Distribution of the number of accessible states in a random deterministic automaton

Arnaud Carayol, Cyril Nicaud

Distribution of the number of accessible states in a random deterministic automaton

Arnaud Carayol[1] **and Cyril Nicaud**[1]

1 **Université Paris-Est, LIGM, CNRS** {carayol,nicaud}@univ-mlv.fr

Abstract

We study the distribution of the number of accessible states in deterministic and complete automata with n states over a k-letters alphabet. We show that as n tends to infinity and for a fixed alphabet size, the distribution converges in law toward a Gaussian centered around $v_k n$ and of standard deviation equivalent to $\sigma_k\sqrt{n}$, for some explicit constants v_k and σ_k. Using this characterization, we give a simple algorithm for random uniform generation of accessible deterministic and complete automata of size n of expected complexity $O(n\sqrt{n})$, which matches the best methods known so far. Moreover, if we allow a ε variation around n in the size of the output automaton, our algorithm is the first solution of linear expected complexity. Finally we show how this work can be used to study accessible automata (which are difficult to apprehend from a combinatorial point of view) through the prism of the simpler deterministic and complete automata. As an example, we show how the average complexity in $O(n \log\log n)$ for Moore's minimization algorithm obtained by David for deterministic and complete automata can be extended to accessible automata.

1998 ACM Subject Classification F.2 Analysis of algorithms and problem complexity

Keywords and phrases finite automata, random sampling, average complexity

Digital Object Identifier 10.4230/LIPIcs.STACS.2012.194

1 Introduction

The structure of an automaton with n states over a k-letter alphabet is simply a deterministic and complete finite automaton with states in $[n] = \{1, \dots, n\}$ over the alphabet $\{a_1, \dots, a_k\}$. The state 1 is always assumed to be the unique initial state. We do not take final states into account as we are only interested in the structure of the automaton and not in the accepted language. We denote by $\mathcal{T}_{n,k}$ (or $\mathcal{T}_n$ if k is understood) the set of all such transition structures. As structures in $\mathcal{T}_{n,k}$ can alternatively be described by k-tuples of mappings from $[n]$ to $[n]$ (*i.e.* the i-th mapping corresponds to the action of the transitions labeled by a_i), the cardinal of $\mathcal{T}_{n,k}$ is $|\mathcal{T}_{n,k}| = n^{kn}$. An accessible automaton is a structure in $\mathcal{T}_{n,k}$ such that all states are accessible from the initial state 1. We denote by $\mathcal{A}_{n,k}$ (or $\mathcal{A}_n$ if k is understood) the set of all accessible automata in $\mathcal{T}_{n,k}$. The accessible automaton of a structure in $\mathcal{T}_{n,k}$ is obtained by restricting the structure to its

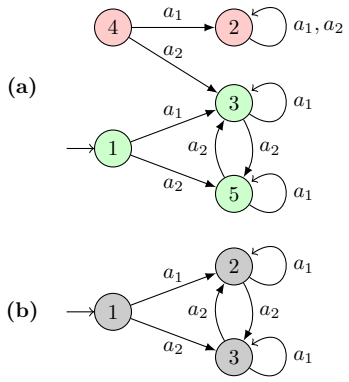

Figure 1 A structure **(a)** of $\mathcal{T}_{5,2}$ (a) and its accessible automaton **(b)** in $\mathcal{A}_{3,2}$.

29th Symposium on Theoretical Aspects of Computer Science (STACS'12).
Editors: Christoph Dürr, Thomas Wilke; pp. 194–205
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany
SYMPOSIUM ON THEORETICAL ASPECTS OF COMPUTER SCIENCE

set $\{s_1 < \ldots < s_j\}$ of accessible states and by renaming the state s_i as i for all $i \in [j]$, as depicted in Fig. 1. Since $s_1 = 1$, the resulting automaton belongs to $\mathcal{A}_{j,k}$.

In this article, we study the distribution of the size of the accessible automaton of a random structure of $\mathcal{T}_{n,k}$, as n tends to infinity. The alphabet size $k \geq 2$ is assumed to be fixed and in particular does not depend on n (the case $k = 1$ is quite different and can be analyzed using known results on random mappings [9]). For all $n \geq 1$, we consider the random variable X_n describing the size of the accessible automaton of a structure in $\mathcal{T}_{n,k}$, for the uniform distribution. The probability for X_n to take value i, for $i \in [n]$, is given by the following formula first obtained in [16]:

$$P(X_n = i) = \frac{\overbrace{\binom{n-1}{i-1}}^{\text{labels of acc. states}} \cdot \overbrace{|\mathcal{A}_{i,k}|}^{\text{acc. aut.}} \cdot \overbrace{n^{k(n-i)}}^{\text{remaining transitions}}}{n^{kn}}. \tag{1}$$

Indeed if we fix the accessible automaton, it remains to choose the labels in $[n]$ for its states (as the initial state is always labeled by 1, we have $\binom{n-1}{i-1}$ choices) and the target for the $k(n-i)$ transitions that take their source outside of the accessible component (kn total transitions for the structure minus the ki of the accessible automaton). For these transitions, all n choices of target are valid. An important consequence of this formula is that two accessible automata in $\mathcal{A}_{i,k}$ appear in the same number of structures of $\mathcal{T}_{n,k}$, for any $i \in [n]$.

n	100	1000	10000
$\mathbb{E}[X_n]$(k=2)	79.6356	796.663	7967.41
$\mathbb{E}[X_n]$(k=3)	94.0138	940.489	9404.40
$\mathbb{E}[X_n]$(k=4)	97.9746	980.137	9801.89

k	2	3	4
v_k	0.796812	0.940479	0.980176

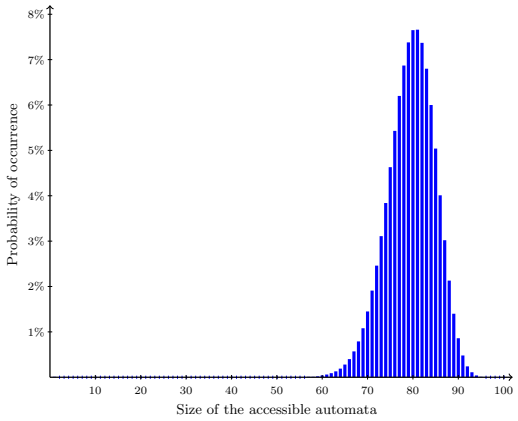

Figure 2 On the top left, an approximation of the average size of the accessible automaton based on 10000 randomly generated structures from $\mathcal{T}_{n,k}$. On the bottom left, the values of the constant $v_k = 1 + \frac{1}{k}W_0(-ke^{-k})$ for different values of k. On the right, the graphical representation of X_{100}.

