_{n-\ell-i,m-1}|$  and  $m(m-1)|S_{n-\ell-i,m-2}|$  respectively. We conclude by noticing that

$$\binom{n-\ell-i-1}{m-1} + 2\binom{n-\ell-i-1}{m-2} + \binom{n-\ell-i-1}{m-3} = \binom{n-\ell-i+1}{m-1} \quad \square$$

In the following lemmas, we establish important properties that link a word and its borders.

**Lemma 4.** *For  $1 \leq i < \ell$ , there are at most  $|A|^{\ell-i}$  pairs of non-empty words  $(u, v)$  such that  $|u| = \ell$ ,  $|v| = i$  and  $v$  is a border of  $u$ .*

**Proof.** Since  $v$  is a border of  $u$ , then  $\ell - i$  is a period of  $u$  (see [14] p.270). Therefore the  $\ell - i$  first letters of  $u$  completely define  $u$  and  $v$ , hence there are at most  $|A|^{\ell-i}$  possible pairs.  $\square$

**Lemma 5.** *For  $1 \leq j < i < \ell$  such that  $i \leq \frac{2}{3}\ell$  or  $j \leq \frac{i}{2}$ , there are at most  $|A|^{\ell-\frac{i}{2}-j}$  triples of non-empty words  $(u, v, w)$  with  $|u| = \ell$ ,  $|v| = i$ ,  $|w| = j$  such that  $v$  is a border of  $u$  and  $w$  is a border of  $v$ .*

**Proof.** If  $i \leq \frac{\ell}{2}$ , since  $w$  is a border of  $v$ , there are at most  $|A|^{i-j}$  possible pairs  $(v, w)$  satisfying the conditions of the lemma. Since  $v$  is a border of  $u$  and  $\ell \geq 2i$ ,  $u$  can be defined with only  $\ell - 2i$  letters for fixed  $v$ . There are hence at most  $|A|^{\ell-i-j} < |A|^{\ell-\frac{i}{2}-j}$  possible triples  $(u, v, w)$ .

When  $i > \frac{\ell}{2}$  and  $j \leq \frac{i}{2}$ , since  $v$  is a border of  $u$ , there are at most  $|A|^{\ell-i}$  possible pairs  $(u, v)$ . Since  $-j \geq -\frac{i}{2}$  we get  $\ell - i \leq \ell - \frac{i}{2} - j$  and  $|A|^{\ell-i} \leq |A|^{\ell-\frac{i}{2}-j}$ .

Finally when  $\frac{\ell}{2} < i \leq \frac{2}{3}\ell$  and  $j > \frac{i}{2}$ , since  $w$  is a border of  $v$ , there are at most  $|A|^{i-j}$  possible pairs  $(v, w)$ . Since  $\ell - i$  is a period of  $u$ ,  $v$  completely define  $u$ . And since  $i - j \leq \ell - \frac{i}{2} - j$ , there are at most  $|A|^{\ell-\frac{i}{2}-j}$  possible triples  $(u, v, w)$ .  $\square$

**Proposition 6.** *For  $1 \leq j < i < \ell$  such that  $i > \frac{2}{3}\ell$  and  $j > \frac{i}{2}$  and for any triple of words  $(u, v, w)$  with  $|u| = \ell$ ,  $|v| = i$ ,  $|w| = j$  such that  $v$  is a border of  $u$  and  $w$  is a border of  $v$ , there exist a primitive word  $x$ , with  $1 \leq |x| \leq \ell - i$ , a prefix  $x_0$  of  $x$  and integers  $p > q > s > 0$  such that  $u = x^p x_0$ ,  $v = x^q x_0$  and  $w = x^s x_0$ .*

**Proof.** Since  $v$  is a border of  $u$ , then  $\ell - i$  is a period of  $u$ . Let  $x$  be the unique primitive word such that  $x^k$  is the prefix of  $u$  of length  $\ell - i$ , for some positive integer  $k$ . Then there exist a prefix  $x_0$  of  $x$  and a positive integer  $p$  such that  $u = x^p x_0$ . Since  $v$  is a suffix of  $u$  of length  $i$ , then  $v = x^{p-k} x_0$ . And since  $\ell - i < i$ , then  $p - k > 0$ . Since  $w$  is a prefix of  $v$  and  $\ell - i < \frac{i}{2} < j$ , then  $w = x^s x_1$  where  $s > 0$  and  $x_1$  is a prefix of  $x$ .

It remains to prove that  $x_1 = x_0$ . Since  $w$  is a suffix of  $v$ , then there exist a suffix  $x_2$  of  $x$  and  $r \geq 0$  such that  $w = x_2 x^r x_0$ . If  $x_2$  is empty, then the result follows. Otherwise  $x^r x_0$  is a border of  $w$ , then  $w$  is a power of  $x_2$  and  $x$  is a power of  $x_2$ . But

$x$  cannot be an integral power of  $x_2$  since it is primitive. Therefore  $x = x_2^t x'_2$  where  $t > 0$  and  $x'_2$  is a prefix of  $x_2$ . Since  $x_2$  is a suffix of  $x$ , then there exists a proper suffix  $x''_2$  of  $x_2$  such that  $x_2 = x''_2 x'_2$ . Since  $x'_2$  is a prefix of  $x_2$ , then  $x_2 = x'_2 x''_2$ . And since  $x''_2 x'_2 = x'_2 x''_2$ , then  $x'_2$  and  $x''_2$  are integral powers of a same word [13]. This leads to a contradiction with the fact that  $x$  is primitive.  $\square$

### 3. Lower Bounds

We first introduce the subsets  $\mathcal{S}_{n,m}^{(p)}$  of  $\mathcal{S}_{n,m}$  that will be used to establish average lower bound results. Let  $n \geq (2p+1)m$ , denote by  $\mathcal{S}_{n,m}^{(p)}$  the set of tuples  $(u_1, \dots, u_m)$  belonging to  $\mathcal{S}_{n,m}^{(p)}$  and such that, for every  $i \in \{1, \dots, m\}$ ,  $|u_i| > 2p$  and whose prefixes (resp. suffixes) of length  $p$  of the  $u_i$  are pairwise disjoint. Note that the set of words defined by a sequence in  $\mathcal{S}_{n,m}^{(p)}$  is a bifix (*i.e.*, prefix and suffix) code.

We prove next that almost all sequences of  $\mathcal{S}_{n,m}$  are in  $\mathcal{S}_{n,m}^{(\lfloor \log n \rfloor)}$ , and that the state complexity of the set associated to a sequence in  $\mathcal{S}_{n,m}^{(\lfloor \log n \rfloor)}$  is asymptotically equivalent to  $n$ .

**Lemma 7.** *For any fixed  $m \geq 1$ ,  $|\mathcal{S}_{n,m}^{(p)}| \sim |\mathcal{S}_{n,m}|$  since  $n \rightarrow \infty$  and  $p \rightarrow \infty$ , with  $p = o(n)$ .*

**Proof.** Let  $\mathcal{P}_m^{(p)}$  be the set of sequences of  $m$  distinct words of length  $p$ . Since  $p$  tends to infinity, the cardinality of  $\mathcal{P}$  satisfies

$$|\mathcal{P}_m^{(p)}| = |A|^p (|A|^p - 1) \cdots (|A|^p - m + 1) \sim |A|^{mp}.$$

By separating prefixes and suffixes of length  $p$  in elements of  $\mathcal{S}_{n,m}^{(p)}$ , we find that  $|\mathcal{S}_{n,m}^{(p)}| = |\mathcal{P}_m^{(p)}| \times |\mathcal{S}_{n-2p,m}| \times |\mathcal{P}_m^{(p)}|$ . As a consequence, when  $n$  and  $p$  tend to infinity with  $p = o(n)$ , we have:

$$|\mathcal{S}_{n,m}^{(p)}| \sim |A|^{mp} \binom{n-2mp-1}{m-1} |A|^{n-2mp} |A|^{mp} \sim \binom{n-1}{m-1} |A|^n.$$

The lemma then follows by Equation (1) of Proposition 1.  $\square$

**Lemma 8.** *For any sequence  $S$  in  $\mathcal{S}_{n,m}$ , every singleton  $\{v\}$ , with  $v$  a suffix of a word in  $S$ , is a left quotient of the finite language associated to  $S$ . Moreover, if  $S \in \mathcal{S}_{n,m}^{(p)}$  then there are at least  $n - 2pm$  such suffixes. As a consequence, the state complexity of  $S$  is at least  $n - 2pm$ .*

