Pointed *versus* Singular Boltzmann Samplers: a Comparative Analysis*

Olivier Bodini†, Antoine Genitrini‡ and Nicolas Rolin†.
{Olivier.Bodini;Nicolas.Rolin}@lipn.univ-paris13.fr and Antoine.Genitrini@lip6.fr.

November 16, 2015

Since the last two decades huge systems (such as giant graphs, big data structures, ...) have played a central rôle in computer science, and with the technology improvements, those large objects are now massively used in practice. In order to handle them we need to analyse some typical properties of models of large objects. One way to study typical behaviours consists in generating random objects to get some experimental results on their properties. A new technique has been introduced ten years ago: the Boltzmann sampling. It has been presented by Duchon *et al*, and is based on automatic interpretation in terms of samplers of the specification of the combinatorial objects under study.

One of the core problem in Boltzmann sampling lies in the distribution of the object sizes, and the choice of some parameters in order to get the more appropriate size distribution. From this choice depends the efficiency of the sampling. Moreover some additional ideas allows to improve the efficiency, one of them is based on some anticipated rejections, the other one on the combinatorial differentiation of the specification. Anticipated rejection consists during the recursive building of a random object to kill the process as soon as we are sure to exceed the maximum target size, rather than waiting until the natural end of the process. In the original paper, while both approaches have been presented, and used on the same kind of structures, the methods are not compared.

We propose in this paper a detailed comparison of both approaches, in order to understand precisely which method is the more efficient.

1 Introduction

Since the last two decades huge systems (such as giant graphs, big data structures,...) have played a central rôle in computer science, and with the technology improvements, those large objects are now massively used in practice. In order to handle them we need to analyse some typical properties of models of large objects. One way to study typical behaviours consists in generating random objects to get some experimental results on their properties. We emphasize two major facts that exhibit the fundamental rôle of random sampling. First, in the context of testing, especially in software testing, uniform generation of huge graphs enables to detect errors [CD09], particularly to detect unexpected overflows that wouldn't have been given away on smaller instances. Another

*This research was partially supported by the A.N.R. projects *MAGNUM*, ANR 2010-BLAN-0204 and by MetACOnc (2015–2018).

†Laboratoire d'Informatique de Paris-Nord, CNRS UMR 7030 - Institut Galilée - Université Paris-Nord, 99, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France.

‡Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris. Antoine.Genitrini@lip6.fr

application consists in constructing large instances that allow to understand how big typical objects do behave: let us mention few papers based on this approach, e.g. in group theory [BNW08], in the context of formal languages [BN07] or in statistical physics [BFP10].

The first algorithms that uniformly sample combinatorial structures are *ad hoc* methods which are dedicated to a specific class of objects. Let us, for example, mention Rémy's algorithm [Rém85], recently improved in [BBJ13, BBJ14] that builds uniformly binary trees. Another usual way to sample uniformly at random combinatorial structures is based on the recursive method developed in the book [NW78] and then revisited in the context of *analytic combinatorics* [FZC94]. This approach relies on a systematic process to generate combinatorial objects. Most of the time it is a bit less efficient than *ad hoc* methods, however it has the nice property to adapt to many families of objects that are usually qualified as decomposable (or specifiable) objects.

Since 2004, a new technique has been developed: the Boltzmann sampling. It was presented by Duchon *et al*, in [DFLS04], and is based on the decomposition of the combinatorial objects, through their specification. Then the approach has been extensively explored in order to fit to much more constructions, see the papers [FFP07, Duc11, BRS12, BP10] to have an idea of some new developments. Boltzmann sampling plays a central rôle in the context of random generation, because of its simplicity, its genericity and its efficiency. All these fundamental properties are however combined with a constraint: in order to be efficient, the generator builds objects with an approximate size instead of the size fixed by the user. In many cases, this constraint is not a problem at all: e.g., when one is interested in testing a huge system, it does no matter if the size of the simulations is around one million instead of being exactly one million.

The Boltzmann samplers deal with a parameter to generate objects. The tuning of this parameter allows for example to build objects whose expected size is the fixed size, one wants to reach. Once the parameter is chosen, a tricky point is the evaluation of generating series in specific values. Some studies are presenting improvements, based on the quadratic iterative Newton method [PSS12], or a curve approach of the problem [BLR15] in order to overcome these difficulties.

Core problems in Boltzmann sampling lie in the distribution of the object sizes, and the choice of the parameter in order to get the best size distribution. An interesting distribution would have its mass concentrated around its mean and the selected parameter would yield a good probability that a randomly generated object falls in a size range near the mean. But many combinatorial structures do not exhibit such nice distributions, and have rather their distribution mass around extreme values, like the distributions usually called with *big tails*. In such cases, even a smart choice of the parameter does not give an interesting complexity. A first step to deal with these problems is based on some anticipated rejection. Thus during the recursive building of a random object we stop the process as soon as we are sure to exceed the maximum target size, rather than waiting until the natural end of the process. The threshold is easy to determine by computing dynamically the size of the object under construction.