Our main technical contribution is to describe the law of X_n for large values of n. As hinted by Fig. 2, we show that the average size of the accessible automaton $\mathbb{E}[X_n]$ is equivalent to $v_k n$ where v_k is a constant depending[1] on the size of the alphabet k. We also show that the standard deviation is equivalent to $\sigma_k\sqrt{n}$, where σ_k is also a constant depending on k. As shown in Fig. 2, the shape of the repartition of the size of the accessible automaton for a fixed n looks like a Gaussian. This type of behavior is quite common with combinatorial objects: it is the case for instance for the number of cycles in a random permutation of size n, for the number of occurrences of a fixed pattern in a random string of length n, ... (see

[1] Recall k is assumed to be fixed in our asymptotic analysis.

[10, p. 683]). It is formally captured by the notion of convergence in distribution to the normal (or Gaussian) law. More precisely, we are going to show that in a random structure of size n, once it has been centered by its mean and scaled by its standard deviation, the distribution of the size of the accessible automaton is asymptotically Gaussian. Note that standard analytic methods [10] cannot be directly applied here since there are no known expressions for the associated generating function.

Our interest in studying the distribution of the size of the accessible automaton is not only motivated by its fundamental nature but also by its rich implications in the algorithmic and combinatorial study of accessible automata. To substantiate our claim, we provide three applications of our theoretical results.

Our first application deals with the problem of uniform generation of deterministic and complete accessible automata. This problem is declined in two variants: the exact one and the ε-approximated one for $\varepsilon \in (0,1)$. The exact generation problem asks to generate uniformly at random an automaton of $\mathcal{A}_m$, for a given size $m \geq 1$, whereas the ε-approximated one asks for an automaton in $\mathcal{A}_{m'}$ for some $m' \in [(1-\varepsilon)m, (1+\varepsilon)m]$ where m is given; the ε-approximated also requires that two automata of the same size have the same probability to be generated. The first solution [17, 4] to the exact generation problem, based on an adaptation of the recursive method [18], has a complexity in $O(m^2)$ (consisting of a preprocessing in $O(m^2)$ and a computation in $O(m)$). In [1], another solution based on a representation of deterministic and complete accessible automata by words was proposed, with a complexity in $O(m^2)$. This complexity was later improved in [3], using methods based on combinatorial bijections and Boltzmann sampling [7], which gives an expected complexity of $O(m\sqrt{m})$. This last work was then adapted to generate possibly incomplete automata [2], with the same expected complexity. Note that the best known upper-bound for the ε-approximated problem is also $O(m\sqrt{m})$ as all known solutions to the ε-approximated problem are in fact solutions to the exact problem.

We propose a very simple algorithm for generating accessible automata of size m whose expected complexity $O(m\sqrt{m})$ matches the best known upper bounds. This algorithm consists in generating uniformly at random a transition structure in $\mathcal{T}_{n,k}$ with $n = \left\lfloor \frac{m}{v_k} \right\rfloor$ states and then to compute its accessible automaton. If it is of size m we output it, and otherwise we restart the process. The correctness of the algorithm follows from the above remark that two accessible automata of size m appear as accessible automata in the same number of structures of $\mathcal{T}_n$. The probability to obtain an accessible automaton of size exactly m is in $\Theta(\frac{1}{\sqrt{m}})$ and hence the average number of iterations of the algorithm is in $O(\sqrt{m})$. As every iteration can be computed in linear time, the expected complexity is in $O(m\sqrt{m})$.

Slightly modifying the algorithm to output the automaton when its size belongs to the interval $[(1-\varepsilon)m, (1+\varepsilon)m]$ yields a solution to the ε-approximation problem with an expected complexity of $O(m)$. We also show that this algorithm can readily be adapted to generate minimal automata (using a recent result on the asymptotic number of minimal automata [11]), with the same expected complexity for the exact version and an expected complexity in $O(n \log\log n)$ for the approximated version.

The second application concerns the formula expressing the asymptotic number of automata in $\mathcal{A}_{n,k}$, as n tends to infinity. In [14], Korshunov established that:

$$|\mathcal{A}_{n,k}| \sim E_k n! \left\{ {kn \atop n} \right\} \quad \text{with} \quad E_k = \frac{1 + \sum_{r=1}^{\infty} \frac{1}{r} \binom{kr}{r-1} (e^{k-1}\lambda_k)^{-r}}{1 + \sum_{r=1}^{\infty} \binom{kr}{r} (e^{k-1}\lambda_k)^{-r}} \text{ and } \lambda_k = \frac{e^{kv_k} - 1}{e^{k-1} v_k^k}, \quad (2)$$

where $\left\{ {kn \atop n} \right\}$ designates the Stirling numbers of the second kind: $\left\{ {kn \atop n} \right\}$ is the number of different

ways to partition kn elements into n non-empty sets. Recently in [15], Lebensztayn gave a simplified expression of the constant E_k, using the theory of Lagrange inversion applied in the context of generalized binomial series. Using our main result and the simple fact that $\sum_{i=1}^{n} P(X_n = i) = 1$, we obtain another proof of his simplified expression for E_k. Note that we do use Korshunov's equivalent to obtain our results but never the expression of the constant E_k given in Eq. (2).

