Brzozowski Algorithm Is Generically Super-Polynomial Deterministic Automata

Sven de Felice, Cyril Nicaud

Brzozowski Algorithm is Generically Super-Polynomial for Deterministic Automata*

Sven De Felice[1] and Cyril Nicaud[1]

LIGM, Université Paris-Est & CNRS, 77454 Marne-la-Vallée Cedex 2, France
sdefelic@univ-mlv.fr, nicaud@univ-mlv.fr

Abstract. We study the number of states of the minimal automaton of the mirror of a rational language recognized by a random deterministic automaton with n states. We prove that, for any $d > 0$, the probability that this number of states is greater than n^d tends to 1 as n tends to infinity. As a consequence, the generic and average complexities of Brzozowski minimization algorithm are super-polynomial for the uniform distribution on deterministic automata.

1 Introduction

Brzozowski proved [5] that determinizing a trim co-deterministic automaton which recognizes a language $\mathcal{L}$ yields the minimal automaton of $\mathcal{L}$. This can be turned into a simple minimization algorithm: start with an automaton, compute its reversal, determinize it and reverse the result in order to obtain a co-deterministic automaton recognizing the same language. A last determinization gives the minimal automaton, by Brzozowski's property.

The determinization steps use the classical subset construction, which is well-known to be of exponential complexity in the worst-case. The co-deterministic automaton $\mathcal{A}_n$ of Fig. 1 is a classical example of such a combinatorial explosion: it has n states and its minimal automaton has 2^{n-1} states.

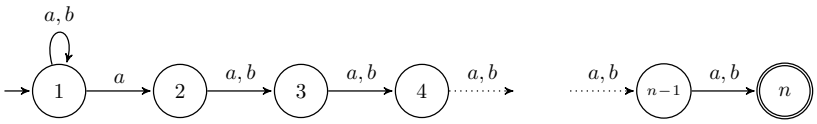

Fig. 1. Determinizing this co-deterministic automaton $\mathcal{A}_n$ with n states, which recognizes A^*aA^{n-2}, yields a minimal automaton with 2^{n-1} states.

How good is Brzozowski minimization algorithm? If the input is a non-deterministic automaton, the combinatorial explosion can be unavoidable, as

* This work is supported by the French National Agency (ANR) through ANR-10-LABX-58 and through ANR-2010-BLAN-0204.

for $\mathcal{A}_n$, and this algorithm can be a good solution (see [17] for an experimental comparison of Brzozowski algorithm versus determinization combined with Hopcroft algorithm). However, if the input is a deterministic automaton, Brzozowski algorithm still has exponential worst-case complexity, which is easily seen by taking the reverse of $\mathcal{A}_n$ as input. Since there exists polynomial solutions to minimize deterministic automata, such as Hopcroft algorithm [13] which runs in time $\mathcal{O}(n \log n)$, there is no use for Brzozowski algorithm in the deterministic case, unless the combinatorial explosion happens very rarely.

Let $\mathcal{L}$ be the language recognized by a n-state deterministic automaton taken uniformly at random. In this article we estimate the typical number of states of the minimal automaton of the mirror $\tilde{\mathcal{L}}$ of $\mathcal{L}$. More precisely, we prove that this quantity is generically super-polynomial, that is, for any $d > 0$, the probability that there are more than n^d states in the minimal automaton of $\tilde{\mathcal{L}}$ tends to 1 as n tends to infinity.

As a consequence, Brzozowski algorithm has super-polynomial generic and average complexity when used on deterministic automata, for the uniform distribution: the combinatorial explosion is almost always met during the process.

Some related works. The interest in statistical properties of random deterministic automata started with the work of Korshunov [14], who studied their combinatorics and exhibited some of their typical behavior. In recent years, an increased activity on the topic aimed at giving mathematical proofs for phenomena observed experimentally. For instance, it was proved in [1, 8] that the average complexity of Moore algorithm, another minimization algorithm, is significantly better than its worst-case complexity, making this algorithm a reasonable solution in practice. The reader can find some results on the average state complexity of operations under different settings in [16, 3]. Let us also mention the recent article [2], in the same area, which focus on quantifying the probability that a random deterministic automaton is minimal.

2 Preliminaries

For any $n \geq 1$, let $[n]$ denote the set $\{1, \ldots, n\}$. If E is a finite set, we denote its cardinality by $|E|$ and its power set by 2^E. A sequence of non-negative real numbers $(x_n)_{n\geq 1}$ *grows super-polynomially* (or is *super-polynomial*) when, for every $d > 0$, there exists some n_d such that for every $n \geq n_d$, $x_n \geq n^d$.