In [DFLS04], two different methods are proposed in order to overcome the problems induced by big tail distributions. (1) A method based on some pointing operation: the idea is to modify the size distribution in order to get some more interesting distribution. (2) A method based on singular sampling: some anticipated rejection is necessary and the parameter is taken to be the dominant singularity associated to the combinatorial objects. This last approach allows to avoid the choice of the parameter and the heavy evaluation of the generating series. While both approaches have been presented, in the original paper [DFLS04], on the same kind of structures, only the complexities for the specific case of the singular exponent of simple family of trees have been computed. Furthermore no quantitative comparison is given there. It isn't clear in the literature which method is the most efficient. We propose in this paper a detailed comparison of both approaches in more general cases (not only for the singular exponent of simple family of trees).

The paper is organized as follows: Section 2 gives the contextual definitions and notions about Boltzmann sampling. Sections 3 and 4 contain the time complexity analysis of the different methods of sampling. Finally, Section 5 is devoted to the quantitative comparison, in terms of time complexity, of the two methods of generation, both in the general structures first and then in the case of aperiodic and strongly connected grammars. The comparison starts from a theoretical

point of view and we conclude the section by experiments to check that the asymptotic theoretical values are already observed for objects that can be sampled in practice.

2 Context of Boltzmann samplers

In this section, we recall the classical definitions in the context of combinatorial structures and Boltzmann samplers. In the paper we will deal with unlabelled combinatorial classes, thus classes associated to ordinary generating functions. However, the adaptation to the labelled structures is straightforward.

2.1 Boltzmann distributions

A *combinatorial class* $\mathcal{C}$ is a set of objects, with a size function, denoted by $|\cdot| : \mathcal{C} \to \mathbb{N}$ and such that for every integer n, the subset $\mathcal{C}_n$ of objects of size n. Its cardinality is finite and denoted by C_n. We define the *ordinary generating function* of a combinatorial class $\mathcal{C}$ to be:

$$C(z) = \sum_{\gamma \in \mathcal{C}} z^{|\gamma|} = \sum_{n \geq 0} C_n z^n.$$

Let ρ be the *dominant singularity* of the generating function $C(z)$.

Definition 2.1. *Let $\mathcal{C}$ be an unlabelled combinatorial class, whose ordinary generating function is $C(z)$. Its associated Boltzmann model, depending on a real parameter $0 < x < \rho$ or $x = \rho$ in some cases, is the distribution over the objects of $\mathcal{C}$, such that the probability of an object γ is*

$$\mathbb{P}_x(\gamma) = \frac{x^{|\gamma|}}{C(x)}, \text{ where } C(x) \text{ is the generating function of } \mathcal{C} \text{ evaluated in } x.$$

Obviously note that the distribution is uniform among all objects of the same size. Hence, the probability of getting an object of size n, denoted by $\mathbb{P}_x(N = n)$ and the expected size, denoted by $\mathbb{E}_x(N)$, of the object, for the parameter x, satisfy:

$$\mathbb{P}_x(N = n) = \frac{C_n x^n}{C(x)}, \qquad \text{and} \qquad \mathbb{E}_x(N) = \frac{x\ C'(x)}{C(x)}.$$

2.2 Boltzmann samplers

A *Boltzmann sampler* is a random generator of objects according to their Boltzmann distribution. Since the size of a generated object is random, in order to sample an object of a given size n, trials are repeated until getting an object of the right size. To obtain a generation [DFLS04] that is more efficient, usually, objects with an approximation on the size are accepted. For example, an object whose size belongs to the range $n(1-\varepsilon)$ and $n(1+\varepsilon)$ is accepted, where ε is the tolerance on the size.

We will restrict the study to combinatorial classes whose generating functions are $\Delta - singular$. Remark that $\Delta - singular$ functions are associated to combinatorial decomposable objects, as argued in [FZC94].

Definition 2.2. *A function $C(z)$ analytic at 0 and with a finite radius of analyticity $\rho > 0$ is said* Δ-singular *if it satisfies the two following conditions.*

(i) The function admits ρ as its only singularity on $|z| = \rho$ and it is continuable in a domain

$$\Delta(r, \theta) = \{z \mid z \neq \rho, |z| < r, arg(z - \rho) \notin (-\theta, \theta)\},$$

for some $r > \rho$ and some θ satisfying $0 < \theta < \frac{\pi}{2}$. The set $\Delta(r, \theta)$ is called a Δ-domain.