The last application and the main perspective of this work is the study of combinatorial properties of accessible automata (which are difficult to apprehend from a combinatorial point of view) through the prism of the simpler structures of $\mathcal{T}_{n,k}$. This approach seems particularly well suited for the average case analysis of classical algorithms on finite automata. We give two examples of asymptotic properties of structures that can be transfered to accessible automata. In particular, we show that the average complexity in $O(n \log\log n)$ for Moore's minimization algorithm recently obtained for structures by David in [6] can be extended, using our result, to accessible automata.

2 Preliminaries

2.1 Deterministic and complete automata

A deterministic and complete transition structure for an automaton over a finite alphabet Γ is a tuple (Q, q_0, δ, F) where Q is a finite set of states, $q_0 \in Q$ is the initial state, $\delta : Q \times \Gamma \mapsto Q$ is the transition function and $F \subseteq Q$ is the set of final sets. For all $n \geq 1$ and $k \geq 2$, we denote by $\mathcal{T}_{n,k}$ the set of all structures over the alphabet $\{a_1, \ldots, a_k\}$ with states in $[n] = \{1, \ldots, n\}$, such that 1 is the initial state and with an empty set of final states.

A structure in $\mathcal{T}_{n,k}$ is said to be accessible (accessible automaton for short) if all its states can be reached from the initial state 1. We denote[2] by $\mathcal{A}_{n,k}$ the set of all accessible automata in $\mathcal{T}_{n,k}$. For a more detailed introduction to finite automata, we refer the reader to [13].

The equivalent of Korshunov given in Eq. (2) involves the Stirling numbers of the second kind $\left\{ {kn \atop n} \right\}$. In [12], Good establishes the following equivalent as n tends to infinity and for a fixed k:

$$\left\{ {kn \atop n} \right\} \sim \frac{(kn)!(e^{\rho_k} - 1)^n}{n! \rho_k^{kn} \sqrt{2\pi kn(1 - ke^{-\rho_k})}} \quad \text{with} \quad \rho_k = k + W_0(-ke^{-k}), \tag{3}$$

where the classical Lambert function W_0 [5] is implicitly defined by $W_0(x)e^{W_0(x)} = x$ and $W_0(x) \geq -1$ for all $x \geq -e^{-1}$. Alternatively, ρ_k is the unique positive solution of $\rho_k = k - ke^{-\rho_k}$.

Using Eq. (3) and Stirling's formula in Korshunov's equivalent (cf. Eq. (2)), we obtain:

$$|\mathcal{A}_{n,k}| \sim E_k \alpha_k \beta_k^n n^{kn} \quad \text{with } \alpha_k = \frac{1}{\sqrt{1 - ke^{-\rho_k}}} \text{ and } \beta_k = \frac{k^k(e^{\rho_k} - 1)}{\rho_k^k e^k}. \tag{4}$$

2.2 Elements of probability

Let us first recall some basic definitions of probability theory (see [8, 10] for more details). If X is a real valued random variable, we denote by $\mathbb{E}[X]$ its expected value and by $\mathbb{V}[X]$ its

[2] Instead of labeled automata, we could consider unlabeled automata: the set $\mathcal{A}^u_{n,k}$ of deterministic and complete automata with n states over $\{a_1, \ldots, a_k\}$ up to isomorphism. As deterministic and accessible automata do not admit non-trivial automorphisms, we have $|\mathcal{A}_{n,k}| = (n-1)!|\mathcal{A}^u_{n,k}|$. Remark that this property does not hold for non-accessible structures or non-deterministic automata.

variance, when they exist. The standard deviation is $\sqrt{\mathbb{V}[X]}$.

▶ **Definition 1.** Let $(X_n)_{n\geq 1}$ be a sequence of real valued random variables and X be a real valued random variable. We say that X_n *converges in distribution to* X when for every $t \in \mathbb{R}$, $P(X_n \leq t) \to P(X \leq t)$ as $n \to \infty$.

▶ **Definition 2.** Let $(X_n)_{n\geq 1}$ be a sequence of random variables such that $\mathbb{E}[X_n]$ and $\mathbb{V}[X_n]$ exist for all $n \geq 1$. We say that X_n *is asymptotically Gaussian* when the standardized random variable $X_n^* = \frac{X_n - \mathbb{E}[X_n]}{\sqrt{\mathbb{V}[X_n]}}$ converges in distribution to the normal distribution $\mathcal{N}(0,1)$ of parameters 0 and 1, defined by, for any $t \in \mathbb{R}$, $P\big(\mathcal{N}(0,1) \leq t\big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx$.

3 Distribution of the size of the accessible component

In this section we state and prove our main result from which we derive all announced properties of this paper. This result, in particular, explains the Gaussian shape of Fig. 2.

▶ **Theorem 3** (Asymptotically Gaussian)**.** *Let X_n be the random variable associated with the size of the accessible part in a structure of $\mathcal{T}_{n,k}$. Then X_n is asymptotically Gaussian with expected value and standard deviation asymptotically equivalent to $v_k\, n$ and $\sigma_k\sqrt{n}$ respectively, with*

$$v_k = 1 + \frac{1}{k} W_0(-ke^{-k}) \quad \textit{and} \quad \sigma_k = \sqrt{\frac{v_k(1-v_k)}{kv_k - k + 1}}. \tag{5}$$