**Proof.** Let  $v, v'$  with  $v \neq v'$  be two prefixes of the words  $u \in S$  and  $u' \in S$  respectively, such that  $p < |v| \leq |u| - p$  and  $p < |v'| \leq |u'| - p$ . Let  $w$  and  $w'$  be the suffixes associated to  $v$  and  $v'$  respectively (*i.e.*  $u = vw$  and  $u' = v'w'$ ). We claim that  $w \neq w'$ . Indeed, if  $w = w'$  then  $u = u'$ , since the suffixes of length  $p$  of two distinct words in  $S$  are distinct and  $|w| > p$ . Hence  $v = v'$  since they are both prefixes of length  $|u| - |w|$  of  $u$ . Therefore,  $v^{-1}S = \{w\}$ ,  $v'^{-1}S = \{w'\}$  and  $\{w\} \neq \{w'\}$ . As a consequence, all the left quotients of  $S$  defined by such prefixes

are distinct. This concludes the proof since there are  $n - 2pm$  such prefixes of words in  $S$ .  $\square$

The proof of the following result is a direct consequence of Lemma 7, Lemma 8 and by Equation (2) of Proposition 1:

**Proposition 9.** *For any fixed  $m \geq 1$ , the average state complexity of an element in  $Set_{n,m}$  is asymptotically equivalent to  $n$  as  $n$  tends to infinity.*

**Proposition 10 (Union)** *For the uniform distribution over the pairs  $(X_1, X_2)$  of  $Set_{n_1, m_1} \times Set_{n_2, m_2}$ , the average state complexity of  $X_1 \cup X_2$  is bounded below by a function equivalent to  $n_1 + n_2$  when both  $n_1$  and  $n_2$  tend to infinity.*

**Proof.** It is sufficient to prove the result for pairs of sequences. The property can then be extended to pairs of sets using Equation (2) of Proposition 1. Assume without loss of generality that  $n_1 \leq n_2$  and consider the subset  $\mathcal{X} \subset \mathcal{S}_{n_1, m_1}^{(p)} \times \mathcal{S}_{n_2, m_2}^{(p)}$ , with  $p = \lfloor \log n_1 \rfloor$ , defined by

$$\mathcal{X} = \{(X_1, X_2) \in \mathcal{S}_{n_1, m_1}^{(p)} \times \mathcal{S}_{n_2, m_2}^{(p)} \mid X_1 \cup X_2 \in \mathcal{S}_{n_1+n_2, m_1+m_2}^{(p)}\}.$$

In other words, all prefixes (resp. suffixes) of length  $p$  of words either in  $X_1$  or in  $X_2$  are distinct.

For any fixed  $X_1 \in \mathcal{S}_{n_1, m_1}^{(p)}$ , the number of sequences  $X_2 \in \mathcal{S}_{n_2, m_2}^{(p)}$  such that  $(X_1, X_2) \in \mathcal{X}$  is asymptotically equal to  $|\mathcal{S}_{n_2, m_2}^{(p)}|$ . This can be established with similar arguments as those presented in the proof of Lemma 7. And applying Lemma 7, we obtain  $|\mathcal{X}| \sim |\mathcal{S}_{n_1, m_1}^{(p)}| \cdot |\mathcal{S}_{n_2, m_2}^{(p)}| \sim |\mathcal{S}_{n_1, m_1}| \cdot |\mathcal{S}_{n_2, m_2}|$ . Moreover for every  $(X_1, X_2) \in \mathcal{X}$ , we have  $X_1 \cup X_2 \in \mathcal{S}_{n_1+n_2-2, m_1+m_2}^{(p)}$ . Therefore, by Lemma 8 the state complexity of  $X_1 \cup X_2$  is at least  $n_1 + n_2 - 2(m_1 + m_2)\lfloor \log n_1 \rfloor$ . This concludes the proof since this inequality holds for almost all pairs of sequences.  $\square$

**Proposition 11 (Concatenation)** *For the uniform distribution over the pairs  $(X_1, X_2)$  of  $Set_{n_1, m_1} \times Set_{n_2, m_2}$ , the average state complexity of  $X_1 \cdot X_2$  is bounded below by a function equivalent to  $n_1 + n_2$ , when both  $n_1$  and  $n_2$  tend to infinity.*

**Proof.** As for the union construction, it is sufficient to establish the result for pairs of sequences. Let  $X_1 \in \mathcal{S}_{n_1, m_1}^{(\lfloor \log n_1 \rfloor)}$  and  $X_2 \in \mathcal{S}_{n_2, m_2}^{(\lfloor \log n_2 \rfloor)}$ . Assume first that  $m_2 = 1$ , and that  $X_2 = (x)$ . The left quotients of  $X_1 \cdot X_2$  are either of the form  $\{vx\}$ , with  $v$  a suffix of a word in  $X_1$  or  $\{v\}$ , with  $v$  a suffix of  $x$ . By Lemma 8 there are at least  $n_1 - 2m_1 \log n_1 + n_2 + 1$  such classes.

Assume now that  $m_2 \geq 2$ . Let  $u$  be an element of  $X_1$ . Since  $X_1$  is a prefix code, then, for any word  $v \in A^*$ ,  $uv \in X_1 \cdot X_2$  if and only if  $v \in X_2$ . Therefore, when  $w$  ranges over all the prefixes of words in  $X_2$ , the left quotient  $(uw)^{-1}(X_1 \cdot X_2) = w^{-1}X_2$  ranges over all the left quotients of  $X_2$  and these quotients are singletons. Hence by Lemma 8 there are at least  $n_2 - 2m_2 \lfloor \log n_2 \rfloor$  distinct left quotients of  $X_1 \cdot X_2$  which are singletons.



Let  $w$  be a prefix of a word of  $X_1$  of length at least  $\lfloor \log n_1 \rfloor$ . For any  $u \in X_1$  and any  $v \in X_2$ , if  $w$  is a prefix of  $uv$  then either  $w$  is a prefix of  $u$  or  $u$  is a prefix of  $w$ . The latter case is not possible since  $X_1$  is a prefix code. Hence  $u = ws$  for some word  $s \in A^*$ , and  $w^{-1}(X_1 \cdot X_2) = s \cdot X_2$ . Let  $u = ws$  and  $u' = w's'$  be two words of  $X_1$  such that  $|s| \geq \lfloor \log n_1 \rfloor$  and  $|s'| \geq \lfloor \log n_1 \rfloor$ . If  $s \cdot X_2 = s' \cdot X_2$ , let  $y$  and  $y'$  be two elements of  $X_2$  such that  $sy = s'y'$ , then  $y = y'$  since  $X_2$  is a suffix code and consequently  $s = s'$ . So, for such suffixes  $s$ , the sets  $s \cdot X_2$  are distinct and there are at least  $n_1 - 2m_1 \lfloor \log n_1 \rfloor$  such left quotients. Since these quotients are not singleton, there are at least  $n_1 - 2m_1 \lfloor \log n_1 \rfloor + n_2 - 2m_2 \lfloor \log n_2 \rfloor$  left quotients of  $X_1 \cdot X_2$ . This concludes the proof since this inequality holds for almost all pairs of sequences.  $\square$

**Proposition 12 (Star)** *For the uniform distribution over the sets  $X$  of  $\text{Set}_{n,m}$  the average state complexity of  $X^*$  is bounded below by a function equivalent to  $n$ , when  $n$  tends to infinity.*

**Proof.** As previously we only establish the result for sequences. Recall that if  $X$  is a prefix code, then the minimal automaton of  $X$  has only one final state. Therefore the state complexity of  $X^*$  when  $X$  is a prefix code of state complexity  $n$  is either  $n$  or  $n - 1$  (see [2] Proposition 2.4, p. 95). We conclude the proof since, by Lemma 8, every  $S \in \mathcal{S}_{n,m}^{(\lfloor \log n \rfloor)}$  has a state complexity greater than  $n - 2n \lfloor \log n \rfloor$  and, by Lemma 7, almost all elements of  $\mathcal{S}_{n,m}$  belongs to  $\mathcal{S}_{n,m}^{(\lfloor \log n \rfloor)}$ .  $\square$

## 4. Average State Complexity of the Union and the Concatenation

### 4.1. Average State Complexity of the Union

Due to the structure of finite languages, it is not difficult to compute the state complexity of their union:

**Theorem 13 (Union)** *For the uniform distribution over the pairs  $(X_1, X_2)$  of  $\text{Set}_{n_1, m_1} \times \text{Set}_{n_2, m_2}$  the average state complexity of  $X_1 \cup X_2$  is equal to  $(n_1 + n_2) + \mathcal{O}(1)$  when both  $n_1$  and  $n_2$  tend to infinity.*

**Proof.** This result directly comes from the inequalities  $|X_1 \cup X_2| \leq |X_1| + |X_2|$  and  $\|X_1 \cup X_2\| \leq \|X_1\| + \|X_2\|$  together with the lower bound of Proposition 10.  $\square$

Note that the state complexity of the union is the same in the average case and in the worst case.

### 4.2. Average State Complexity of the Concatenation

In the following we prove that the average state complexity of the concatenation of two finite languages is linear in the sum of their lengths.