(ii) For z tending to ρ in the Δ-domain, $C(z)$ satisfies a singular expansion of the form

$$C(z) \underset{z\to\rho}{=} P(z) + c_0(1 - z/\rho)^{-\alpha}) + o((1 - z/\rho)^{-\alpha}), \alpha \in \mathbb{R}\backslash\{0, -1, -2, ...\},$$

where $P(z)$ is a polynomial; c_0 a constant and $-\alpha$ is called the singular exponent *of $C(z)$.*

Using the Flajolet-Odlysko transfer theorems, (cf. [FS09, chapter VI] for details),

$$C_n := [z^n]C(z) \underset{n\to\infty}{\sim} \frac{c_0}{\Gamma(\alpha)}\rho^{-n}n^{\alpha-1}. \tag{1}$$

Note that for $-\alpha > 0$, the size distributions of objects have a heavy tail: such distributions are usually called *peaked distributions*. While the distributions induced by $-\alpha < 0$ are called *flat distributions*. In Figure 1 some distribution are represented, according to the value of the parameter. The blue dashed lines correspond to peaked distributions (for this example $-\alpha = 1/2$) while the red lines are flat distributions (here $-\alpha = -1/2$). Note that in the case of peaked distribution, a large probability is dedicated to small objects. In the case of a flat distribution, the mass of the probability is spread over a larger set of objects.

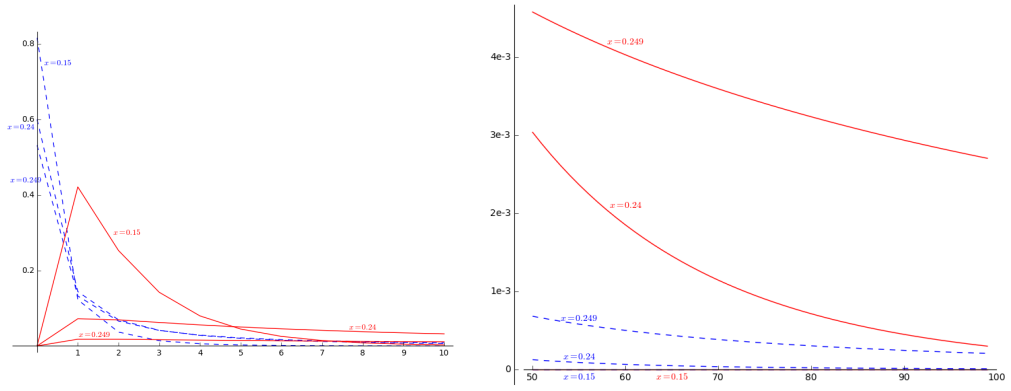

Figure 1: *Boltzmann distributions with different parameters x (a peaked distribution in blue dashed lines; a flat distribution in red lines)*

In order to generate objects according to their Boltzmann distribution, Duchon *et al* [DFLS04] have presented three samplers. The first one, that we will call *the basic Boltzmann sampler*, is based on the best parameter x_n as possible, in order to maximize the chances to get an object of the appropriate size. The second sampler, called *pointed Boltzmann sampler* deals with flat distributions of objects. In order to turn a peak distribution (induced by trees for example) to a flat distribution, we can mark some atoms of the structures, then generate such marked structures and finally avoid the marked nodes. Obviously, this method is still uniform among objects of the same size, even if the global distribution is now different from the original Boltzmann one. Finally, the third generator, called *singular Boltzmann sampler*, is based on peaked distributions of structures. In this case, we can take the dominant singularity as parameter.

2.3 Complexity of Boltzmann samplers

For a Boltzmann sampler, the complexity is usually measured by the cumulated size of the objects that are rejected because their sizes are not in the good range. This measure is relevant, because in Boltzmann sampling, the time, the space and the number of random bits that are used are usually growing linearly with the cumulated sizes of the sampled objects (cf. [DFLS04]).

Considering such a complexity measure, we can easily improve the sampler by avoiding the construction the objects larger that the maximum size fixed. Thus by rejecting an object as soon

as one of its part exceeded the maximum size. This procedure is called *anticipated rejection*. Let us first present a result proved in [DFLS04], but that is central for our study. It gives the complexities of the approximate Boltzmann samplings without or with anticipated rejection.

Lemma 2.1. *[DFLS04] Let $C(z)$ be the generating function of a class $\mathcal{C}$, and let $C^{<n_1}$, $C^{>n_2}$ and $C^{[n_1,n_2]}$ be the generating functions for the subclasses of objects with size, respectively, smaller than n_1, greater than n_2, and between n_1 and n_2. The cumulative size, T_n, of the rejected objects (whose size is either smaller than n_1 or larger than n_2) with parameter x satisfies the following first and second moments.*

$$\mathbb{E}(T_n) = \frac{xC'^{<n_1}(x) + xC'^{>n_2}(x)}{C^{[n_1,n_2]}(x)} \text{ and}$$

$$\mathbb{E}(T_n^2) = \frac{x^2C''^{<n_1}(x) + x^2C''^{>n_2}(x)}{C^{[n_1,n_2]}(x)} + 2\mathbb{E}(T_n)^2 + \mathbb{E}(T_n).$$