3.1 Outline of the proof of Theorem 3

In this section we present the ideas of the proof of Theorem 3. To shorten the presentation, we do not derive the asymptotic value of the expected value and variance before establishing the convergence in distribution to the Gaussian law. Initially, we estimated the values of the expected value and variance from Eq. (10) and Eq. (11) respectively. As shown in the next section, the proof of Theorem 3 can be reduced to the following statements: for all $\ell \in \{0,1,2\}$ and for all $t \in \mathbb{R}$, as n tends to infinity we have

$$\sum_{i=1}^{\lfloor v_k n\rfloor + \lfloor t\sqrt{n}\rfloor} \left(\frac{i - \lfloor v_k\, n\rfloor}{\sqrt{n}}\right)^{\ell} \cdot P(X_n = i) \;\longrightarrow\; \frac{1}{\sigma_k\sqrt{2\pi}} \int_{-\infty}^{t} x^{\ell} \cdot \exp\left(-\frac{x^2}{2\sigma_k^2}\right) dx, \tag{6}$$

$$\sum_{i=\lfloor v_k n\rfloor + \lfloor t\sqrt{n}\rfloor}^{n} \left(\frac{i - \lfloor v_k\, n\rfloor}{\sqrt{n}}\right)^{\ell} \cdot P(X_n = i) \;\longrightarrow\; \frac{1}{\sigma_k\sqrt{2\pi}} \int_{t}^{\infty} x^{\ell} \cdot \exp\left(-\frac{x^2}{2\sigma_k^2}\right) dx, \tag{7}$$

and there exists a positive constant $C > 0$ such that, for every i and n such that $1 \leq i \leq n$,

$$P(X_n = i) \leq \frac{C}{\sqrt{n}} \quad \text{and} \quad P(X_n = \lfloor v_k n\rfloor) \sim \frac{E_k \alpha_k \sqrt{v_k}}{\sqrt{2\pi n(1-v_k)}}. \tag{8}$$

3.1.1 Expected value and variance

Assuming that Eq. (6), Eq. (7) and Eq. (8) hold, we show how to establish Theorem 3.

For the expected value of X_n, we consider the sum of Eq. (6) and Eq. (7) for $\ell = 1$ and $t = 0$:

$$\frac{\mathbb{E}(X_n) - \lfloor v_k\, n\rfloor}{\sqrt{n}} = \sum_{i=1}^{n} \frac{i - \lfloor v_k\, n\rfloor}{\sqrt{n}} \cdot P(X_n = i) \longrightarrow \frac{1}{\sigma_k\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{x \cdot \exp\left(-\frac{x^2}{2\sigma_k^2}\right)}_{\text{odd function}} dx = 0$$

This proves that $\mathbb{E}[X_n] = \lfloor v_k\, n \rfloor + o(\sqrt{n}) = v_k\, n + o(\sqrt{n})$.

Similarly for the variance of X_n, we consider the sum of Eq. (6) and Eq. (7) for $\ell = 2$ and $t = 0$, we prove $\mathbb{E}\left[(X_n - \lfloor v_k\, n \rfloor)^2\right] \sim \sigma_k^2\, n$. It follows that $\mathbb{V}[X_n] \sim \sigma_k^2\, n$.

For the convergence in distribution to a normal distribution, we have, using the two equivalents obtained previously, that $P(X_n^* \leq t) \sim P(X_n \leq v_k\, n + \sigma_k\, t\sqrt{n})$. The error terms can be handled with Eq. (8), so that using Eq. (6) for $\ell = 0$ gives the result.

Note that we cannot deduce $\mathbb{E}[X_n] \sim v_k\, n$ and $\mathbb{V}[X_n] \sim \sigma_k^2\, n$ from the case $\ell = 0$ only, since the convergence in distribution of Y_n to Y does not necessarily imply that $\mathbb{E}[Y_n] \to \mathbb{E}[Y]$ or $\mathbb{V}[Y_n] \to \mathbb{V}[Y]$. In particular here, one can prove that not all the moments of X_n^* converge.

3.1.2 Reducing the range of the sums of Eq. (6) and Eq. (7)

Our first step of the proof is to show there exist two reals a and b such that $\frac{1}{e} < a < v_k < b < 1$ and

$$\sum_{i=1}^{\lfloor an \rfloor} P(X_n = i) + \sum_{i=\lfloor bn \rfloor}^{n} P(X_n = i) = o\left(\frac{1}{n}\right). \tag{9}$$

This is proved using classical upper bounds for binomial coefficients and for the number of automata in Eq. (1). As $\left(\frac{i - \lfloor v_k\, n \rfloor}{\sqrt{n}}\right)^\ell \in O(n^{\ell/2})$, this also shows that for $\ell \in \{1, 2\}$,

$$\sum_{i=1}^{\lfloor an \rfloor} \left(\frac{i - \lfloor v_k\, n \rfloor}{\sqrt{n}}\right)^\ell P(X_n = i) + \sum_{i=\lfloor bn \rfloor}^{n} \left(\frac{i - \lfloor v_k\, n \rfloor}{\sqrt{n}}\right)^\ell P(X_n = i) = o(1).$$

For the reminder of the proof, we fix a and b and we denote by I_n the set $\{\lfloor an \rfloor + 1, \ldots, \lfloor bn \rfloor\}$.

3.1.3 Equivalent of $P(X_n = i)$ for $i \in I_n$.

Our starting point is Eq. (1), which states that $P(X_n = i) = \frac{i}{n}\binom{n}{i}|\mathcal{A}_{i,k}|n^{-ki}$. For any integer i in I_n we can use the equivalent for $|\mathcal{A}_{i,k}|$ of Eq. (4) and Stirling's formula to obtain the following equivalent of $P(X_n = i)$:

$$P(X_n = i) \sim \frac{E_k \alpha_k}{\sqrt{2\pi n}}\, g\left(\frac{i}{n}\right)\left[f\left(\frac{i}{n}\right)\right]^n, \text{ with } f(x) = \frac{x^{(k-1)x}\beta_k^x}{(1-x)^{1-x}} \text{ and } g(x) = \sqrt{\frac{x}{1-x}}. \tag{10}$$

The constants α_k and β_k of Eq. (4) can be reformulated in terms of v_k using the facts that $v_k = \frac{\rho_k}{k}$ and $v_k = 1 - e^{-kv_k}$. We have $\alpha_k = (1 - ke^{-kv_k})^{-1/2}$ and $\beta_k = \frac{1}{(1-v_k)v_k^{k-1}e^k}$. As i belongs to I_n, $\frac{i}{n}$ belongs to $[a, b]$. On $[a, b]$ the function g is continuous and positive: it has little influence on the analysis. The situation is different for f because it is raised to the power n in the expression. When n grows, the distribution of probabilities is concentrated around the unique point v_k of $[a, b]$ where f reaches its maximum 1. The function f is positive on $[a, b]$, increasing on $[a, v_k]$ and decreasing on $[v_k, b]$, with $f(v_k) = 1$ and $f'(v_k) = 0$, as shown in Fig. 3. Notice that, since g is bounded on $[a, b]$ and

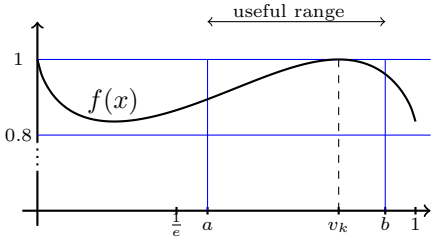

Figure 3 The variations of f on $[0, 1]$.