**Theorem 14 (Concatenation)** *For the uniform distribution over the pairs  $(X_1, X_2)$  of  $\text{Set}_{n_1, m_1} \times \text{Set}_{n_2, m_2}$ , the average state complexity of  $X_1 \cdot X_2$  is equal to  $(n_1 + n_2) + \mathcal{O}(1)$  when both  $n_1$  and  $n_2$  tend to infinity.*

Note that Proposition 11 already gives the lower bound  $(n_1 + n_2) + \mathcal{O}(1)$ . The rest of this section is devoted to the proof of the upper bound: From a nondeterministic automata recognizing  $X_1 \cdot X_2$ , we bound from above the number of states of its associated deterministic automaton obtained by the subset construction, which is greater than or equal to the state complexity of  $X_1 \cdot X_2$ .

#### 4.2.1. Construction

We associate to the finite languages  $X_1$  and  $X_2$  the automata  $\mathcal{T}_{X_1}$  and  $\mathcal{T}_{X_2}$  defined in Section 2.1. Let  $\mathcal{A}_{X_1 \cdot X_2}$  be the nondeterministic automaton defined by

$$\mathcal{A}_{X_1 \cdot X_2} = (A, (\text{Pr}(X_1) \times \{\emptyset\}) \cup (\{\emptyset\} \times \text{Pr}(X_2)), T'_{X_1} \cup T'_{X_2} \cup T, (\varepsilon, \emptyset), F),$$

where  $T'_{X_1} = \{((u, \emptyset), a, (ua, \emptyset)) \mid (u, a, ua) \in T_{X_1}\}$ ,  $T'_{X_2} = \{((\emptyset, v), a, (\emptyset, va)) \mid (v, a, va) \in T_{X_2}\}$ ,  $T = \{((u, \emptyset), a, (\emptyset, a)) \mid u \in X_1, a \in \text{Pr}(X_2)\}$  and  $F = \{\emptyset\} \times X_2$  (note that  $\varepsilon \notin X_2$ ). This automaton recognizes  $X_1 \cdot X_2$ . We denote by  $\mathcal{A}_{S \cdot T}$  the automaton defined for the set of elements of any two sequences  $S$  and  $T$  by the above construction. For any two finite sets of words  $X_1, X_2 \subset A^*$  (resp. any two sequences  $S, T$ ), we denote by  $\mathcal{D}_{X_1 \cdot X_2}$  (resp.  $\mathcal{D}_{S \cdot T}$ ) the accessible deterministic automaton obtained from the automaton  $\mathcal{A}_{X_1 \cdot X_2}$  (resp.  $\mathcal{A}_{S \cdot T}$ ) making use of the subset construction.

**Lemma 15.** *For any two finite sets of non-empty words  $X_1, X_2 \subset A^*$ , the states of the deterministic automaton  $\mathcal{D}_{X_1 \cdot X_2}$  recognizing  $X_1 \cdot X_2$  are couples  $(u, Z)$  in  $(\text{Pr}(X_1) \cup \{\emptyset\}) \times \mathcal{P}(\text{Pr}(X_2))$  and they satisfy the following properties:*

- *If  $u \in \text{Pr}(X_1)$ , there exists a unique  $Z \in \mathcal{P}(\text{Pr}(X_2))$  such that  $(u, Z)$  is a state of  $\mathcal{D}_{X_1 \cdot X_2}$ .*
- *If  $u = \emptyset$  and  $Z = \{v_1, \dots, v_\ell\}$ , then for each  $i, j$  in  $\{1, \dots, \ell\}$ , there exist  $x_i, x_j \in X_1$  and  $p_i, p_j \in X_2$  such that  $x_i p_i = x_j p_j$ . In particular, if  $v$  is the longest word in  $Z$ , for any  $i \in \{1, \dots, \ell\}$ ,  $v = w_i v_i$ , with  $w_i \in X_1^{-1} X_1$ .*

**Proof.** The first property comes from the structure of the automaton  $\mathcal{T}_{X_1}$ . Indeed, for any  $u \in \text{Pr}(X_1)$ , there is only one path from the initial state to  $u$  in  $\mathcal{T}_{X_1}$  and therefore only one path from the initial state to a state of the form  $(u, Z)$  in  $\mathcal{D}_{X_1 \cdot X_2}$ .

Let  $(\emptyset, Z)$  be a state in  $\mathcal{D}_{X_1 \cdot X_2}$ . Since  $(\emptyset, Z)$  is accessible from the initial state, for any word  $u \in Z$  there exists a path labelled by  $u$  from the initial state to  $(\emptyset, Z)$  in  $\mathcal{D}_{X_1 \cdot X_2}$ . Therefore, by construction of  $\mathcal{D}_{X_1 \cdot X_2}$ , there exist  $x \in X_1$  and  $p \in \text{Pr}(X_2)$  such that  $u = xp$ .  $\square$

We prove the result for pairs of sequences instead of pairs of sets and use Equation (2) of Proposition 1 to generalize it to pairs of sets. In the sequel we denote by

$\mathcal{S}$  the product  $\mathcal{S}_{n_1, m_1} \times \mathcal{S}_{n_2, m_2}$ . Given  $u \in A^* \cup \{\emptyset\}$ ,  $Z \in \mathcal{P}(A^*)$  and  $(S_1, S_2) \in \mathcal{S}$ , we denote by  $\mathfrak{Det}(S_1 \cdot S_2, (u, Z))$  the property:  $(u, Z)$  is the label of a state in  $\mathcal{D}_{S_1 \cdot S_2}$ .

To find an upper bound on the average number of states of the deterministic automaton  $\mathcal{D}_{S_1 \cdot S_2}$  when the sequence  $S_1$  ranges over the set  $\mathcal{S}_{n_1, m_1}$  and  $S_2$  ranges over the set  $\mathcal{S}_{n_2, m_2}$ , we count the states of all automata according to their labels. More precisely we want to estimate the sum<sup>a</sup>

$$\Delta = \sum_{(S_1, S_2) \in \mathcal{S}} \#\mathcal{D}_{S_1 \cdot S_2} = \sum_{(S_1, S_2) \in \mathcal{S}} \sum_{u \in (A^* \cup \{\emptyset\})} \sum_{Z \in \mathcal{P}(A^*)} [\mathfrak{Det}(S_1 \cdot S_2, (u, Z))].$$

Taking into account the cardinality of the labels of the states, we find

$$\begin{aligned} \Delta &= \sum_{(S_1, S_2) \in \mathcal{S}} \sum_{u \in A^*} \sum_{Z \in \mathcal{P}(Pr(X_2))} [\mathfrak{Det}(S_1 \cdot S_2, (u, Z))] \\ &\quad + \sum_{(S_1, S_2) \in \mathcal{S}} \sum_{v \in A^*} [\mathfrak{Det}(S_1 \cdot S_2, (\emptyset, \{v\}))] \\ &\quad + \sum_{(S_1, S_2) \in \mathcal{S}} \sum_{u \in A^+} \sum_{Z \subset A^*, |Z| \geq 2} [\mathfrak{Det}(S_1 \cdot S_2, (\emptyset, Z))]. \end{aligned}$$

By Lemma 15 the number of states labelled by  $(u, Z)$  with  $u \neq \emptyset$  is equal to the cardinality of  $Pr(X_1)$ , and therefore smaller than or equal to  $n_1 + 1$ . Hence, we get

$$\sum_{(S_1, S_2) \in \mathcal{S}} \sum_{u \in A^*} \sum_{Z \in \mathcal{P}(Pr(X_1))} [\mathfrak{Det}(S_1 \cdot S_2, (u, Z))] \leq \sum_{(S_1, S_2) \in \mathcal{S}} (n_1 + 1) \leq (n_1 + 1)|\mathcal{S}|.$$

Moreover, if  $(\emptyset, \{v\})$  is a label of a state, then  $v$  is in  $Pr(X_2)$ , therefore we obtain

$$\sum_{(S_1, S_2) \in \mathcal{S}} \sum_{u \in A^*} [\mathfrak{Det}(S_1 \cdot S_2, (\emptyset, v))] \leq (n_2 + 1)|\mathcal{S}|.$$

It remains to study  $\Gamma$ , with

$$\Gamma = \sum_{(S_1, S_2) \in \mathcal{S}} \sum_{u \in A^+} \sum_{Z \subset A^*, |Z| \geq 2} [\mathfrak{Det}(S_1 \cdot S_2, (\emptyset, Z))].$$

Let  $Z \subset A^*$  be the subset of non-empty words, with  $|Z| \geq 2$ . By Lemma 15, if  $(\emptyset, Z)$ , with  $|Z| \geq 2$ , is the label of a state of an automaton  $\mathcal{D}_{S_1 \cdot S_2}$ , then  $Z$  belongs to a set  $Q_{u, v}$ , for some  $u, v$  in  $Z$  such that  $v$  is a proper suffix of  $u$ . Therefore, we have

$$\Gamma = \sum_{(S_1, S_2) \in \mathcal{S}} \sum_{u \in A^+} \sum_{v \in \text{Suff}(u)} \sum_{Z \in Q_{u, v}} [\mathfrak{Det}(S_1 \cdot S_2, (\emptyset, Z))].$$

Changing the order of the sums, we get

$$\Gamma = \sum_{u \in A^+} \sum_{v \in \text{Suff}(u)} \sum_{Z \in Q_{u, v}} \sum_{(S_1, S_2) \in \mathcal{S}} [\mathfrak{Det}(S_1 \cdot S_2, (\emptyset, Z))].$$

<sup>a</sup>The operator  $[\ ]$  is defined by  $[P] = 1$  if the property  $P$  is true and 0 otherwise.