2.1 Automata

Basic definitions. Let A be a finite alphabet, an automaton $\mathcal{A}$ is a tuple (Q, δ, I, F), where Q is its finite set of states, $I \subseteq Q$ is its set of initial states and $F \subseteq Q$ is its set of final states. Its transition function is a (partial) map from $Q \times A$ to 2^Q. A transition of $\mathcal{A}$ is a tuple $(p, a, q) \in Q \times A \times Q$, which we write $p \xrightarrow{a} q$, such that $q \in \delta(p, a)$. The map δ is classically extended by morphism to $Q \times A^*$. We denote by $\mathcal{L}(\mathcal{A})$ the set of words recognized by $\mathcal{A}$.

A deterministic and complete automaton is an automaton such that $|I| = 1$ and for every $p \in Q$ and $a \in A$, $|\delta(p, a)| = 1$; for such an automaton we consider that δ is a (total) map from $Q \times A^*$ to Q to simplify the notations.

A state p in an automaton is *accessible* (resp. *co-accessible*) when there is a path from an initial state to p (resp. from p to a final state). The *accessible part* (resp. *co-accessible* part) of an automaton is the set of its accessible states (resp. co-accessible states). A *trim* automaton is an automaton whose states are all accessible and co-accessible. If $\mathcal{A}$ is an automaton, we denote by $\mathrm{Trim}(\mathcal{A})$ the automaton obtained after removing states that are not accessible or not co-accessible.

For any automaton $\mathcal{A} = (Q, \delta, I, F)$, we denote by $\tilde{\mathcal{A}}$ the *reverse* of $\mathcal{A}$, which is the automaton $\tilde{\mathcal{A}} = (Q, \tilde{\delta}, F, I)$, where $p \xrightarrow{a} q$ is a transition of $\tilde{\mathcal{A}}$ if and only if $q \xrightarrow{a} p$ is a transition of $\mathcal{A}$. The automaton $\tilde{\mathcal{A}}$ recognizes the mirror[1] of $\mathcal{L}(\mathcal{A})$. An automaton is *co-deterministic* when its reverse is deterministic.

Recall that the *minimal automaton* of a rational language $\mathcal{L}$ is the smallest deterministic and complete automaton[2] that recognizes $\mathcal{L}$. To each rational language $\mathcal{L}$ corresponds a minimal automaton, which is unique up to isomorphism.

Subset construction and Brzozowski algorithm. If $\mathcal{A} = (Q, \delta, I, F)$ is a non-deterministic automaton, it is classical that the subset automaton of $\mathcal{A}$ defined by

$$\mathcal{B} = \big(2^Q, \gamma, \{I\}, \{X \in 2^Q \mid F \cap X \neq \emptyset\}\big)$$

is a deterministic automaton that recognizes the same language, where for every $X \in 2^Q$ and every $a \in A$, $\gamma(X, a) = \cup_{p \in X} \delta(p, a)$. This is of course still true if we only take the accessible part of $\mathcal{B}$, and this is not a difficulty when implementing it, since the accessible part of $\mathcal{B}$ can be built on the fly, using the rule for γ in a depth-first traversal of $\mathcal{B}$ starting from I. We denote by $\mathrm{Subset}(\mathcal{A})$ the accessible part of the subset automaton of $\mathcal{A}$.

In [5], Brzozowski established the following result:

Theorem 1 (Brzozowski). *If $\mathcal{A}$ is a trim co-deterministic automaton then $\mathit{Subset}(\mathcal{A})$ is the minimal automaton of $\mathcal{L}(\mathcal{A})$.*

This theorem readily yields an algorithm to compute the minimal automaton of the language recognized by an automaton $\mathcal{A}$, based on the subset construction: since $\mathcal{B} = \mathrm{Subset}(\mathrm{Trim}(\tilde{\mathcal{A}}))$ is a deterministic automaton recognizing the mirror of $\mathcal{L}(\mathcal{A})$, then $\mathrm{Subset}(\mathrm{Trim}(\tilde{\mathcal{B}}))$ is the minimal automaton of $\mathcal{L}(\mathcal{A})$.

2.2 Combinatorial structures

Permutations. A *permutation* of size n is a bijection from $[n]$ to $[n]$. A size-n permutation σ can be represented by a directed graph of set of vertices $[n]$, with an edge $i \to j$ whenever $\sigma(i) = j$. As σ is a bijection, such a graph is always a union of cycles. The *order* of a permutation is the smallest positive integer m

[1] If $u = u_0 \dots u_{n-1}$ is a word of length n, the *mirror* of u is the word $\tilde{u} = u_{n-1} \dots u_0$.
[2] Minimal automata are not always required to be complete in the literature.

such that σ^m the identity. It is equal to the least common multiple (lcm) of the lengths of its cycles.