$|f| \leq 1$ on this interval, we have for all $i \in I_n$ that $P(X_n = i) \leq \frac{C}{\sqrt{n}}$ for some $C \geq 0$. This, and Eq (9) for values of i outside of I_n, proves the first part of Eq. (8).

We now set $i = \lfloor v_k n \rfloor + j$ to center around $\lfloor v_k n \rfloor$. Using Taylor's formula near v_k on $n \ln f(x)$ and remarking that $f''(v_k) = -\frac{1}{\sigma_k^2}$ we get:

$$f\left(\frac{\lfloor v_k n \rfloor + j}{n}\right)^n = \exp\left(-\frac{j^2}{2\sigma_k^2 n}\right)\left(1 + O\left(\frac{j^3}{n^2}\right) + O\left(\frac{1}{n}\right)\right). \tag{11}$$

This equation and Eq. (10) for $j = 0$ proves the second part of Eq. (8).

The function $x \mapsto e^{-x^2/2\sigma^2}$ appears in this formula, applied to $x = \frac{j}{\sqrt{n}}$, from which we will eventually obtain the asymptotic Gaussian shape. This hints that everything meaningful happens at scale $\sqrt{n}$ around $\lfloor v_k n \rfloor$. We now want to consider the sum where Eq. (11) is useful, that is, on a range where it contains every window of scale $\sqrt{n}$ and also where $\frac{j^3}{n^2}$ is not too big, for the Gaussian approximation to hold. For these reasons[3], we take a window of scale $n^{5/9}$ for j (we have $\sqrt{n} \ll n^{5/9} \ll n^{3/2}$). One can verify that the contribution outside of this window is negligible:

$$\sum_{i=\lfloor an \rfloor + 1}^{\lfloor v_k n \rfloor - \lfloor n^{5/9} \rfloor} P(X_n = i) + \sum_{i=\lfloor v_k n \rfloor + \lfloor n^{5/9} \rfloor}^{\lfloor bn \rfloor} P(X_n = i) = o\left(\frac{1}{n}\right). \tag{12}$$

For the first sum, we use that f is increasing on $[a, v_k]$, so that we can bound from above $P(X_n = i)$ by its value computed from the estimation of Eq. (11) with $i = v_k n - n^{5/9}$; this is enough to obtain the result. The second sum is calculated similarly.

3.1.4 Approximation by an integral at scale $\sqrt{n}$ around $v_k n$

At this point we have reduced the range of the sum to $\{\lfloor v_k n \rfloor - \lfloor n^{5/9} \rfloor, \ldots, \lfloor v_k n \rfloor + \lfloor n^{5/9} \rfloor\}$, and we aim at proving the following result: for all $t \in \mathbb{R}$,

$$\sum_{j=-\lfloor n^{5/9} \rfloor}^{\lfloor t\sqrt{n} \rfloor} \left(\frac{j}{\sqrt{n}}\right)^\ell P(X_n = \lfloor v_k n \rfloor + j) \xrightarrow[n\to\infty]{} \frac{E_k \alpha_k\, g(v_k)}{\sqrt{2\pi}} \int_{-\infty}^{t} x^\ell \cdot \exp\left(-\frac{x^2}{2\sigma_k^2}\right) dx. \tag{13}$$

In the working range of this section, we can use both Eq. (10) and Eq. (11). By Taylor's formula, $g(v_k + O(\frac{1}{n})) = g(v_k) + O(\frac{1}{n})$, and therefore, for all $j \in \{-\lfloor n^{5/9} \rfloor, \ldots, \lfloor n^{5/9} \rfloor\}$,

$$P(X_n = \lfloor v_k n \rfloor + j) = \frac{E_k \alpha_k\, g(v_k)}{\sqrt{2\pi n}} \exp\left(-\frac{j^2}{2\sigma_k^2 n}\right)\left(1 + O(n^{-1/3}) + O(\kappa_n)\right), \tag{14}$$

for some positive sequence $(\kappa_n)_{n\geq 1}$ that tends to 0 as n tends to infinity, which comes from Eq. (10): it is the maximum of the error term for $j \in I_n$.

Let h_ℓ be the function defined on $\mathbb{R}$ by $h_\ell(x) = x^\ell \cdot \exp\left(-\frac{x^2}{2\sigma_k^2}\right)$ and let $(\omega_n)_{n\geq 1}$ be a sequence of positive reals[4] with $\omega_n \to +\infty$, $\omega_n \cdot \kappa_n \to 0$ and $\omega_n \cdot n^{-1/9} \to 0$ as $n \to \infty$. Using this properties, one can obtain from Eq. (14) that for any fixed real t,

$$\sum_{j=-\lfloor \omega_n \sqrt{n} \rfloor}^{\lfloor t\sqrt{n} \rfloor} \left(\frac{j}{\sqrt{n}}\right)^\ell P(X_n = \lfloor v_k n \rfloor + j) = \frac{E_k \alpha_k\, g(v_k)}{\sqrt{2\pi n}} \sum_{j=-\lfloor \omega_n \sqrt{n} \rfloor}^{\lfloor t\sqrt{n} \rfloor} h_\ell\left(\frac{j}{\sqrt{n}}\right) + o(1).$$

---

[3] There are other technical reasons for which $n^{\frac{5}{9}}$ is a better choice than others n^λ with $\frac{1}{2} < \lambda < \frac{2}{3}$, but these are the main ones.