Partitioning the sum  $\Gamma$  into  $\Gamma_1 \cup \Gamma_2$ , depending on whether the word  $v$  is prefix of  $u$  or not, we obtain:

$$\begin{aligned}\Gamma_1 &= \sum_{u \in A^+} \sum_{v \in \text{Suff}(u) \setminus \text{Pref}(v)} \sum_{Z \in Q_{u,v}} \sum_{(S_1, S_2) \in \mathcal{S}} \llbracket \mathcal{D}\text{et}(S_1 \cdot S_2, (\emptyset, Z)) \rrbracket, \\ \Gamma_2 &= \sum_{u \in A^+} \sum_{v \in \text{Bord}(u)} \sum_{Z \in Q_{u,v}} \sum_{(S_1, S_2) \in \mathcal{S}} \llbracket \mathcal{D}\text{et}(S_1 \cdot S_2, (\emptyset, Z)) \rrbracket.\end{aligned}$$

To prove Theorem 16, we shall establish that  $\Gamma_1$  and  $\Gamma_2$  are both  $\mathcal{O}(|\mathcal{S}|)$ .

•  **$\Gamma_1$  is in  $\mathcal{O}(|\mathcal{S}|)$ :** For any  $u \in A^+$ , for any  $v \in \text{Suff}(u) \setminus \text{Pref}(u)$  and for any  $Z \in Q_{u,v}$ , the number of pairs of sequences  $(S_1, S_2) \in \mathcal{S}$  such that  $\mathcal{D}_{S_1 \cdot S_2}$  contains a state labelled by  $(\emptyset, Z)$  is at most

$$m_1 |A|^{n_1 - |u| + |v|} \binom{n_1 - |u| + |v|}{m_1 - 1} \times m_2 (m_2 - 1) |A|^{n_2 - |u| - |v|} \binom{n_2 - |u| - |v| + 1}{m_2 - 1}.$$

The left part is a consequence of Lemma 2,  $v^{-1}u$  being a suffix of an element in  $X_1$ ; the right part is a consequence of Lemma 3,  $v$  and  $u$  being prefixes of two distinct elements in  $X_2$ . Hence,  $\Gamma_1$  is bounded above by

$$\sum_{u \in A^+} \sum_{\substack{v \in \text{Suff}(u) \\ v \notin \text{Pref}(u)}} \sum_{Z \in Q_{u,v}} m_1 m_2 (m_2 - 1) |A|^{n_1 + n_2 - 2|u|} \binom{n_1 - |u| + |v|}{m_1 - 1} \binom{n_2 - |u| - |v| + 1}{m_2 - 1}.$$

Moreover if  $u$  is the longest word in  $Z$ , then by Lemma 15,  $Z$  must be a subset of  $(X_1^{-1}X_1)u$  for  $(\emptyset, Z)$  to be the label of a state. But  $|X_1^{-1}X_1| \leq m_1^2$ . Therefore, setting  $|u| = \ell$  and  $|v| = i$ , we obtain:

$$\Gamma_1 \leq \sum_{\ell=2}^{n_2 - m_2 + 1} |A|^\ell \sum_{i=1}^{\ell-1} 2^{m_1^2} m_1 m_2 (m_2 - 1) |A|^{n_1 + n_2 - 2\ell} \binom{n_1 - \ell + i}{m_1 - 1} \binom{n_2 - \ell - i + 1}{m_2 - 1}.$$

Since  $1 \leq i \leq \ell - 1$  and  $\ell \geq 2$ ,  $\binom{n_1 - \ell + i}{m_1 - 1} \leq \binom{n_1 - 1}{m_1 - 1}$  and  $\binom{n_2 - \ell - i + 1}{m_2 - 1} \leq \binom{n_2 - 2}{m_2 - 1}$ . Thus

$$\Gamma_1 \leq D_{m_1, m_2} |A|^{n_1 + n_2} \binom{n_1 - 1}{m_1 - 1} \binom{n_2 - 2}{m_2 - 1} \sum_{\ell=2}^{n_2 - m_2 + 1} |A|^{-\ell} (\ell - 1),$$

where  $D_{m_1, m_2}$  only depends on  $m_1$  and  $m_2$ . Since  $\sum_{\ell=2}^{\infty} |A|^{-\ell} (\ell - 1)$  is a convergent series, it is bounded above by a constant  $M$ . Therefore, by Equation (1) of Proposition 1,  $\Gamma_1 \leq M D_{m_1, m_2} |\mathcal{S}|$  or, in other words,  $\Gamma_1 = \mathcal{O}(|\mathcal{S}|)$ .

•  **$\Gamma_2$  is in  $\mathcal{O}(|\mathcal{S}|)$ :** For any  $u \in A^+$ , any  $v \in \text{Bord}(u)$  and any  $Z \in Q_{u,v}$ , the number of pairs of sequences  $(S_1, S_2) \in \mathcal{S}$  such that  $\mathcal{D}_{S_1 \cdot S_2}$  contains a state labelled by  $(\emptyset, Z)$  is at most

$$m_1 |A|^{n_1 - |u| + |v|} \binom{n_1 - |u| + |v|}{m_1 - 1} \times m_2 |A|^{n_2 - |u|} \binom{n_2 - |u|}{m_2 - 1}.$$

Both the left and the right parts are consequences of Corollary 2.1,  $v^{-1}u$  being a suffix of an element in  $X_1$  and  $u$  being a prefix of a word in  $X_2$ . Hence, we get

$$\Gamma_2 \leq \sum_{u \in A^+} \sum_{v \in \text{Bord}(u)} \sum_{Z \in Q_{u,v}} m_1 m_2 |A|^{n_1 + n_2 - 2|u| + |v|} \binom{n_1 - |u| + |v|}{m_1 - 1} \binom{n_2 - |u|}{m_2 - 1}.$$

As for  $\Gamma_1$ , the number of subsets  $Z$  of  $Q_{u,v}$  that can appear in a label of a state in the automaton is at most  $2^{m_1^2}$ . Therefore setting  $|u| = \ell$  and  $|v| = i$ , by Lemma 4, we obtain:

$$\Gamma_2 \leq \sum_{\ell=2}^{n_2-m_2+1} \sum_{i=1}^{\ell-1} |A|^{\ell-i} 2^{m_1^2} m_1 m_2 |A|^{n_1+n_2-2\ell+i} \binom{n_1-\ell+i}{m_1-1} \binom{n_2-\ell}{m_2-1}.$$

Hence, there exists  $E_{m_1, m_2}$  such that

$$\Gamma_2 \leq E_{m_1, m_2} \binom{n_1-1}{m_1-1} \binom{n_2-2}{m_2-1} |A|^{n_1+n_2} \sum_{\ell=2}^{n_2-m_2+1} |A|^{-\ell} (\ell-1).$$

Thus  $\Gamma_2 \leq |S| E_{m_1, m_2} \sum_{\ell=2}^{\infty} (\ell-1) |A|^{-\ell}$  or, in other words,  $\Gamma_2 = \mathcal{O}(|S|)$ . This concludes the proof since, putting all together,  $\Delta = (n_1 + n_2 + \mathcal{O}(1))|S|$ .

## 5. Average State Complexity of the Star

In the following we study the average state complexity of the star of finite languages.

**Theorem 16 (Star)** *For the uniform distribution over the sets  $X$  of  $\text{Set}_{n,m}$  the average state complexity of  $X^*$  is in  $\Theta(n)$  when  $n$  tends to infinity. Moreover, if the cardinality of the alphabet is greater than or equal to 3, this state complexity is asymptotically equivalent to  $n$ .*

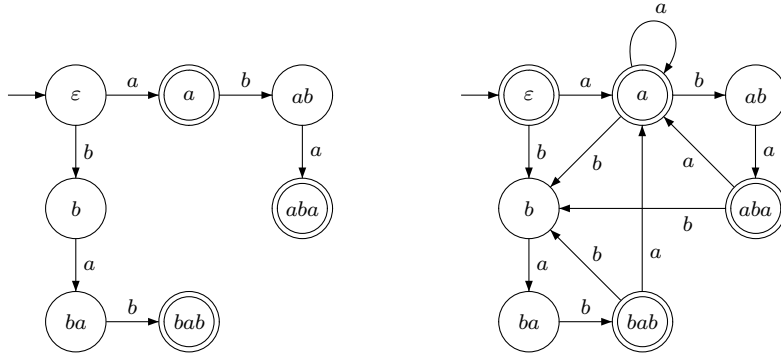
In order to prove Theorem 16 we show that the average number of states of the deterministic automaton  $\mathcal{D}_X$  (defined in the next section) recognizing  $X^*$  is linear in the length of  $X$  and that, if the alphabet is of cardinality greater than two, this complexity is smaller than or equal to  $n + \mathcal{O}(1)$ . The result holds for the average state complexity of  $X^*$  since, for each  $X$  in  $\text{Set}_{n,m}$ , the size of the minimal automaton  $\mathcal{M}_X$  of  $X^*$  is smaller than or equal to the size of  $\mathcal{D}_X$ .