Mappings. A *mapping* of size n is a total function from $[n]$ to $[n]$. As done for permutations, a mapping f can be seen as a directed graph with an edge $i \to j$ whenever $f(i) = j$. Such a graph is no longer a union of cycles, but a union of cycles of trees (trees whose roots are linked into directed cycles), as depicted in Fig. 2. Let f be a size-n mapping. An element $x \in [n]$ is a *cyclic point* of f when there exists an integer $i > 0$ such that $f^i(x) = x$. The *cyclic part* of a mapping f is the permutation obtained when restricting f on its set of cyclic points. The *normalized cyclic part* of f is obtained by relabelling the c cyclic points of f by elements of $[c]$ while keeping their relative order (see Fig 2).

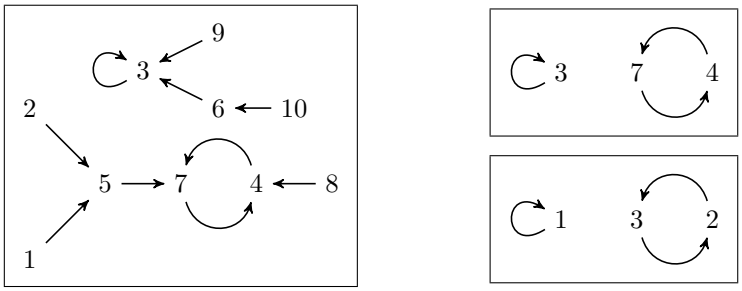

Fig. 2. A mapping of $\{1, \ldots, 10\}$ seen as a directed graph on the left. Its cyclic part is depicted on the upper right, and its normalized cyclic part on the lower right. The normalization is obtained by relabelling the 3 vertices with elements of $\{1, 2, 3\}$, while keeping the relative order; hence $3 \mapsto 1$, $4 \mapsto 2$ and $7 \mapsto 3$.

Automata as combinatorial structures. In the sequel, A is always a fixed alphabet with $k \geq 2$ letters. Let $\mathfrak{A}_n$ denote the set of all deterministic and complete automata on A whose set of states is $[n]$ and whose initial state is 1. Such an automaton $\mathcal{A}$ is characterized by the tuple (n, δ, F). A *transition structure* is an automaton without final states, and we denote by $\mathfrak{T}_n$ the set of n-state transition structures with the same label restrictions as for $\mathfrak{A}_n$. If $\mathcal{A} \in \mathfrak{A}_n$, an a-cycle of $\mathcal{A}$ is a cycle of the mapping induced by a, i.e. $p \mapsto \delta(p, a)$. If $\mathcal{C}$ is an a-cycle of length ℓ, the *word associated to* $\mathcal{C}$ is the word u of length ℓ on the alphabet $\{0, 1\}$ obtained as follows: if x is the smallest element of $\mathcal{C}$, $u_i = 1$ if and only if $\delta(x, a^i) \in F$, for $i \in \{0, \ldots, \ell - 1\}$. In other words, one starts at x and follows the cycle, writing a 1 when the current state is final and a 0 otherwise. An a-cycle is *primitive* when its associated word u is primitive, that is, when u cannot be written $u = v^m$ for some word v and some integer $m \geq 2$.

2.3 Probabilities on automata and genericity

A *probabilistic model* is a sequence $(\mathbb{P}_n)_{n\geq 1}$ of probability measures on the same space. A property P is said to be *generic* for the probabilistic model $(\mathbb{P}_n)_{n\geq 1}$ when the probability that P is satisfied tends to 1 as n tends to infinity.

In our settings, we work on a set E of combinatorial objects with a notion of size, and we will only consider probabilistic models where the support of $\mathbb{P}_n$ is the finite set E_n of size-n objects. The *uniform model* (or *uniform distribution* which is a slight abuse of notation since there is one distribution for each n) on a set $E = \cup_{n\geq 1} E_n$ is defined for any $e \in E_n$ by $\mathbb{P}_n(\{e\}) = \frac{1}{|E_n|}$. The reader is referred to [12] for more information on combinatorial probabilistic models.

For any $0 < b < 1$, the *Bernoulli model of parameter* b (or just a *Bernoulli model* for short) on deterministic automata is the model where an automaton of size-n is obtained by first drawing an element of $\mathfrak{T}_n$ under the uniform distribution, then choosing whether each state is final or not with probability b, independently: the probability of an element $\mathcal{A} \in \mathfrak{A}_n$ with f final states is by definition $\frac{b^f(1-b)^{n-f}}{|\mathfrak{T}_n|}$. The uniform distribution on deterministic automata is obtained by choosing $b = \frac{1}{2}$.