[4] For instance $\omega_n = \min\{\log n, -\log \kappa_n\}$ for n large enough.

Here we recognize a Riemann sum of step $\frac{1}{\sqrt{n}}$. Using that h'_ℓ is bounded on $\mathbb{R}$, we can therefore prove that it can be approximated by an integral (see Figure 4) as follows:

$$\frac{1}{\sqrt{n}} \sum_{j=-\lfloor \omega_n \sqrt{n} \rfloor}^{\lfloor t\sqrt{n} \rfloor} h_\ell\left(\frac{j}{\sqrt{n}}\right) = \int_{-\omega_n}^{t} h_\ell(x)dx + O\left(\frac{\omega_n}{\sqrt{n}}\right). \tag{15}$$

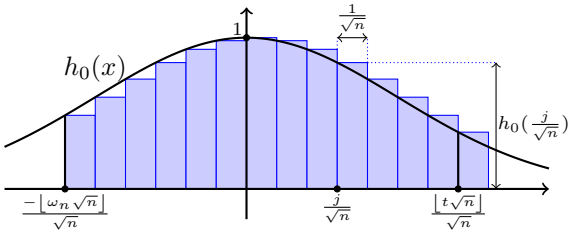

Figure 4 The sum is equal to the area in blue. It is well approximated, when n grows, by the surface below the curve between $-\omega_n$ and t, since the rectangles' width is smaller and smaller. For our function h_ℓ, the error term of this approximation is in $O(\frac{\omega_n}{\sqrt{n}})$, including the last rectangle.

It remains to estimate the sum for j in $\{-\lfloor n^{5/9} \rfloor, \ldots, \lfloor t\sqrt{n} \rfloor\}$. We use integral bounds, which are similar to Riemann sums, to obtain that there exists a constant $D > 0$ such that

$$\sum_{j=-\lfloor n^{5/9} \rfloor}^{-\lfloor \omega_n \sqrt{n} \rfloor} \left(\frac{j}{\sqrt{n}}\right)^\ell P(X_n = \lfloor v_k \rfloor\ n + j) \leq D \int_{-\infty}^{-\omega_n} |h_\ell(x)|dx.$$

Since $-\omega_n \to -\infty$, this part is asymptotically negligible, completing the proof of Eq. (13).

Hence, using Eq. (9) and Eq. (12) we obtain the proof that Eq. (6) holds. The same techniques can also be applied in order to prove Eq. (7).

3.1.5 Another proof of Lebensztayn's theorem [15]

Observe that

$$1 = \sum_{i=1}^{n} P(X_n = i) = \sum_{i=1}^{\lfloor v_k n \rfloor} P(X_n = i) + \sum_{\lfloor v_k n \rfloor}^{n} P(X_n = i) - \underbrace{P(X_n = \lfloor v_k n \rfloor)}_{O(n^{-1/2}) \text{ by Eq. (8)}}.$$

Hence, from what we have just proven, by taking $\ell = 0$ and $t = 0$ in Eq. (6) and Eq. (7), we obtain that

$$\sum_{i=1}^{n} P(X_n = i) \quad \xrightarrow[n\to\infty]{} \quad \frac{E_k \alpha_k\, g(v_k)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h_0(x)dx.$$

But the left quantity is equal to 1, and the right part can be computed and is equal to $E_k \alpha_k\, g(v_k)\sigma_k$. Hence, $E_k \alpha_k\, g(v_k)\sigma_k = 1$, and after basic simplifications we obtain the following expression for E_k, which is much simpler than Eq. (2):

$$E_k = \frac{1}{\alpha_k\, g(v_k)\sigma_k} = k + \frac{k-1}{v_k}.$$

Note that we only needed to know that E_k exists to obtain the formula above, yielding another proof of Lebensztayn's theorem that does not use Korshunov's complicated expression for E_k.

4 Algorithms for random sampling

In this section we describe random generation algorithms for deterministic and complete accessible automata, which are all variations on the same rejection algorithm[5]. As explained in the introduction, our main algorithm `RandomSampler(`m`,`k`)` (presented in Fig. 5) generates at random structures in $\mathcal{T}_{n,k}$ for $n = \left\lfloor \frac{m}{v_k} \right\rfloor$ and extracts their accessible automata. The algorithm rejects until the accessible automaton is of size m. Recall that the accessible automaton of a structure is obtained by restricting the structure to its set $\{s_1 < \ldots < s_j\}$ of accessible states and by relabeling the state s_i as i for all $i \in [j]$.

```
1 v_k ← 1 + (1/k) W_0(−k e^{−k})
2 n ← ⌊m / v_k⌋
3 repeat
4 |   for (p, a) ∈ [n] × {a_1, ..., a_k} do
5 |   |   δ_T(p, a) ←Uniform([n])
6 |   A ← accessible automaton of T
7 until A has m states
8 return A
```

`RandomSampler(`m`,`k`)`

```
1 v ← e^{−1}
2 repeat
3 |   n ← ⌊m / v⌋
4 |   for (p, a) ∈ [n] × {a_1, ..., a_k} do
5 |   |   δ_T(p, a) ←Uniform([n])
6 |   A ←accessible automaton of T
7 |   v ← max(|A| / n, e^{−1})
8 until A has m states
9 return A
```

`AdaptiveRandomSampler(`m`,`k`)`

Figure 5 Random samplers for deterministic and complete accessible automata. $\delta_T(p, q)$ is the target of the transition starting at p and labeled by a in T.

Let us now analyze the expected complexity of `RandomSampler(`m`,`k`)`. The computation of v_k can be achieved [5] using a truncation of the formula $W_0(x) = \sum_{i=1}^{\infty} \frac{(-i)^{i-1}}{i!} x^i$, which converges exponentially fast for $|x| < \frac{1}{e}$ and hence in particular for $x = -ke^{-k}$. Hence if we keep the m first terms only, we have enough precision for the rest of the algorithm. A breadth-first search algorithm is used to compute the accessible part in time $\Theta(m)$, since k is fixed. The relabeling necessary to obtain the accessible automaton can also be computed in $\Theta(m)$.