### 5.1. Construction

Let  $X \subset A^*$  be a finite set of words. The automaton  $\mathcal{T}_X$  defined in Section 2.1. recognizes the set  $X$  and the automaton  $\mathcal{A}_X = (A, \text{Pr}(X), \mathcal{T}_X \cup T, \{\varepsilon\}, X \cup \{\varepsilon\})$ , where  $T = \{(u, a, a) \mid u \in X, a \in A \cap \text{Pr}(X)\}$  recognizes  $X^*$  (see Fig.1). We denote by  $\mathcal{A}_S$  the automaton defined for the set of elements of any sequence  $S$  by the above construction. In such an automaton, only the states labelled by a letter have more than one incoming transition.

For any finite set of words  $X \subset A^*$  (resp. any sequence  $S$ ), we denote by  $\mathcal{D}_X$  (resp.  $\mathcal{D}_S$ ) the accessible deterministic automaton obtained from the automaton  $\mathcal{A}_X$  (resp.  $\mathcal{A}_S$ ) making use of the subset construction and by  $\mathcal{M}_X$  the minimal automaton of  $X^*$ .

**Lemma 17.** *For any finite set of words  $X \subset A^*$ , the states of the deterministic automaton  $\mathcal{D}_X$  recognizing  $X^*$  are non-empty subsets  $\{u_1, \dots, u_l\}$  of  $\text{Pr}(X)$  such that, for all  $i, j \in \{1, \dots, l\}$ ,*


 Fig. 1. The automata  $\mathcal{T}_X$  and  $\mathcal{A}_X$ , for  $X = \{a, aba, bab\}$ 

- either  $u_i$  is a suffix of  $u_j$  or  $u_j$  is a suffix of  $u_i$ ,
- there exist  $x_0, \dots, x_{h_i}, y_0, \dots, y_{h_j} \in X$  such that  $x_0 \cdots x_{h_i} u_i = y_0 \cdots y_{h_j} u_j$ .

**Proof.** If  $\{u_1, \dots, u_l\}$  is a state of  $\mathcal{D}_X$ , then, for each  $i$ ,  $u_i$  is a prefix of a word of  $X$  by construction. Since every state in  $\mathcal{D}_X$  is accessible, then there exists a path from the initial state  $\{\varepsilon\}$  to  $\{u_1, \dots, u_l\}$  with label  $\alpha$ . By definition of subset construction, for each  $u_i$ , there exists in  $\mathcal{A}_X$  a path  $p_i$  with label  $\alpha$  from the initial state  $\varepsilon$  to the state  $u_i$ . Moreover the path  $p_i$  must have as suffix a path with label  $u_i$ , starting at a final state and ending at  $u_i$ . So, for each  $i$ , there exist  $x_0, \dots, x_{h_i} \in X$  such that  $\alpha = x_0 \cdots x_{h_i} u_i$ , concluding the proof of the second item.  $\square$

**Corollary 18.** Let  $X$  be a finite set and  $u, v \in A^*$ ,  $|u| > |v|$ . If  $\mathcal{D}_X$  has a state containing  $u$  and  $v$  then  $u$  and  $v$  are prefixes of two words in  $X$  and there exists  $w \in \text{Suff}(X)X^* \cup X^+$  such that  $u = vw$ .

## 5.2. Upper Bound

To prove the result on sets it is sufficient to prove it on sequences since

$$\frac{1}{|\text{Set}_{n,m}|} \sum_{X \in \text{Set}_{n,m}} \#\mathcal{D}_X = \frac{1}{m! |\text{Set}_{n,m}|} \sum_{S \in \mathcal{S}_{n,m}^\#} \#\mathcal{D}_S \leq \frac{1}{m! |\text{Set}_{n,m}|} \sum_{S \in \mathcal{S}_{n,m}} \#\mathcal{D}_S$$

and then Equation (2) of Proposition 1 leads to the conclusion.

Let  $Y \subset A^*$  and  $S \in \mathcal{S}_{n,m}$ . Recall that we denote by  $\mathfrak{Det}(S, Y)$  the property:  $Y$  is the label of a state of  $\mathcal{D}_S$ .

To find an upper bound for the average number of states of the deterministic automaton  $\mathcal{D}_S$  when the sequence  $S$  ranges over the set  $\mathcal{S}_{n,m}$ , we count the states of all automata according to their labels. More precisely we want to estimate the sum

$$\sum_{S \in \mathcal{S}_{n,m}} \#\mathcal{D}_S = \sum_{S \in \mathcal{S}_{n,m}} \sum_{Y \subset A^*} [\mathfrak{Det}(S, Y)].$$

Taking into account the cardinality of the labels of the states, we find

$$\sum_{S \in \mathcal{S}_{n,m}} \#\mathcal{D}_S = \sum_{S \in \mathcal{S}_{n,m}} \sum_{|Y|=1} [\mathfrak{Det}(S, Y)] + \sum_{S \in \mathcal{S}_{n,m}} \sum_{|Y| \geq 2} [\mathfrak{Det}(S, Y)].$$

The first sum deals with states labelled by a single word. Since, for each  $S \in \mathcal{S}_{n,m}$ , the words that appear in the labels of states of  $\mathcal{D}_S$  are prefixes of words of  $S$ , we have

$$\sum_{S \in \mathcal{S}_{n,m}} \sum_{|Y|=1} [\mathfrak{Det}(S, Y)] = \sum_{S \in \mathcal{S}_{n,m}} \sum_{\substack{u \text{ prefix of} \\ \text{a word of } S}} [\mathfrak{Det}(S, \{u\})] \leq (n+1)|\mathcal{S}_{n,m}|.$$

It remains to study the sum

$$\Delta = \sum_{S \in \mathcal{S}_{n,m}} \sum_{|Y| \geq 2} [\mathfrak{Det}(S, Y)].$$

Let  $Y \subset A^*$  be a non-empty set which is not a singleton. By Lemma 17, if  $Y$  is the label of a state of an automaton  $\mathcal{D}_S$ , then  $Y$  belongs to a set  $Q_{u,v}$ , for some non-empty word  $u$  and some proper suffix  $v$  of  $u$ . Therefore, we have

$$\Delta = \sum_{S \in \mathcal{S}_{n,m}} \sum_{u \in A^+} \sum_{v \in \text{Suff}(u)} \sum_{Y \in Q_{u,v}} [\mathfrak{Det}(S, Y)].$$

Changing the order of the sums, we obtain

$$\Delta = \sum_{u \in A^+} \sum_{v \in \text{Suff}(u)} \sum_{Y \in Q_{u,v}} \sum_{S \in \mathcal{S}_{n,m}} [\mathfrak{Det}(S, Y)].$$

We then partition the sum  $\Delta$  into  $\Delta_1 + \Delta_2$  depending on whether the word  $v$  is prefix of  $u$  or not:

$$\Delta_1 = \sum_{u \in A^+} \sum_{v \in \text{Bord}(u)} \sum_{Y \in Q_{u,v}} \sum_{S \in \mathcal{S}_{n,m}} [\mathfrak{Det}(S, Y)], \quad (5)$$

$$\Delta_2 = \sum_{u \in A^+} \sum_{v \in \text{Suff}(u) \setminus \text{Pref}(u)} \sum_{Y \in Q_{u,v}} \sum_{S \in \mathcal{S}_{n,m}} [\mathfrak{Det}(S, Y)]. \quad (6)$$

To prove Theorem 16, we study the asymptotic behavior of  $\Delta_1$  and  $\Delta_2$ .

### 5.3. The Case of Alphabets with at Least Three Letters

The following lemmas will be used to prove the second part of Theorem 16. They rely on the condition  $w \in \text{Suff}(X)X^* \cup X^+$  of Corollary 18.

**Lemma 19.** *Let  $u, v$  be two words in  $A^+$  such that  $v$  is a suffix of  $u$ , but not a prefix of  $u$  and let  $w$  be the word such that  $u = wv$ . Setting  $|u| = \ell$  and  $|v| = i$ , then there are at most*

$$C_m |A|^{n-2\ell} \binom{n-2\ell+1}{m-1} + C_m A^{n-\ell-i} \binom{n-\ell-i}{m-2}$$

*sequences  $S$  in  $\mathcal{S}_{n,m}$  such that  $u$  et  $v$  are prefixes of two words in  $S$  and such that  $w \in \text{Suff}(S)S^* \cup S^+$ .*

**Proof.** Let  $S$  be a sequence that satisfies the conditions of the lemma, We consider two cases, depending on whether  $w \in \text{Suff}(S)$  or not.