3 Main results

Our main result is Theorem 2 below, which gives a super-polynomial lower bound for the generic number of states of the minimal automaton of the mirror.

Theorem 2. *Consider a Bernoulli model for automata on an alphabet with at least two letters. For any $d > 0$, the minimal automaton of the mirror of $\mathcal{L}$, where $\mathcal{L}$ is the language recognized by a random deterministic n-state automaton, generically has a super-polynomial number of states.*

This directly yields the generic complexity of Brzozowski algorithm, and therefore its average case complexity. It also emphasizes that, in our case, the generic complexity analysis is more precise than the average case analysis: a negligible proportion of bad cases could also have lead to a bad average complexity.

Corollary 1 (Average complexity). *For any fixed alphabet with at least two letters, the generic and average complexity of Brzozowski algorithm is super-polynomial for Bernoulli models on deterministic automata.*

Proof. It is generically super-polynomial by Theorem 2. Hence for any $d > 0$, the complexity is greater than n^{d+1} with probability more than $\frac{1}{2}$, for n large enough. Thus, the average complexity is bounded from below by $\frac{1}{2}n^{d+1} > n^d$ for n large enough. □

Lemma 1 below is the main ingredient of the proof of Theorem 2, as it allows to focus on a-cycles only, which contains enough information to exhibit a lower bound. The other letters are necessary to prove that such a-cycles are accessible in a random automaton, as we shall see in Section 4.

Lemma 1. *Let $\mathcal{A} \in \mathfrak{A}_n$ be a deterministic automaton that contains m primitive a-cycles $\mathcal{C}_1, \ldots \mathcal{C}_m$ of length at least two that are all accessible. The minimal automaton of $\mathcal{L}(\tilde{\mathcal{A}})$ has at least $lcm(|\mathcal{C}_1|, \ldots, |\mathcal{C}_m|)$ states.*

Proof. By Theorem 1, the minimal automaton of the mirror of $\mathcal{L}(\mathcal{A})$ is obtained by determinizing the reverse of the accessible part of $\mathcal{A}$. Since the a-cycles are accessible, they are still there after removing the non-accessible part. Moreover, as they are primitive and of length at least two, they necessarily contain at least one final state. Hence, they are also co-accessible.

Let $\mathcal{C} = \cup_{j\in[m]}\mathcal{C}_j$ and let σ be the permutation of $\mathcal{C}$ defined by $\sigma(x) = y$ if and only if $\delta(y, a) = x$. This permutation is well defined, since every element of $\mathcal{C}$ has a unique preimage by a that lies in $\mathcal{C}$. We are interested in the natural action of σ on the subsets of $\mathcal{C}$: let F be the set of final states of $\mathcal{A}$, which is also the set of initial states of $\tilde{\mathcal{A}}$, and consider the set $X = \mathcal{C} \cap F$. Let ℓ be the size of the orbit of X under the action of $< \sigma >$. We have $\sigma^\ell(X) = X$. Let $\mathcal{C}_j$ be one of the cycles and let $X_j = \mathcal{C}_j \cap X$. The set $\mathcal{C}_j$ is stable under the action of σ, and $X_j \subseteq X$, thus $\sigma^\ell(X_j) = X_j$. Hence, the size of its orbit under the action of $< \sigma >$ divides ℓ. Moreover, since $\mathcal{C}_j$ is primitive, there are exactly $|\mathcal{C}_j|$ elements in the orbit of X_j, and thus $|\mathcal{C}_j|$ divides ℓ for every $j \in [m]$. Hence, ℓ is the lcm of the cycles' lengths. Therefore, by looking at the intersection of $\tilde{\delta}(F, a^i)$ with $\mathcal{C}$, for $i \geq 0$, there are at least $\mathrm{lcm}(|\mathcal{C}_1|, \ldots, |\mathcal{C}_m|)$ accessible states in $\mathrm{Subset}(\tilde{\mathcal{A}})$. □

4 Accessibility in random transition structures

The very first part of the algorithm is to remove useless states, and in particular states that are not accessible. The precise study of the number of accessible states in a random transition structure has been done in [6]: if X_n is the random variable associated with the number of accessible states, the expectation of X_n is equivalent to $v_k \cdot n$, for some explicit constant v_k, and the distribution is asymptotically Gaussian. In the sequel, we only need the following weaker result established in [6]:

Lemma 2. *There exists two real numbers α and β, with $0 < \alpha < \beta < 1$ such that the number of accessible states in a random transition structure of size n is generically in the interval $[\alpha n, \beta n]$.*

In order to use Lemma 1, we need to exhibit large enough primitive a-cycles in a random deterministic automaton in the proof of Theorem 2. This can only work if those cycles are in the accessible part of the automaton, which is established in Proposition 1 below. The proof directly follows a more general idea given by Andrea Sportiello in a private communication.