Hence the expected complexity is a linear function of m times the expected number of iterations. The expected number of iterations is the expected value of the number of tries to obtain the event $X_n = m$ which is equal to $\frac{1}{P(X_n=m)}$. Using Eq. (10) and Eq. (11), we have that $P(X_n = m)$ is equivalent to $\frac{E_k \alpha_k g(v_k)}{\sqrt{2\pi n}}$. The expected number of iterations is therefore in $\Theta(\sqrt{n})$. This leads to the expected complexity stated in the theorem below.

▶ **Theorem 4.** *For any fixed integer $k \geq 2$, the expected complexity of* `RandomSampler(`m`,`k`)` *is in $\Theta(m^{3/2})$.*

4.1 Approximate Sampling

If we relax the condition of Line 7 in `RandomSampler(`m`,`k`)` to keep A when its number of states is in $[m-\varepsilon\sqrt{m}, m+\varepsilon\sqrt{m}]$, we obtain algorithm `ApproxRandomSampler(`m`,`k`,`ε`)`. Notice that `ApproxRandomSampler(`m`,`k`,`ε`)` outputs automata of different sizes, in $[m - \varepsilon\sqrt{m}, m +$

[5] Some prefer the name "pseudo-algorithm" since it may never halt; but this event has probability 0.

$\varepsilon\sqrt{m}]$. However, if we only consider automata of fixed size $m' \in [m - \varepsilon\sqrt{m}, m + \varepsilon\sqrt{m}]$, `ApproxRandomSampler(`$m$`,`$k$`,`$\varepsilon$`)` is a uniform generator for accessible automata of size m'.

As for `RandomSampler(`m`,`k`)`, the expected complexity is a linear function of m times the average number of iterations, which depends on $P(m - \varepsilon\sqrt{n} \leq X_n \leq m + \varepsilon\sqrt{n})$. For any $\varepsilon > 0$, the average number of iterations of `ApproximateSampling(`m`)` is

$$P(m - \varepsilon\sqrt{n} \leq X_n \leq m + \varepsilon\sqrt{n}) \sim \left(\frac{1}{\sigma_k\sqrt{2\pi}} \int_{-\varepsilon}^{\varepsilon} \exp\left(-\frac{x^2}{2\sigma_k^2}\right) dx\right) = \Theta(1). \tag{16}$$

More precisely, Equation (16) shows that the average number of iterations is in $\Theta(\varepsilon^{-1})$.

▶ **Theorem 5.** *For any fixed integer $k \geq 2$ and real $\varepsilon > 0$, the expected complexity of* `ApproxRandomSampler(`m`,`k`,`ε`)` *is in $\Theta(m)$.*

Remark that the usual approximated range interval is $[m(1-\varepsilon), m(1+\varepsilon)]$: the algorithm we propose is much more precise.

4.2 Avoiding the computation of the constant v_k.

It is possible to avoid the explicit computation of v_k, using the self-adaptive algorithm `AdaptiveRandomSampler(`m`,`k`)` presented in Fig. 5. The idea is that if n is large enough, generating a structure of size n and computing the size n' of its accessible automaton yields an estimation of v_k by $\frac{n'}{n}$, which is likely to be precise. The approximated sampler `AdaptiveApproxRandomSampler(`m`,`k`,`ε`)` is defined similarly.

Though needing more iterations than the first versions[6], these two adaptive algorithms have the same expected complexity as stated in the following theorem.

▶ **Theorem 6.** *For any fixed integer $k \geq 2$ and real $\varepsilon > 0$, the expected complexities of* `AdaptiveRandomSampler(`m`,`k`)` *and* `AdaptiveApproxRandomSampler(`m`,`k`,`ε`)` *are respectively $\Theta(m^{3/2})$ and $\Theta(m)$.*

Note that it could be tempting to replace v_k by a fixed approximation. For instance, one could take 0.8 for v_2. It is easy to show that doing so results in an asymptotically exponential number of rejections. For instance, when taking $v_2 = 0.8$, if we use $f(0.8)^n$ to estimate the proportion of additional rejects, we see that for automata of size 100,000 we do approximatively 7 times as many rejections, but moving to automata of size 1,000,000 this number jumps to approximatively 142,000,000 times as many. This underlines the importance of mathematical analysis not only to study algorithms but also to devise them.

4.3 Sampling random accessible minimal automata

Recently, in [11] it was shown that the probability for an automaton of $\mathcal{A}_{n,k}$ to be minimal tends toward some constant $\lambda_k > 0$ as n tends to infinity.

So if we replace the condition of Line 8 of `RandomSampler(`m`,`k`)` by "$\mathcal{A}$ has m states and is minimal", we obtain a random sampler for accessible and minimal automata. This is strictly equivalent to first use `RandomSampler(`m`,`k`)` and then apply a rejection algorithm to keep minimal automata only; the induced distribution on minimal automata is therefore the uniform distribution.

[6] Simulations for a two-letter alphabet seem to indicate that at most twice as much iterations are required, on average.

If we use Moore's algorithm (which has an average complexity in $O(n \log\log n)$, see Section 5) to test for minimality, we obtain an expected complexity in $\Theta(m^{3/2}+m\log\log m) = \Theta(m^{3/2})$. For the `ApproxRandomSampler(`$m$`,`$k$`,`$\varepsilon$`)`, we obtain an approximated sampler for minimal accessible automata with an expected complexity in $\Theta(m\log\log m)$.

5 Application to analysis of algorithm

The main perspective of this work is to help analyze the average complexities of algorithms that deal with accessible automata using average complexities of this same algorithms on structures.

A common technique for such studies is to isolate sets of inputs with non-typical behaviors, and then prove that the contribution of such sets to the average complexity is negligible, because an input belongs to such a set with a small probability. In our context, it is usually much easier to prove such properties for structures rather than for automata, since they are simpler combinatorial objects. In this section, we briefly describe a general scheme that can be used in these situations: under some general conditions, properties that sufficiently rarely hold for structures still rarely hold for automata.