If  $w \in \text{Suff}(S)$ , then there exist three words  $x_u, x_v$  and  $x_w$  in  $S$  such that  $u$  is a prefix of  $x_u$ ,  $v$  is a prefix of  $x_v$  and  $w$  is a suffix of  $x_w$ . Since  $v$  is not a prefix of  $u$ , then  $x_u \neq x_v$ . We distinguish three cases:

- If  $x_u = x_w$  and  $|u| + |w| \geq |x_u|$ , setting with  $j = |x_u|$ , the sequence  $S$  can be built, from a sequence in  $\mathcal{S}_{n-j, m-1}$  having  $v$  as prefix of one of its words, by adding  $x_u$  at some position. Hence, using Lemma 2, there are at most  $m(m-1) \binom{n-i-j}{m-2} |A|^{n-j-i}$  such sequences. Since  $j$  ranges from  $\ell + 1$  to  $2\ell - 1$ , the total number of such sequences is bounded above by  $K_m \binom{n-i-\ell}{m-2} |A|^{n-\ell-i}$  for some  $K_m$  that only depends on  $m$ .
- If  $x_v = x_w$  and  $|v| + |w| \geq |x_v|$ , then using similar arguments as the previous ones, we can prove that there are at most  $K'_m \binom{n-i-\ell}{m-2} |A|^{n-\ell-i}$  such sequences.
- In all other cases, the sequence can be built from a sequence of  $\mathcal{S}_{n-(\ell-i), m}$  having  $u$  and  $v$  as prefixes of two of its words, by adding  $w$  at the end of some element. Hence, by Lemma 3, there are at most  $K''_m \binom{n-2\ell+1}{m-1} |A|^{n-2\ell}$  such sequences.

If  $w \notin \text{Suff}(S)$ , then  $w \in (\text{Suff}(S) \cup \{\varepsilon\})X^+$ . Therefore there exist a word  $x_w \in S$  that is a suffix of  $w$  and two words  $x_u$  and  $x_v$  having respectively  $u$  and  $v$  as prefixes. Since  $|w| < |u|$ , then  $|x_w| < |x_u|$  and the words  $x_u$  and  $x_w$  are distinct. Since  $v$  is not a prefix of  $u$ , then  $x_u$  and  $x_v$  are distinct too. Denote by  $j$  the length of  $x_w$ . We now distinguish whether  $x_v = x_w$  or not.

If  $x_v \neq x_w$ , since  $x_w$  is a word belonging to the sequence, by Lemma 3, the number of sequences that satisfies the properties is at most  $m(m-1) \binom{n-i-\ell-j+1}{m-2} |A|^{n-\ell-i-j}$ . Summing for  $j$  from 1 to  $\ell - i$ , we find that there are at most  $L_m |A|^{n-i-\ell} \binom{n-i-\ell}{m-2}$  such sequences, where  $L_m$  only depends on  $m$ .

If  $x_v = x_w$ , then, by Lemma 2, there are at most  $B'_m \binom{n-j-\ell}{m-2} |A|^{n-j-\ell}$  such sequences. Summing for  $j$  from  $i$  to  $\ell - i - 1$ , we find that there are at most  $L'_m |A|^{n-i-\ell} \binom{n-i-\ell}{m-2}$  such sequences, where  $L'_m$  only depends on  $m$ .

Adding all the contributions, we get the announced upper bound.  $\square$

**Lemma 20.** *Let  $u, v$  be two words in  $A^+$  such that  $v$  is a proper border of  $u$  and let  $w$  be the word such that  $u = vw$ . Setting  $|u| = \ell$  and  $|v| = i$ , then there are at most*

$$D_m |A|^{n-2\ell+i} \binom{n-2\ell+i}{m-1} + D_m |A|^{n-\ell} \binom{n-\ell-1}{m-2}$$

*sequences  $S$  in  $\mathcal{S}_{n, m}$  such that  $u$  et  $v$  are prefixes of two words in  $S$  and such that  $w \in \text{Suff}(S)S^* \cup S^+$ .*

**Proof.** We consider two cases depending on whether  $w \in \text{Suff}(S)$  or not.



If  $w \in \text{Suff}(S)$ , then there exist  $x_u$  and  $x_w$  in  $S$  such that  $u$  is a prefix of  $x_u$  and  $w$  is a proper suffix of  $x_w$ . The number of such sequences with  $x_u = x_w$  and  $|u| + |w| \leq |x_u|$  is smaller than or equal to  $m \binom{n-j-1}{m-2} |A|^{n-j}$ . Summing for  $j$  from  $\ell + 1$  to  $2\ell - 1$ , we find that there are at most  $E_m \binom{n-\ell-1}{m-2} |A|^{n-\ell}$  such sequences, for some  $E_m$  depending only on  $m$ . On the other hand, if  $x_u \neq x_w$  or  $|u| + |w| > |x_u|$ , the number of sequences is smaller than or equal to  $E'_m \binom{n-\ell-(\ell-i)+1}{m-1} |A|^{n-\ell-(\ell-i)}$ .

If  $w \notin \text{Suff}(S)$ , then  $w \in (\text{Suff}(S) \cup \{\varepsilon\})X^+$  and there exists a word  $x_w$  in  $S$  that is a suffix of  $w$ . Setting  $|x_w| = j$ , by Lemma 2, there are at most  $F_m \binom{n-\ell-j}{m-2} |A|^{n-\ell-j}$  such sequences. Summing for  $j$  from 1 to  $\ell - i$ , we find that there are at most  $F'_m \binom{n-\ell-1}{m-2} |A|^{n-\ell}$  such sequences.

Adding all the contributions, we get the announced upper bound.  $\square$

In the sequel, we prove that  $\Delta_1$  and  $\Delta_2$  defined by Equations (5) and (6) (p.15) are both in  $\mathcal{O}(|\mathcal{S}_{n,m}|)$ .

By Corollary 18 and Lemma 20, one has  $\Delta_1 \leq \Delta_{1,1} + \Delta_{1,2}$  with

$$\begin{aligned} \Delta_{1,1} &= \sum_{u \in A^+} \sum_{v \in \text{Bord}(u)} \sum_{Y \in Q_{u,v}} D_m |A|^{n-2|u|+|v|} \binom{n-2|u|+|v|}{m-1}, \\ \Delta_{1,2} &= \sum_{u \in A^+} \sum_{v \in \text{Bord}(u)} \sum_{Y \in Q_{u,v}} D_m |A|^{n-|u|} \binom{n-|u|-1}{m-2}. \end{aligned}$$

Setting  $|u| = \ell$  and  $|v| = i$  and using Lemma 4, we get

$$\Delta_{1,1} \leq \sum_{\ell=2}^{n-m+1} \sum_{i=1}^{\ell-1} |A|^{\ell-i} 2^{i-1} D_m |A|^{n-2\ell+i} \binom{n-2\ell+i}{m-1}.$$

Since, for  $2 \leq \ell \leq n - m + 1$  and  $1 \leq i \leq \ell - 1$ ,  $\binom{n-2\ell+i}{m-1} \leq \binom{n-3}{m-1}$ , we obtain

$$\Delta_{1,1} \leq \frac{1}{2} D_m |A|^n \binom{n-3}{m-1} \left( \sum_{\ell=2}^{\infty} |A|^{-\ell} \right) \left( \sum_{i=1}^{\infty} |A|^{-i} 2^i \right).$$

Finally, since  $|A| \geq 3$ , then  $\Delta_{1,1}$  is in  $\mathcal{O}(|\mathcal{S}_{n,m}|)$ .

The same arguments lead to

$$\Delta_{1,2} \leq \sum_{\ell=2}^{n-m+1} \sum_{i=1}^{\ell-1} |A|^{\ell-i} 2^{i-1} D_m |A|^{n-\ell} \binom{n-\ell-1}{m-2}.$$

Moreover, since  $\sum_{j=m-2}^N \binom{j}{m-2} = \binom{N+1}{m-1}$ , we obtain

$$\Delta_{1,2} \leq \frac{1}{2} D_m |A|^n \binom{n-2}{m-1} \left( \sum_{i=1}^{\infty} |A|^{-i} 2^i \right)$$

or, in other words, since  $|A| \geq 3$ ,  $\Delta_{1,2} = \mathcal{O}(|\mathcal{S}_{n,m}|)$ .

Thus using exactly the same kind of computations, one can prove by Lemma 19 that  $\Delta_2$  is in  $\mathcal{O}(|\mathcal{S}_{n,m}|)$ .

#### 5.4. For Binary Alphabets

We now prove that the average state complexity of the star of a finite language on a binary alphabet is linear. More precisely we show that  $\Delta_1$  and  $\Delta_2$  defined by Equations (5) and (6) (p.15) are both in  $\mathcal{O}(n|\mathcal{S}_{n,m}|)$ .