Furthermore, by improving the sampler with some anticipated rejection, we get:

$$\mathbb{E}(T_n) = \frac{xC'^{<n_1}(x) + n_2C^{>n_2}(x)}{C^{[n_1,n_2]}(x)} \text{ and}$$

$$\mathbb{E}(T_n^2) = \frac{x^2C''^{<n_1}(x) + n_2(n_2-1)C^{>n_2}(x)}{C^{[n_1,n_2]}(x)} + 2\mathbb{E}(T_n)^2 + \mathbb{E}(T_n).$$

Note that in the average complexity computations $\mathbb{E}(T_n)$, in the second case, with some anticipated rejection, no object of size larger than n_2 is generated. In fact while the construction of an object, once its partial size is larger than n_2, it is directly rejected. Thus the term $xC'^{>n_2}(x)$ computed in the average complexity of the basic sampler (without anticipated rejection) is replaced by $n_2C'^{>n_2}(x)$ for the sampler with anticipated rejection.

Proof. The following proof is given in [DFLS04].
The probability generating function related to the approximate Boltzmann sampler with rejection targeted at $[n_1, n_2]$ is

$$F(u,x) = \sum_{k>0} \mathbb{P}(T_n = k)u^k.$$

From the decomposition of a call to the sampler into a sequence of unsuccessful trials, (each one contributing to T_n) and then followed by a final successful trial (that does not contribute to T_n), we get:

$$F(u,x) = \left(1 - \frac{1}{C(x)}\left(C^{<n_1}(ux) + C^{>n_2}(x)u^{n_2}\right)\right)^{-1} \frac{C^{[n_1,n_2]}(x)}{C(x)}.$$

Then the expectation of the cost is given by $\mathbb{E}(T_n) = \frac{u\partial}{\partial u}F(u,x)_{|u=1}$, and the second moment by $\mathbb{E}(T_n^2) = \left(\frac{u\partial}{\partial u} + \frac{u^2\partial^2}{\partial u^2}\right)F(u,x)_{|u=1}$. By observing that $C(x) - C^{<n_1}(x) - C^{>n_2}(x) = C^{[n_1,n_2]}(x)$, the results follow immediately. □

This lemma allows to compute the expected cost of an approximate Boltzmann sampler and its second moment with anticipated rejection as long as the quantities $xC'^{<n_1}(x)$, $n_2C^{>n_2}(x)$, $C^{[n_1,n_2]}(x)$ and $x^2C''^{<n_1}(x)$ can be calculated.

Let us end this section by recalling some notations associated to each sampler, that have been introduced in [DFLS04]. Let $\mu C(x,n,\varepsilon)$ represents the Boltzmann sampler for the combinatorial class $\mathcal{C}$, without anticipated rejection, with parameter x, size n and size-tolerance ε. And let $\nu C(x,n,\varepsilon)$ be the analogous Boltzmann sampler with anticipated rejection.

We turn now to the computations of the sampler complexities. In order to classify our study, we partition the combinatorial structures according to the value of their singular exponent.

3 Objects presenting a flat distribution

In this section, we are considering combinatorial structures, whose singular exponent $-\alpha$ is negative. Such objects present a size distribution that is flat. For example, surjections or regular languages belong to the case of flat distributions.

To sample an object of size n, the canonical way of choosing an appropriate parameter x_n is such that the expected size of the generated objects equals or is close to n. Thus let x_n be the solution of $\mathbb{E}_{x_n}(N) = x_n$. In the case when $-\alpha < 0$, we get $x_n \underset{n\to\infty}{\sim} \rho(1 - \frac{\alpha}{n})$, hence we choose x_n to be equal to $\rho(1 - \frac{\alpha}{n})$.

3.1 Basic Boltzmann sampler without anticipated rejection

Theorem 3.1. *Let $\mathcal{C}$ be a combinatorial class whose generating function is Δ-singular with an exponent $-\alpha < 0$. Then the cumulated size T_n of the objects generated and rejected by an approximate Boltzmann sampler without anticipated rejection, $\mu C(x_n, n, \varepsilon)$, satisfies:*

$$\mathbb{E}(T_n) \underset{n\to\infty}{\sim} n\kappa(\varepsilon, \alpha), \quad \text{and} \quad \mathbb{E}(T_n^2) \underset{n\to\infty}{\sim} 2\mathbb{E}(T_n)^2 + n^2\bar{\kappa}(\varepsilon, \alpha),$$

where

$$\kappa(\varepsilon, \alpha) = \frac{\int_{w=0}^{1-\varepsilon} w^\alpha e^{-\alpha w} dw + \int_{w=1+\varepsilon}^{\infty} w^\alpha e^{-\alpha w} dw}{\int_{w=1-\varepsilon}^{1+\varepsilon} w^{\alpha-1} e^{-\alpha w} dw},$$

and

$$\bar{\kappa}(\varepsilon, \alpha) = \frac{\int_{w=0}^{1-\varepsilon} w^{\alpha+1} e^{-\alpha w} dw + \int_{w=1+\varepsilon}^{\infty} w^{\alpha+1} e^{-\alpha w} dw}{\int_{w=1-\varepsilon}^{1+\varepsilon} w^{\alpha-1} e^{-\alpha w} dw}.$$

Proof. Let us first compute estimations for the next three quantities:
$x_n C'^{<n(1-\varepsilon)}(x_n)$, $x_n C'^{>n(1+\varepsilon)}(x_n)$ and $C^{[n(1-\varepsilon),n(1+\varepsilon)]}(x_n)$.