The idea is the following: let $\mathcal{P}$ be a property of automata (accessible or not) such that if the property holds for the accessible automaton of a structure it also holds for the structure itself. A property such as "being accessible" obviously does not satisfy this requirement but a property such as "having a sink state" does. In the following we explain and illustrate why, if one can afford a $\sqrt{m}$ multiplier, the negligibility of such a property can be transfered from structures to automata.

Let $p_{\mathcal{A}}(m)$ and $p_{\mathcal{T}}(n)$ denote the probabilities that $\mathcal{P}$ holds for a size-m automaton and for a size-n structure, respectively. Then, if A_T denotes the accessible automaton of the structure T,

$$p_{\mathcal{A}}(m) = \frac{|\{T \in \mathcal{T}_n : |A_T| = m \text{ and } A_T \text{ satisfies } \mathcal{P}\}|}{|\{T \in \mathcal{T}_n : |A_T| = m\}|} \leq \frac{|\{T \in \mathcal{T}_n : T \text{ satisfies } \mathcal{P}\}|}{|\{T \in \mathcal{T}_n : |A_T| = m\}|}.$$

The last quantity is equal to $\frac{p_{\mathcal{T}}(n)}{P(X_n=m)}$ by multiplying and dividing the quantity by $|\mathcal{T}_n|$. By taking $n = \left\lfloor \frac{m}{v_k} \right\rfloor$, and using Eq. (8), we obtain that for any such property $\mathcal{P}$,

$$p_{\mathcal{A}}(m) \leq p_{\mathcal{T}}\left(\left\lfloor \frac{m}{v_k} \right\rfloor\right) \cdot O(\sqrt{m}). \tag{17}$$

We now give two examples to illustrate how Eq. (17) can be used. First, we prove that the probability that an automaton has a sink state is asymptotically negligible:

▶ **Lemma 7.** *For the uniform distribution, the probability that an automaton with m states on an alphabet with $k \geq 2$ letters has a sink state is in $O(n^{3/2-k})$.*

Proof. As remarked previously Eq. (17) holds for this property. The probability that a structure with n states has at least one sink state is at most n^{1-k}, since every given state is a sink state with probability n^{-k}. This conclude the proof by taking $n = \lfloor m/v_k \rfloor$. ◀

The same technique can be used to prove a deeper result on Moore's minimization algorithm. David [6] proved that for the uniform distribution on structures with n states on an alphabet with at least two letters, the average complexity of Moore's algorithm is $O(n\log\log n)$.

His result can almost readily be adapted to the uniform distribution on accessible automata, using the method described above. First notice that when applied to a structure T, the complexity of Moore's algorithm is greater than or equal to its complexity for the accessible automaton of T. David's proof relies on showing that the probability that a structure needs more than $\Theta(n \log\log n)$ instructions is small enough to have a negligible contribution to the average complexity. With some care, one can show that his error terms can be handled with the $O(\sqrt{m})$ multiplier of Eq. (17), giving the result.

Acknowledgements: The second author is supported by the ANR project MAGNUM: ANR-2010-BLAN-0204.

References

1. Marco Almeida, Nelma Moreira, and Rogério Reis. Enumeration and generation with a string automata representation. *Theor. Comput. Sci.*, 387:93–102, 2007.
2. Frédérique Bassino, Julien David, and Cyril Nicaud. Enumeration and random generation of possibly incomplete deterministic automata. *Pure Math. and Appl.*, 19:1–16, 2008.
3. Frédérique Bassino and Cyril Nicaud. Enumeration and Random Generation of Accessible Automata. *Theor. Comput. Sci.*, 381:86–104, 2007.
4. Jean-Marc Champarnaud and Thomas Paranthoën. Random generation of DFAs. *Theor. Comput. Sci.*, 330:221–235, 2005.
5. Robert M. Corless, Gaston H. Gonnet, David E. G. Hare, David J. Jeffrey, and Donald E. Knuth. On the Lambert W function. In *Adv. Comp. Math.*, pages 329–359, 1996.
6. Julien David. The Average Complexity of Moore's State Minimization Algorithm is $O(n \log\log n)$. In *Proc. of MFCS '10*, LNCS, pages 318–329, 2010.
7. Philippe Duchon, Philippe Flajolet, Guy Louchard, and Gilles Schaeffer. Boltzmann samplers for the random generation of combinatorial structures. *Combinatorics, Probability & Computing*, 13(4-5):577–625, 2004.
8. William Feller. *An Introduction to Probability Theory and Its Applications I.* Wiley, 1968.
9. Philippe Flajolet and Andrew M. Odlyzko. Random mapping statistics. In *Adv. Crypt.*, pages 329–354, 1990.
10. Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics.* Cambridge University Press, 2009.
11. Frédérique Bassino and Julien David and Andrea Sportiello. Asymptotic Enumeration of Minimal Automata. In *Proc. of STACS '12*, LIPIcs, 2012.
12. Irving J. Good. An asymptotic formula for the differences of the powers at zero. *Ann. Math. Statist.*, 32:249–256, 1961.
13. John E. Hopcroft and Jeffrey Ullman. *Introduction to automata theory, languages and computation.* Addison-Wesley, 1980.
14. Aleksey Korshunov. Enumeration of finite automata. *Problemy Kibernetiki*, 34:5–82, 1978.
15. Elcio Lebensztayn. On the asymptotic enumeration of accessible automata. *Discr. Math. & Theor. Comp. Sc.*, 12(3):75–80, 2010.
16. Valery A. Liskovets. The number of initially connected automata. *Cybernetics*, 4:259–262, 1969. english translation of Kibernetika (3) 1969, 16-19.
17. Cyril Nicaud. Comportement en moyenne des automates finis et des langages rationnels. *PhD Thesis, Université Paris VII*, 2000.
18. Albert Nijenhuis and Herbert S. Wilf. *Combinatorial Algorithms.* Academic Press, 1978.