By Lemma 3, we have

$$\Delta_2 \leq \sum_{u \in A^+} \sum_{v \in \text{Suff}(u) \setminus \text{Pref}(u)} \sum_{Y \in Q_{u,v}} m(m-1)2^{n-|u|-|v|} \binom{n-|u|-|v|+1}{m-1}.$$

Since  $|Q_{u,v}| = 2^{|v|-1}$ , with  $\ell = |u|$  and  $i = |v|$ , we get

$$\Delta_2 \leq m(m-1) \sum_{\ell=2}^{n-m+1} 2^\ell \sum_{i=1}^{\ell-1} 2^{i-1} 2^{n-\ell-i} \binom{n-\ell-i+1}{m-1}.$$

Moreover, since  $\sum_{\ell=2}^{n-m+1} \sum_{i=1}^{\ell-1} \binom{n-\ell-i+1}{m-1} = \binom{n-1}{m-1}$ , then  $\Delta_2 \leq \frac{m(m-1)}{2} 2^n \binom{n-1}{m-1}$  and, by Equation (1) of Proposition 1,  $\Delta_2$  is in  $\mathcal{O}(n|\mathcal{S}_{n,m}|)$ .

We partition now the sum  $\Delta_1$  into two sums  $\Delta_{1,1}$  and  $\Delta_{1,2}$  depending on whether the set  $Y$  contains exactly two elements or not (and therefore belongs to some set  $Q_{u,v,w}$ ). More precisely, we have

$$\Delta_{1,1} = \sum_{u \in A^+} \sum_{v \in \text{Bord}(u)} \sum_{S \in \mathcal{S}_{n,m}} [\mathfrak{Det}(S, \{u, v\})]$$

and

$$\Delta_{1,2} = \sum_{u \in A^+} \sum_{v \in \text{Bord}(u)} \sum_{w \in \text{Suff}(v)} \sum_{Y \in Q_{u,v,w}} \sum_{S \in \mathcal{S}_{n,m}} [\mathfrak{Det}(S, Y)].$$

Using Lemma 2 and Lemma 4, and since  $\sum_{\ell=2}^{n-m+1} \binom{n-\ell}{m-1} = \binom{n-1}{m-1}$ , we obtain

$$\Delta_{1,1} \leq \sum_{\ell=2}^{n-m+1} \sum_{i=1}^{\ell-1} m \binom{n-\ell}{m-1} 2^{n-\ell} 2^{\ell-i} \leq m 2^n \binom{n-1}{m-1}.$$

Consequently, by Equation (1) of Proposition 1,  $\Delta_{1,1}$  is in  $\mathcal{O}(n|\mathcal{S}_{n,m}|)$ .

We decompose next the sum  $\Delta_{1,2}$  into the sum  $B_{1,2} + N_{1,2}$  depending on whether  $w$  is a prefix (and therefore a border) of  $v$  or not.

When  $w$  is not a prefix of  $v$ , by Lemma 3, the number of sequences  $S \in \mathcal{S}_{n,m}$  such that  $u$  and  $w$  are prefixes of two distinct words of  $S$  is smaller than or equal to  $m(m-1)2^{n-\ell-j} \binom{n-\ell-j+1}{m-1}$ .

Since, by Lemma 4, there are less than  $2^{\ell-i}$  pairs  $(u, v)$  such that  $v$  is a border

of  $u$  and since  $|Q_{u,v,w}| = 2^{|w|-1}$ , we get:

$$\begin{aligned} N_{1,2} &= \sum_{u \in A^+} \sum_{v \in \text{Bord}(u)} \sum_{w \in \text{Suff}(v) \setminus \text{Pref}(v)} \sum_{Y \in Q_{u,v,w}} \sum_{S \in \mathcal{S}_{n,m}} [\mathfrak{Det}(S, Y)] \\ &\leq m(m-1) \sum_{\ell=3}^{n-m+1} \sum_{i=2}^{\ell-1} \sum_{j=1}^{i-1} 2^{\ell-i} 2^{j-1} 2^{n-\ell-j} \binom{n-\ell-j+1}{m-1} \\ &\leq \frac{m(m-1)}{2} 2^n \sum_{\ell=3}^{n-m+1} \sum_{i=2}^{\ell-1} 2^{-i} \sum_{j=1}^{i-1} \binom{n-\ell-j+1}{m-1}. \end{aligned}$$

Since  $\binom{n-\ell-j+1}{m-1} \leq \binom{n-\ell}{m-1}$ , we obtain

$$N_{1,2} \leq \frac{m(m-1)}{2} 2^n \sum_{\ell=3}^{n-m+1} \binom{n-\ell}{m-1} \sum_{i=2}^{\ell-1} (i-1) 2^{-i}.$$

Because of the convergence of the series,  $\sum_{i=2}^{\ell-1} (i-1) 2^{-i}$  is bounded. Therefore, since  $\sum_{\ell=3}^{n-m+1} \binom{n-\ell}{m-1} = \binom{n-2}{m-2}$  and  $|\mathcal{S}_{n,m}| = \binom{n-1}{m-1} 2^n$ , we have  $N_{1,2} = \mathcal{O}(n |\mathcal{S}_{n,m}|)$ .

When  $w$  is prefix of  $v$ , the associated sum  $B_{1,2}$  is partitioned into the following sums:

$$B_{1,2} = \sum_{u \in A^+} \sum_{v \in \text{Bord}(u)} \sum_{w \in \text{Bord}(v)} \sum_{Y \in Q_{u,v,w}} \sum_{S \in \mathcal{S}_{n,m}} [\mathfrak{Det}(S, Y)] = B'_{1,2} + B''_{1,2}$$

with

$$B'_{1,2} = \sum_{u \in A^+} \sum_{\substack{v \in \text{Bord}(u) \\ |v| > \frac{2}{3}|u|}} \sum_{\substack{w \in \text{Bord}(v) \\ |w| > \frac{|v|}{2}}} \sum_{Y \in Q_{u,v,w}} \sum_{S \in \mathcal{S}_{n,m}} [\mathfrak{Det}(S, Y)]$$

and  $B''_{1,2} = B_{1,2} \setminus B'_{1,2}$ . Using Lemma 4, the fact that  $|Q_{u,v,w}| = 2^{|w|-1}$  and relaxing the constraints on the lengths of the words  $v$  and  $w$ , we get

$$B''_{1,2} \leq \sum_{\ell=3}^{n-m+1} \sum_{i=2}^{\ell-1} \sum_{j=1}^{i-1} m \binom{n-\ell}{m-1} 2^{n-\ell} 2^{\ell-\frac{i}{2}-j} 2^{j-1}.$$

Since  $\sum_{i=2}^{\ell-1} (i-1) 2^{-\frac{i}{2}}$  is bounded by a constant  $M$ ,

$$B''_{1,2} \leq mM 2^{n-1} \sum_{\ell=3}^{n-m+1} \binom{n-\ell}{m-1}.$$

Finally since  $\sum_{\ell=3}^{n-m+1} \binom{n-\ell}{m-1} = \binom{n-2}{m-2}$  and  $|\mathcal{S}_{n,m}| = \binom{n-1}{m-1} 2^n$ , then  $B''_{1,2}$  is in  $\mathcal{O}(n |\mathcal{S}_{n,m}|)$ .

Now by Lemma 2 and since  $|Q_{u,v,w}| = 2^{|w|-1}$ , we get:

$$B'_{1,2} \leq \sum_{u \in A^+} \sum_{\substack{v \in \text{Bord}(u) \\ |v| > \frac{2}{3}|u|}} \sum_{\substack{w \in \text{Bord}(v) \\ |w| > \frac{|v|}{2}}} 2^{|w|-1} m \binom{n-|u|}{m-1} 2^{n-|u|}.$$

Moreover, by Proposition 6, the words  $u, v$  and  $w$  of respective lengths  $\ell, i$  and  $j$  are powers of a same primitive word  $x$ . We set  $u = x^p x_0$ ,  $v = x^q x_0$  and  $w = x^s x_0$ , with  $p > q > s > 0$  and  $x_0 \in \text{Pr}(x)$ . Let  $r$  be the length of  $x$ , then there are less than  $2^r$  such words  $x$ . And since  $1 \leq r \leq \ell - i$  and  $i > \frac{2}{3}\ell$ , then  $r < \frac{\ell}{3}$ . As a consequence, the lengths of  $v$  and  $w$  can be written  $i = \ell - hr$  where  $1 \leq h < \ell/3r$  and  $j = \ell - h'r$  where  $h < h' < \frac{1}{2}(\frac{\ell}{r} + h)$ . Therefore, we get

$$\begin{aligned} B'_{1,2} &\leq \sum_{\ell=3}^{n-m+1} \sum_{r=1}^{\frac{\ell}{3}-1} \sum_{h=1}^{\frac{\ell}{3r}} \sum_{h'=h+1}^{\frac{1}{2}(\frac{\ell}{r}+h)} m \binom{n-\ell}{m-1} 2^{n-\ell} 2^r 2^{\ell-h'r-1} \\ &\leq m 2^{n-1} \sum_{\ell=3}^{n-m+1} \binom{n-\ell}{m-1} \sum_{r=1}^{\frac{\ell}{3}-1} 2^r \sum_{h=1}^{\frac{\ell}{3r}} \sum_{h'=h+1}^{\frac{1}{2}(\frac{\ell}{r}+h)} (2^{-r})^{h'}. \end{aligned}$$

Since  $\sum_{h=1}^{\frac{\ell}{3r}} \sum_{h'=h+1}^{\frac{1}{2}(\frac{\ell}{r}+h)} (2^{-r})^{h'} \leq 4/2^{2r}$  when  $r \geq 1$ , we find

$$B'_{1,2} \leq m 2^{n+1} \sum_{\ell=3}^{n-m+1} \binom{n-\ell}{m-1} \sum_{r=1}^{\frac{\ell}{3}-1} 2^{-r} \leq m 2^{n+1} \sum_{\ell=3}^{n-m+1} \binom{n-\ell}{m-1}.$$

Finally, since  $\sum_{\ell=3}^{n-m+1} \binom{n-\ell}{m-1} = \binom{n-2}{m}$  and  $|\mathcal{S}_{n,m}| = \binom{n-1}{m-1} 2^n$ , we obtain that  $B'_{1,2}$  is in  $\mathcal{O}(n |\mathcal{S}_{n,m}|)$ , concluding the proof.