We use the approximation (1) together with the value for x_n defined in the beginning of Section 3 to calculate the quantities :

$$x_n C'^{<n(1-\varepsilon)}(x_n) = \frac{c_0}{\Gamma(\alpha)} \sum_{k=0}^{\lceil n(1-\varepsilon)\rceil} e^{k\ln(1-\alpha/n)} k^\alpha (1 + o_{k\to\infty}(1))$$

$$= \frac{c_0}{\Gamma(\alpha)} \sum_{k=0}^{\lceil n(1-\varepsilon)\rceil} e^{-\frac{\alpha k}{n}} k^\alpha (1 + o_{k\to\infty}(1)).$$

As the sum $\sum_{k=0}^{\lceil n(1-\varepsilon)\rceil} e^{-\frac{\alpha k}{n}} k^\alpha$ tends to infinity when n tends to infinity, we can pull the error term out of the sum:

$$x_n C'^{<n(1-\varepsilon)}(x_n) \underset{n\to\infty}{=} \frac{c_0}{\Gamma(\alpha)} \left(\sum_{k=0}^{\lceil n(1-\varepsilon)\rceil} e^{-\frac{\alpha k}{n}} k^\alpha \right) (1 + o_{n\to\infty}(1)).$$

Then we can use Euler-MacLaurin summation:

$$\frac{c_0}{\Gamma(\alpha)}\left(\sum_{k=0}^{\lceil n(1-\varepsilon)\rceil} e^{-\frac{\alpha k}{n}} k^\alpha\right)(1+o(1)) = \frac{c_0}{\Gamma(\alpha)}\left(\int_{t=1}^{\lceil n(1-\varepsilon)\rceil} t^\alpha e^{-\alpha t} dt\right)(1+o(1)).$$

Hence,

$$x_n C^{\prime<n(1-\varepsilon)}(x_n) \underset{n\to\infty}{\sim} \frac{c_0 n^{\alpha+1}}{\Gamma(\alpha)} \int_{w=0}^{1-\varepsilon} w^\alpha e^{-\alpha w} dw.$$

We can use Euler-MacLaurin summation of the coefficients of $C(z)$ and the same kind of analysis to give us the following estimations:

$$C^{[n(1-\varepsilon),n(1+\varepsilon)]}(x_n) \underset{n\to\infty}{\sim} \frac{c_0 n^{\alpha}}{\Gamma(\alpha)} \int_{w=1-\varepsilon}^{1+\varepsilon} w^{\alpha-1} e^{-\alpha w} dw,$$

$$x_n^2 C^{\prime\prime<n(1-\varepsilon)}(x_n) \underset{n\to\infty}{\sim} \frac{c_0 n^{\alpha+2}}{\Gamma(\alpha)} \int_{w=0}^{1-\varepsilon} w^{\alpha+1} e^{-\alpha w} dw,$$

and

$$x_n^2 C^{\prime\prime>n(1+\varepsilon)}(x_n) \underset{n\to\infty}{\sim} \frac{c_0 n^{\alpha+2}}{\Gamma(\alpha)} \int_{w=1+\varepsilon}^{\infty} w^{\alpha+1} e^{-\alpha w} dw.$$

It remains now to apply the Lemma 2.1 in order to compute the first and second moment of the complexity measure. □

Obviously, once the combinatorial class $\mathcal{C}$ is fixed (and ϵ also), all these integrals can be explicitly computed and then the comparison of the different methods is direct (see an example in Section 5).

3.2 Basic Boltzmann sampler with anticipated rejection

In all our study, we are aware of efficiency. Let us now look at samplers with anticipated rejection. Then we will be able to compare both approaches. Thus, at each step of the building, we compare the actual size of the components under construction to the maximal allowed size. If we have already exceeded (before the end of the building process) the maximal size, we stop the process, reject the partial construction and start a new one.