Proposition 1. *For the uniform distribution on transition structures of size n, all the a-cycles of lengths greater than $\log n$ are generically accessible.*

Proof. Let $i \in [n]$ and let $\mathcal{A}$ be an accessible transition structure with i states, whose states labels are in $[n]$ and such that 1 labels the initial state. By a

direct counting argument [6], there are exactly $n^{k(n-i)}$ transition structures in $\mathfrak{T}_n$ whose accessible part is $\mathcal{A}$. Let us bound from above the number of such automata having a non-accessible a-cycle of size ℓ: to create such a cycle, one need to choose the ℓ state labels not in the accessible part and how these states are circularly linked using transitions labelled by a. Other transitions can end at any of the n states. There are therefore no more than $\binom{n-i}{\ell}(\ell-1)!\cdot n^{k(n-i)-\ell}$ possibilities. Hence, the probability that it happens, conditioned by having $\mathcal{A}$ as accessible part, is bounded from above by

$$\binom{n-i}{\ell}(\ell-1)!\; n^{-\ell} = \frac{n^{-\ell}}{\ell}(n-i)(n-i-1)\cdots(n-i-\ell+1) \leq (n-i)^{\ell}n^{-\ell}.$$

This bound only depends on the size of the accessible part. Let X_n be the random variable associated with the number of states in the accessible part of a random transition structure. Using the formula above, the probability of having an a-cycle of length equal to ℓ that is not accessible and at least αn accessible states is bounded from above by[3]

$$\sum_{i=\alpha n}^{n} (n-i)^{\ell}n^{-\ell}\cdot \mathbb{P}(X_n = i) \leq (1-\alpha)^{\ell}.$$

Hence the probability of having a non-accessible a-cycle of length at least ℓ and at least αn accessible states is bounded from above by $\sum_{j=\ell}^{(1-\alpha)n}(1-\alpha)^j$ which tends to 0 as ℓ tends to infinity, as the remainder of a converging series. This concludes the proof, using $\ell = \log n$, since by Lemma 2, the accessible part of a transition structure generically has more than αn states. □

5 Proof of Theorem 2

Our proof of Theorem 2 relies on Lemma 1 and on a famous theorem of Erdős and Turán: let O_n be the random variable associated with the order of a random permutation of size n. Erdős and Turán theorem states that the mean value of $\log O_n$ is equivalent to $\frac{1}{2}\log^2 n$, and that when normalized[4], it converges in distribution to the normal law. In the sequel, we shall only need an intermediate result they use to establish their proof, which is the following [10, Eq. (14.3)]:

Proposition 2 (Erdős and Turán). *For the uniform distribution, the order of a random permutation of size n is generically greater than* $\exp(\frac{1}{3}\log^2 n)$.

The idea is to use Proposition 2 to quantify the lcm of the primitive accessible a-cycles in a random automaton, under the Bernoulli model. This requires some care, since not all a-cycles are necessarily accessible or primitive.

[3] For readability we have not use integer parts in the bounds, here and in the sequel; this does not change the results.

[4] Centered around its means and divided by its standard deviation.

5.1 Accessible a-cycles

We first focus on the shape of random automata and therefore work on transition structures. By Proposition 1, all a-cycles of length greater than $\log n$ are generically accessible, and we need to exhibit enough such cycles.

The action of letter a in a uniform element of $\mathfrak{T}_n$ is a uniform size-n random mapping. These objects have been studied intensively, and their typical properties are well-known [11]. We shall need the two following results in the sequel.

Lemma 3. *For any $\epsilon > 0$, the number of cyclic points of a size-n random mapping is generically greater than $n^{\frac{1}{2}-\epsilon}$.*

Proof. Let α be a real number such that $\frac{1}{2} - \epsilon < \alpha < \frac{1}{2}$. Let f be a mapping of size n, and consider the sequence 1, $f(1)$, $f^2(1) = f(f(1))$, ... At some point, $f^i(1)$ is for the first time equal to a $f^j(1)$ for $j < i$, and we have a cycle of length $i-j+1$. This reduces the problem to the Birthday Paradox: we repeatedly draw a random number from $[n]$ (the image of the new iteration of f) until a number is seen twice. Let X_n be the random variable associated with the number of distinct numbers in the sequence, we classically have:

$$\mathbb{P}(X_n \geq m) = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\ldots\left(1 - \frac{m-1}{n}\right).$$

Moreover, for $x \in (0, \frac{1}{2})$, $1 - x \geq \exp(-2x)$, and therefore, for $m \leq \frac{1}{2}n$ we have

$$\mathbb{P}(X_n \geq m) \geq \exp\left(-\frac{2}{n}\sum_{i=1}^{m-1} i\right) = \exp\left(-\frac{m(m-1)}{n}\right).$$