## 6. Remarks on the Average Time Complexity

The state complexity of a language recognized by a nondeterministic automaton with  $n$  states is, in the worst case, equal to  $2^n$ . Therefore, the worst-case time complexity of the determinization is in  $\Omega(2^n)$ . It is thus natural to measure the time complexity according to the size of the output of the algorithm and strive to design algorithms whose efficiency is a function of the size of the output, rather than the input.

The constructions we proposed—to build deterministic automata recognizing the star of a finite language or the concatenation of two finite languages—mainly rely on a classical determinization of some specific nondeterministic automata. The union operation is easy to perform efficiently by considering the union of  $\{u_1, \dots, u_{m_1}\}$  and  $\{v_1, \dots, v_{m_2}\}$  as an element of the set  $\text{Set}_{n_1+n_2, m_1+m_2}$ , and constructing the tree associated to  $\{u_1, \dots, u_{m_1}, v_1, \dots, v_{m_2}\}$ .

The complexity of the subset construction basically depends upon the encoding and the storage of the set of states. At each step, for a given set of states  $P$  and a letter  $a \in A$ , the algorithm computes the set  $P \cdot a$  of states of the initial automaton that can be reached from a state of  $P$  by a transition labelled  $a$ . Then it tests whether this set has already been computed.

Here the automata to be determinized are specific. In both the star and the concatenation constructions, they have the following property: for any accessible set of states  $X$  and every letter  $a$ , the size of  $X \cdot a$  is at most twice the size of  $X$ . More precisely, the automata satisfy the following:

- for the star, the image of a state  $u$  by a letter  $a$  in the nondeterministic automaton is either  $\emptyset$ ,  $a$ ,  $ua$  or  $\{a, ua\}$ ;
- for the concatenation, the image by a letter  $a$  of a state of the form  $(\emptyset, X)$  is  $(\emptyset, X \cdot a)$  and  $X \cdot a$  is of size at most  $|X|$  since the second automaton is deterministic; the image of  $(u, X)$  by a letter  $a$  is  $(z, X')$ , where  $X'$  is either  $X \cdot a$  or  $X \cdot a \cup \{a\}$ .

Hence computing the image of a set of states  $X$  by a letter  $a$  can be performed in time  $\mathcal{O}(P(|X|))$ , where  $P$  is some polynomial.

In order to store the sets of states,  $N + 1$  balanced trees  $\mathcal{T}_0, \dots, \mathcal{T}_N$  are used, where each tree  $\mathcal{T}_i$  contains only subsets of size  $i$ . When a new set of states  $X$  is computed, it is inserted in the tree  $\mathcal{T}_{|X|}$ —in the concatenation case, the size of a state  $(z, X)$  is defined as the size of  $X$ . In order to cover all the possible sizes, it is enough to set  $N = n + 1$  in the star case and  $N = n_2 + 1$  in the concatenation case. Each balanced tree  $T \in \mathcal{T}_i$  contains at most  $\binom{N}{i} \leq N^i$  elements in the star case, and at most  $2\binom{N}{i} \leq 2N^i$  in the concatenation case, since the first coordinate can be either a word or  $\emptyset$ , the word being unique for a given second coordinate. Thus the insertion and search in  $T$  can be performed in  $\mathcal{O}(i \log N)$  comparisons. Since the comparisons can be performed in polynomial time in  $i$ , the overall complexity of the computation of the image of  $X$  by a letter  $a$  (namely, checking if  $X \cdot a$  is in  $\mathcal{T}_{|X \cdot a|}$  and insert it if it is not) is in time  $\mathcal{O}(Q(i) \log N)$ , for some polynomial  $Q$ .

As a consequence, one can show the following results:

- for  $|A| \geq 3$ , the average time complexity of the construction of  $\mathcal{D}_X$  recognizing the star of a finite language  $X$  in  $\text{Set}_{n,m}$  is in  $\mathcal{O}(n \log n)$ ,
- for  $|A| \geq 2$ , the average time complexity of the construction of  $\mathcal{D}_{X_1 X_2}$  recognizing the concatenation of two finite languages  $X_1 \in \text{Set}_{n_1, m_1}$  and  $X_2 \in \text{Set}_{n_2, m_2}$  is in  $\mathcal{O}((n_1 + n_2) \log n_2)$ .

The proof consists in reproducing the proofs of Theorem 16 and Theorem 14, adding a multiplicative factor  $Q(i + 1) \log N$ .

**Conclusion** To manipulate finite languages given by lists of words, deterministic automata are very efficient when the probability distribution we considered here models the input data correctly: the theoretical possible exponential growth in space almost never appears, and deterministic automata can be quickly computed using standard constructions.

## References

- [1] F. Bassino, L. Giambruno, C. Nicaud, The average state complexity of the star of a finite set of words is linear, *International Conference on Developments in Language Theory 2008 (DLT'08)* volume 5257 in Lect. Notes Comput. Sci., 134–145. Springer, 2008.
- [2] J. Berstel, D. Perrin. *Theory of Codes*. Academic Press, 1985.

22 *Bibliography*

- [3] C. Campeanu, K. Culik, K. Salomaa, S. Yu. State complexity of basic operations on finite languages. In *Automata Implementation: 4th International Workshop on Implementing automata (WIA'99)*, Vol. 2214 in Lect. Notes Comput. Sci., 60–70, 2001.
- [4] C. Campeanu, K. Salomaa, S. Yu. State complexity of regular languages: finite versus infinite. In C. S. Calude and G. Paun, eds., *Finite Versus Infinite: Contributions to an Eternal Dilemma*, 53–73, Springer, 2000.
- [5] J. Clément, J.-P. Duval, G. Guaiana, D. Perrin, G. Rindone. Parsing with a finite dictionary. *Theoret. Comput. Sci.*, 340:432–442, 2005.
- [6] K. Ellul, B. Krawetz, J. Shallit, M.-W. Wang. Regular expressions: new results and open problems. *J. Autom. Lang. Combin.*, 10:407–437, 2005.
- [7] P. Flajolet, R. Sedgewick. *Analytic Combinatorics*, Cambridge University Press, 2009.
- [8] H. Gruber, M. Holzer. On the average state and transition complexity of finite languages. *Theoret. Comput. Sci.*, 387:155–166, 2007.
- [9] J. E. Hopcroft, J. D. Ullman *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley Publishing Company, 1979.
- [10] J. L. Ramírez-Alfonsín. Complexity of the Frobenius problem. *Combinatorica*, 16:143–147, 1996.
- [11] J. L. Ramírez-Alfonsín. *The Diophantine Frobenius Problem*. Oxford University Press, 2005.
- [12] J.-Y. Kao, J. Shallit, Z. Xu. The Frobenius problem in a free monoid. In S. Albers and P. Weil, eds., *Symposium on Theoretical Aspects of Computer Science 2008* (Bordeaux), 421–432, [www.stacs-cong.org](http://www.stacs-cong.org).
- [13] M. Lothaire. *Combinatorics on Words*, Vol 17 of Encyclopedia of mathematics and its applications. Addison-Wesley, 1983.
- [14] M. Lothaire. *Algebraic Combinatorics on Words*, Vol 90 of Encyclopedia of mathematics and its applications. Cambridge University Press, 2002.
- [15] M. Lothaire. *Applied Combinatorics on Words*, Vol 104 of Encyclopedia of mathematics and its applications. Cambridge University Press, 2005.
- [16] A. N. Maslov. Estimates of the number of states of finite automata. *Dokl. Akad. Nauk. SSSR*, 194:1266–1268, 1970. (in Russian). English translation in. *Soviet. Math. Dokl.*, 11:1373–1375, 1970.
- [17] S. Yu, Q. Zhuang, K. Salomaa. The state complexities of some basic operations on regular languages. *Theoret. Comput. Sci.*, 125:315–328, 1994.