Theorem 3.2. *Let $\mathcal{C}$ be a combinatorial class whose generating function is Δ-singular with an exponent $-\alpha < 0$. Then the cumulated size T_n of the objects generated and rejected by an approximate Boltzmann sampler with anticipated rejection, $\nu C(x_n, n, \varepsilon)$, satisfies:*

$$\mathbb{E}(T_n) \underset{n\to\infty}{\sim} n\kappa_r(\varepsilon,\alpha), \quad \text{and} \quad \mathbb{E}(T_n^2) \underset{n\to\infty}{\sim} 2\mathbb{E}(T_n)^2 + n^2\bar{\kappa}_r(\varepsilon,\alpha),$$

where

$$\kappa_r(\varepsilon,\alpha) = \frac{\displaystyle\int_{w=0}^{1-\varepsilon} w^\alpha e^{-\alpha w} dw + \int_{w=1+\varepsilon}^{\infty} (1+\varepsilon)\cdot w^{\alpha-1} e^{-\alpha w} dw}{\displaystyle\int_{w=1-\varepsilon}^{1+\varepsilon} w^{\alpha-1} e^{-\alpha w} dw},$$

and

$$\bar{\kappa}_r(\varepsilon,\alpha)=\frac{\int_{w=0}^{1-\varepsilon} w^{\alpha+1}e^{-\alpha w}dw+\int_{w=1+\varepsilon}^{\infty}(1+\varepsilon)\cdot w^{\alpha-1}e^{-\alpha w}dw}{\int_{w=1-\varepsilon}^{1+\varepsilon} w^{\alpha-1}e^{-\alpha w}dw}.$$

The proof is completely analogous to the previous one for the Theorem 3.1.

Proof. Our goal consists in using the Lemma 2.1. Thus we need estimations for $x_nC'^{<n(1-\varepsilon)}(x_n)$, $n(1+\varepsilon)C^{>n(1+\varepsilon)}(x_n)$ and $C^{[n(1-\varepsilon),n(1+\varepsilon)]}(x_n)$. We use an Euler-MacLaurin summation on the coefficients of $C(z)$ to obtain those estimations:

$$x_nC'^{<n(1-\varepsilon)}(x_n)\underset{n\to\infty}{\sim}\frac{c_0n^{\alpha+1}}{\Gamma(\alpha)}\int_{w=0}^{1-\varepsilon} w^{\alpha}e^{-\alpha w}dw,$$

$$n(1+\varepsilon)C^{>n(1+\varepsilon)}(x_n)\underset{n\to\infty}{\sim}\frac{c_0n^{\alpha+1}}{\Gamma(\alpha)}\int_{w=1+\varepsilon}^{\infty} w^{\alpha-1}e^{-\alpha w}dw,$$

$$C^{[n(1-\varepsilon),n(1+\varepsilon)]}(x_n)\underset{n\to\infty}{\sim}\frac{c_0n^{\alpha}}{\Gamma(\alpha)}\int_{w=1-\varepsilon}^{1+\varepsilon} w^{\alpha-1}e^{-\alpha w}dw,$$

and

$$x_n^2C''^{<n(1-\varepsilon)}(x_n)\underset{n\to\infty}{\sim}\frac{c_0n^{\alpha+2}}{\Gamma(\alpha)}\int_{w=1+\varepsilon}^{\infty} w^{\alpha+1}e^{-\alpha w}dw.$$

□

By studying the previous moments for both samplers with or without anticipated rejection the consequences of anticipated rejection is directly readable in the integral formulas.

4 Objects presenting a peaked distribution

In this section, we are considering combinatorial structures, whose singular exponent $-\alpha$ is positive. Such objects present a size distribution that is peaked. Tree structures are examples of this kind of objects with $-\alpha$ a positive rational number, smaller than 1.

4.1 Pointed Boltzmann sampler

A pointed object is a structure with a node that is marked. Thus, a pointing Boltzmann sampler generates a pointed structure, but returns the structure without the mark. Note that the sampler generates objects from a much more large set and then embeds the marked structure in the objects under consideration. It is noticeable that each final object (of a given size) as the same number of associated pointed structures and thus the random distribution is uniform. In fact this operation breaks the symmetry during the generation and turns effectively peaked distributions into flat distributions. Given a combinatorial class $\mathcal{C}$, we define the pointed class (associated to any kind of traversal) is defined as:

$$\mathcal{C}^{\bullet}\underset{n\to\infty}{\sim}\{(\gamma,i)\mid\gamma\in\mathcal{C},i\in\{1,\dots,|\gamma|\}\}.$$

We deduce $|\mathcal{C}_n^{\bullet}|=n|\mathcal{C}_n|$, and the related generating function is $C^{\bullet}(z)=zC'(z)$. Hence, if $\mathcal{C}$ has a singular exponent denoted by $-\alpha$, then $\mathcal{C}^{\bullet}$ has a singular exponent equal to $-\alpha-1$.

Let now suppose $\mathcal{C}$ to be a combinatorial class with singular exponent $-\alpha > 0$. By marking $\lceil \alpha \rceil$ atoms in the structures into consideration we obtain a flat distribution over the multi-marked structures. Thus both Theorems 3.1 and 3.2 can be applied on those multi-marked structures. Once an object is generated, we erase its marks, and then obtain an object of $\mathcal{C}$ uniformly at random among objects of the same size.