Hence $\mathbb{P}(X_n < n^\alpha) = \mathcal{O}(n^{2\alpha-1})$, and since $\alpha < \frac{1}{2}$, there are generically more than n^α distinct iterations. Moreover, by symmetry, if $f^i(1) = f^j(1)$ is the first collision, j is a uniform element of $\{0, \ldots, i-1\}$. Since $n^{\frac{1}{2}-\epsilon}$ is significantly smaller than n^α, the collision is generically not on one of the $n^{\frac{1}{2}-\epsilon}$ last iterations, and the cycle is of length greater than $n^{\frac{1}{2}-\epsilon}$. This concludes the proof, since the number of cyclic points is at least the length of this cycle. □

Lemma 4. *Let $i \in [n]$ and let σ and τ be two permutations of $[i]$. The probability that the normalized cyclic permutation of a uniform size-n random mapping is σ is equal to the probability it is τ.*

Proof. (sketch) This is a folklore result. From its graph representation, one can see that a mapping is uniquely determined by its set $\mathcal{T} = \{T_1, \ldots, T_m\}$ of trees and the permutation of their roots. Conditioned to have $\mathcal{T}$ as set of trees, the normalized cyclic permutation of a random mapping is therefore a uniform permutation. The result follows directly, by the law of total probabilities. □

Hence generically, the number of a-cyclic states is greater than, say, $n^{\frac{1}{3}}$, and conditioned by its size, the normalized cyclic permutation of a random mapping

follows the uniform distribution. We can therefore use the statistical properties of random uniform permutations, which are very well-known as well. In particular, we shall need the following generic upper bound for the number of cycles.

Lemma 5. *For the uniform distribution, a size-n random permutation generically has less than $2\log n$ cycles.*

Proof. The expectation and standard deviation of the number of cycles in a random permutation are well-known (see for instance [12, Example IX.9 p. 644]) and are respectively equivalent to $\log n$ and $\sqrt{\log n}$. It implies by Chebyshev's inequality that a random permutation has generically less than $2\log n$ cycles. □

The following proposition summarizes the results collected so far.

Proposition 3. *Generically in a random size-n transition structure, there are more than $n^{\frac{1}{3}}$ a-cyclic states, organized in less than $2\log n$ a-cycles, and all a-cycles of length greater than $\log n$ are accessible.*

Proof. Let ℓ be an integer such that $n^{\frac{1}{3}} < \ell \leq n$. Let C_n be the random variable associated with the number of a-cyclic points in a random size-n transition structure. Let also N_ℓ be the random variable associated with the number of cycles in a random permutation of size ℓ. By Lemma 5, there exists a non-increasing sequence $(\epsilon_n)_{n\geq 1}$ that tends to 0 such that

$$\mathbb{P}(N_\ell < 2\log \ell) \geq 1 - \epsilon_\ell.$$

Let $\mathfrak{G}_n \subseteq \mathfrak{T}_n$ denote the set of transition structures with more than $n^{\frac{1}{3}}$ a-cyclic states that are organized in less than $2\log n$ a-cycles. If T_n represents an element of $\mathfrak{T}_n$ taken uniformly at random, we have

$$\mathbb{P}(T_n \in \mathfrak{G}_n) = \sum_{\ell=n^{1/3}}^{n} \mathbb{P}(T_n \in \mathfrak{G}_n \mid C_n = \ell) \cdot \mathbb{P}(C_n = \ell).$$

By Lemma 4, $\mathbb{P}(T_n \in \mathfrak{G}_n \mid C_n = \ell) = \mathbb{P}(N_\ell < 2\log n) \geq \mathbb{P}(N_\ell < 2\log \ell)$, since under the condition $C_n = \ell$, the a-cyclic part of T_n is a uniform random permutation of length ℓ. Hence

$$\mathbb{P}(T_n \in \mathfrak{G}_n) \geq (1 - \epsilon_{n^{1/3}}) \sum_{\ell=n^{1/3}}^{n} \mathbb{P}(C_n = \ell) = (1 - \epsilon_{n^{1/3}}) \cdot \mathbb{P}(C_n \geq n^{1/3}).$$

Hence, by Lemma 3, a random transition structure is generically in $\mathfrak{G}_n$. This concludes the proof, since by Proposition 1, all a-cycles of length greater than $\log n$ are generically accessible. □

5.2 Lcm of truncated random permutations

Since we cannot guarantee that small cycles are accessible in a typical transition structure, we need to adapt Proposition 2 to obtain the needed lower bound for the lcm of the lengths of accessible a-cycle. In a size-n permutation, a *large cycle* (resp. *small cycle*) denote a cycle of length greater than (resp. at most) $3\log n$.