Note that the differentiation of some specification could give much more large specification (in its size) however this size growth does not interfere our complexity measure for constructing objects.

4.2 Singular Boltzmann sampler

Another method to deal with the specific case when $0 < -\alpha < 1$ consists in avoiding the computation of the appropriate parameter by choosing the singularity as parameter, since the size distribution has a heavy tail. Note, in this context of singular sampling, the average size of a generated object is infinite. Thus we rely on anticipated rejection to sample an object in a reasonable time.

Theorem 4.1. *Let $\mathcal{C}$ be a combinatorial class whose generating function is Δ-singular with an exponent $0 < -\alpha < 1$. Then the cumulated size T_n of the objects generated and rejected by an approximate Boltzmann sampler with anticipated rejection, $\nu C(\rho, n, \varepsilon)$, satisfies:*

$$\mathbb{E}(T_n) \underset{n\to\infty}{\sim} n\kappa_s(\varepsilon,\alpha), \qquad \text{and} \qquad \mathbb{E}(T_n^2) \underset{n\to\infty}{\sim} 2\mathbb{E}(T_n)^2 + n^2\bar{\kappa}_s(\varepsilon,\alpha),$$

where

$$\kappa_s(\varepsilon,\alpha) = \frac{-\alpha\cdot\left(\frac{(1-\varepsilon)^{\alpha+1}}{\alpha+1} + \frac{(1+\varepsilon)^{\alpha+1})}{-\alpha}\right)}{(1-\varepsilon)^\alpha - (1+\varepsilon)^\alpha},$$

and

$$\bar{\kappa}_s(\varepsilon,\alpha) = \frac{-\alpha\cdot\left(\frac{(1-\varepsilon)^{\alpha+2}}{\alpha+2} + \frac{(1+\varepsilon)^{\alpha+1})}{-\alpha}\right)}{(1-\varepsilon)^\alpha - (1+\varepsilon)^\alpha}.$$

The proof of the result is a consequence of Lemma 2.1, like the for the two previous samplers. However, since the parameter has a simple expression (it is the dominant singularity), then the integral formulas can be explicitly computed.

Proof. In order to use the Lemma 2.1, we need the three following estimations obtained by Euler-MacLaurin summations on the coefficients of $C(z)$.

$$\rho C'^{<n(1-\varepsilon)}(\rho) \underset{n\to\infty}{\sim} \frac{c_0 n^{\alpha+1}(1-\varepsilon)^{\alpha+1}}{(\alpha+1)\Gamma(\alpha)}, \qquad C^{>n(1+\varepsilon)}(\rho) \underset{n\to\infty}{\sim} -\frac{c_0 n^\alpha (1+\varepsilon)^\alpha}{\alpha\Gamma(\alpha)},$$

$$C^{[n(1-\varepsilon),n(1+\varepsilon)]}(\rho) \underset{n\to\infty}{\sim} \frac{c_0 n^\alpha}{\Gamma(\alpha)} \frac{(1+\varepsilon)^\alpha - (1-\varepsilon)^\alpha}{\alpha},$$

and

$$\rho^2 C''^{<n(1-\varepsilon)}(\rho) \underset{n\to\infty}{\sim} \frac{c_0 n^{\alpha+2}(1-\varepsilon)^{\alpha+2}}{(\alpha+2)\Gamma(\alpha)}.$$

□

4.3 Basic Boltzmann sampler

In the case when $-\alpha > 0$, we can still compute an approximation of the parameter x_n solution of $\mathbb{E}_{x_n}(N) = x_n$. However, to obtain this approximation, the computations are much more technical and the time complexities of the basic Boltzmann samplings (with and without anticipated rejection) are worst than in the previous cases. Thus we will not give any details on these cases.

5 Comparison of the methods for peaked distributions

The goal of this section consists in comparing the distinct methods of sampling effectively, by studying in details the previous complexity measures. An easy approach (already presented in [DFLS04]) is to let ε tend to 0 (with some care because of the two consecutive limits on n and then on ε), and then to compare the behaviour of the complexities.

5.1 A general class of objects with $0 < -\alpha < 1$

We are interested in objects of a combinatorial class, whose singular exponent $-\alpha$ belongs to $]0,1[$. By pointing one atom, we can apply the basic Boltzmann samplers (Theorems 3.1 and 3.2) with the exponent $-\alpha - 1$ or we can directly apply the singular Boltzmann sampler. We synthesize the results in the following Figure 2.