Lemma 6. *The lcm of the lengths of the large cycles in a uniform random permutation of size n is generically greater than* $\exp(\frac{1}{4}\log^2 n)$.

Proof. By Lemma 5 there are generically less than $2\log n$ cycles in a random permutation. The number of points in small cycles is therefore generically bounded from above by $6(\log n)^2$. For a given permutation, we split the lengths of its cycles into two sets L and S, whether they are greater than $3\log n$ or not. The order of the permutation is the lcm of the lengths of its cycles, and is therefore bounded from above by $\mathrm{lcm}(L)\cdot\mathrm{lcm}(S)$. Hence

$$\mathrm{lcm}(L) \geq \frac{\mathrm{lcm}(L\cup S)}{\mathrm{lcm}(S)}.$$

By Landau's theorem [15], the maximal order of a permutation of length ℓ is equivalent to $\exp(\sqrt{\ell}\log\ell)$ and therefore bounded from above by $2\exp(\sqrt{\ell}\log\ell)$ for large enough ℓ. Hence, the less than $6(\log n)^2$ points in small cycles form a permutation whose order, which is equal to $\mathrm{lcm}(S)$, is bounded from above by $2\exp(\sqrt{6\log^2 n}\log(6\log^2 n))$. Using this bound and Proposition 2 yields the result: for n large enough, we have a generic lower bound of

$$\begin{aligned}\frac{\exp(\frac{1}{3}\log^2 n)}{2\exp(\sqrt{6\log^2 n}\log(6\log^2 n))} &= \frac{1}{2}\exp\left(\frac{1}{3}\log^2 n - \sqrt{6}\log n\log(6\log^2 n)\right)\\ &\geq \exp\left(\frac{1}{4}\log^2 n\right).\end{aligned}$$

5.3 Primitivity

One last effort is required, as we need to take final states into account and prove the generic primitivity of large cycles in a uniform random permutation under the Bernoulli model. Recall that an a-cycle of final and non-final states is encoded by a word with 1's and 0's (see Section 2.2). The following lemma establishes the needed result.

Lemma 7. *Generically, the a-cycles of length greater than $\log n$ in a random automaton with n states are all primitive.*

Proof. We first follow [7] for words on $\{0,1\}$ under the Bernoulli model of parameter $b\in(0,1)$: if a word u of length n is not primitive, there exist an integer $d\geq 2$ and a word v of length n/d such that $u=v^d$. For such a fixed v with z zeros, the probability that $u=v^d$ is $(1-b)^{dz}b^{n-dz}$. Since there are exactly $\binom{n/d}{z}$ such v, the probability that u is the d-power of a word, for any fixed $d\geq 2$ that divides n, is

$$\sum_{z=0}^{n/d}\binom{n/d}{z}(1-b)^{dz}b^{n-dz} = (b^d+(1-b)^d)^{\frac{n}{d}}.$$

Hence the probability that u is not primitive is bounded from above by the sum of $(b^d+(1-b)^d)^{\frac{n}{d}}$ for $2\leq d\leq n$, which is smaller than $\alpha\lambda^n$, for $\lambda=\sqrt{b^2+(1-b)^2}$

and for some constant $\alpha > 0$. Then, each a-cycle of length greater than $\log n$ is non-primitive with probability bounded from above by $\alpha \log n \cdot \lambda^{\log n}$. By Proposition 3, the probability ϵ_n that there are more than $2 \log n$ a-cycles tends to 0. Hence, the probability of having a non-primitive a-cycle of length greater than $\log n$ is bounded from above by $2\alpha(\log n)^2 \lambda^{\log n} + \epsilon_n$, which tends to 0. □

5.4 Conclusion of the proof

We now have all the ingredients to establish the proof of Theorem 2. By Proposition 3, the a-cycles of a random automaton $\mathcal{A}$ generically form a random permutation of size greater that $n^{\frac{1}{3}}$. Therefore, the large a-cycles are generically of length greater than $3 \log n^{\frac{1}{3}} = \log n$. Since a-cycles of size greater than $\log n$ are generically accessible and primitive by Proposition 1 and Lemma 7, the lcm of the large cycles' lengths is a lower bound for the number of states of the minimal automaton of $\mathcal{L}(\tilde{\mathcal{A}})$, by Lemma 1.

By lemma 4, conditioned by its size, the a-cyclic permutation is a uniform permutation. Using the law of total probability and Lemma 6 we therefore obtain that there are generically more than $\exp(\frac{1}{4} \log^2 n^{\frac{1}{3}})$ states in the minimal automaton of $\mathcal{L}(\tilde{\mathcal{A}})$, concluding the proof.

6 Conclusion and perspectives

In this article we have found generic super-polynomial lower bounds for the mirror operator and for the complexity of Brzozowski algorithm. These results hold for deterministic automata under Bernoulli models, where the shape of the automaton is chosen uniformly at random, and where each state is final with a fixed probability $b \in (0, 1)$.

These probabilistic models are interesting since they contain the uniform distribution on deterministic automata. It is however natural to consider other distribution on automata, and we propose two directions.

The first idea is to change the distribution on final states, in order to have less final states in a typical automaton. Remark that our results and our proofs still hold if $b := b_n$ depends on n, provided there exists $0 < \alpha < \frac{1}{2}$ such that both b_n and $1 - b_n$ are in $\Omega(\frac{1}{n^\alpha})$. This only require to revisit Lemma 7, and to use the $n^{\frac{1}{2}-\epsilon}$ of Lemma 3 instead of the $n^{\frac{1}{3}}$ as we did in this paper. However our proof technique does not work for smaller probabilities such as $b_n = \frac{1}{n^{2/3}}$ because with high probability, the a-cycles have no final states. Trying to handle such distributions with a small number of final states is ongoing work. Note that the other works on random automata [1, 8, 2] also face the same limitation.

The other natural idea is to consider the uniform distribution on accessible deterministic automata and not on deterministic automata. The combinatorics of accessible deterministic automata is more involved [14, 4], but it is sometimes possible to deduce generic properties for the distribution on accessible deterministic automata from the distribution on accessible automata [6]. In our case, this would require to prove that the error terms are all in $o(\frac{1}{\sqrt{n}})$.

Acknowledgment. We would like to thanks Andrea Sportiello for the fruitful technical discussion we had in Cluny, which, amongst many other things, lead to the proof of Proposition 1.

References

1. F. Bassino, J. David, and C. Nicaud. Average case analysis of Moore's state minimization algorithm. *Algorithmica*, 63(1-2):509–531, 2012.
2. F. Bassino, J. David, and A. Sportiello. Asymptotic enumeration of minimal automata. In Dürr and Wilke [9], pages 88–99.
3. F. Bassino, L. Giambruno, and C. Nicaud. The average state complexity of rational operations on finite languages. *International Journal of Foundations of Computer Science*, 21(4):495–516, 2010.
4. F. Bassino and C. Nicaud. Enumeration and random generation of accessible automata. *Theor. Comput. Sci.*, 381(1-3):86–104, 2007.
5. J. A. Brzozowski. Canonical regular expressions and minimal state graphs for definite events. In *Mathematical theory of Automata*, pages 529–561. Polytechnic Press, Polytechnic Institute of Brooklyn, N.Y., 1962. Volume 12 of MRI Symposia Series.
6. A. Carayol and C. Nicaud. Distribution of the number of accessible states in a random deterministic automaton. In Dürr and Wilke [9], pages 194–205.
7. P. Chassaing and E. Z. Azad. Asymptotic behavior of some factorizations of random words, 2010. arXiv:1004.4062v1.
8. J. David. Average complexity of Moore's and Hopcroft's algorithms. *Theor. Comput. Sci.*, 417:50–65, 2012.
9. C. Dürr and T. Wilke, editors. *29th International Symposium on Theoretical Aspects of Computer Science, STACS 2012, February 29th - March 3rd, 2012, Paris, France*, volume 14 of *LIPIcs*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2012.
10. P. Erdős and P. Turán. On some problems of a statistical group-theory I. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 4:175–186, 1965.
11. P. Flajolet and A. M. Odlyzko. Random mapping statistics. In J.-J. Quisquater and J. Vandewalle, editors, *EUROCRYPT*, volume 434 of *Lecture Notes in Computer Science*, pages 329–354. Springer, 1989.
12. P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.
13. J. E. Hopcroft. An $n \log n$ algorithm for minimizing the states in a finite automaton. In Z. Kohavi, editor, *The Theory of Machines and Computations*, pages 189–196. Academic Press, 1971.
14. A. Korshunov. Enumeration of finite automata. *Problemy Kibernetiki*, 34:5–82, 1978. in russian.
15. E. Landau. *Handbuch der lehre von der verteilung der primzahlen*, volume 2. B. G. Teubner, 1909.
16. C. Nicaud. Average state complexity of operations on unary automata. In M. Kutylowski, L. Pacholski, and T. Wierzbicki, editors, *MFCS*, volume 1672 of *Lecture Notes in Computer Science*, pages 231–240. Springer, 1999.
17. D. Tabakov and M. Y. Vardi. Experimental evaluation of classical automata constructions. In *In LPAR 2005, LNCS 3835*, pages 396–411. Springer, 2005.