	Exact average complexity
Pointed B. s. without anticipated rej.	$n \cdot \dfrac{\int_{w=0}^{1-\varepsilon} w^{\alpha+1} e^{(-\alpha-1)w} dw + \int_{w=1+\varepsilon}^{\infty} w^{\alpha+1} e^{(-\alpha-1)w} dw}{\int_{w=1-\varepsilon}^{1+\varepsilon} w^{\alpha} e^{(-\alpha-1)w} dw}$
Pointed B. sampler with anticipated rej.	$n \cdot \dfrac{\int_{w=0}^{1-\varepsilon} w^{\alpha+1} e^{(-\alpha-1)w} dw + \int_{w=1+\varepsilon}^{\infty} (1+\varepsilon)\ w^{\alpha} e^{(-\alpha-1)w} dw}{\int_{w=1-\varepsilon}^{1+\varepsilon} w^{\alpha} e^{(-\alpha-1)w} dw}$
Singular B. sampler	$-\alpha n \left(\frac{(1-\varepsilon)^{\alpha+1}}{\alpha+1} + \frac{(1+\varepsilon)^{\alpha+1})}{-\alpha} \right) \left((1-\varepsilon)^{\alpha} - (1+\varepsilon)^{\alpha} \right)^{-1}$
	Approx. average complexity ($\varepsilon \to 0$ and $\frac{1}{\varepsilon} = o(n)$)
Pointed B. s. without anticipated rej.	$\frac{n}{2\varepsilon} \left(\left(\frac{e}{\alpha+1} \right)^{\alpha+1} \cdot \Gamma(\alpha+1) \right) + o\left(\frac{n}{\varepsilon} \right)$
Pointed B. sampler with anticipated rej.	$\frac{n}{2\varepsilon} \left(\left(\frac{e}{\alpha+1} \right)^{\alpha+1} \cdot \Gamma(\alpha+1) - \frac{1}{\alpha+1} \right) + o\left(\frac{n}{\varepsilon} \right)$
Singular B. sampler	$\frac{n}{2\varepsilon} \cdot \frac{1}{-\alpha(\alpha+1)} + o\left(\frac{n}{\varepsilon} \right)$

Figure 2: *Asymptotics of the average cumulated sizes of rejected objects, when* $0 < -\alpha < 1$.

In the light of this study, we observe that the different optimizations for Boltzmann sampling, already presented in the original paper [DFLS04], are efficient, even if the distinct average complexities do only differ by a constant factor. However, if we are interested in exact Boltzmann sampling, i.e. generating objects of an exact given size, then the previous results are valid for $\varepsilon = \frac{1}{2n}$. Thus in this context, we obtain quadratic complexity samplers.

5.2 Aperiodic and strongly connected grammars: $-\alpha = 1/2$

Let us first recall that in the case of aperiodic and strongly connected grammars, like grammars for trees, the approximate complexity of the Boltzmann singular methods was already presented in [DFLS04], but without explicit comparison of the methods.

By taking $-\alpha = 1/2$ in the previous formulas, when ε tends to 0 (with the restriction $\frac{1}{\varepsilon} = o(n)$). Direct calculations of the average complexities give the following results. In the case of the pointing

sampling without anticipated rejection we get the average complexity equivalent to:

$$\left(\sqrt{\frac{\pi e}{2}}\right) \cdot \frac{n}{\varepsilon}, \qquad \text{with the approximation } \sqrt{\frac{\pi e}{2}} \approx 2.066.$$

The improved case of the pointing sampling with anticipated rejection, we obtain the following equivalent formula for the complexity:

$$\left(\sqrt{\frac{\pi e}{2}} - 1\right) \cdot \frac{n}{\varepsilon}, \qquad \text{with the approximation } \sqrt{\frac{\pi e}{2}} - 1 \approx 1.066.$$

In the case of singular sampling, the complexity measure is asymptotically equivalent to:

$$2 \cdot \frac{n}{\varepsilon}.$$

Thus finally note that the pointing sampling without anticipated rejection is roughly equivalent to the one of the singular sampling. However, pointing sampling with anticipated rejection is almost twice as good!

5.3 Experimental comparison of samplers

Let us take an example of aperiodic and strongly connected grammar. We choose the classical specification of complete binary trees, where the size notion corresponds to the number of internal nodes. its specification is

$$\mathcal{B} = 1 + \mathcal{Z} \times \mathcal{B} \times \mathcal{B}.$$

The corresponding ordinary generating function corresponds to the Catalan numbers series and is

$$B(z) = \frac{1 - \sqrt{1 - 4z}}{2z}.$$

Thus its singular exponent is $-\alpha = 1/2$ and the trees present a peaked distribution. In the Figure 1, the peaked distributions in dashed blue lines are exactly the distribution for the complete binary trees with the parameters $x \in \{0.15, 0.24, 0.249\}$.

Our goal is to observe experimentally if the complexity measures of the samplers behave almost like in the asymptotics case already for small trees (several thousands of nodes). Our comparison will deal with the constructions of trees with one marked node (i.e. trees in $\mathcal{B}^\bullet$) and of trees in $\mathcal{B}$.

Let us first give some properties of the size distributions. First with the parameter $x = 0.249$ not so far from the dominant singularity for $\mathcal{B}$, the single tree of size 0 (the tree reduced to a leaf) has probability approximately 0.532 to be sampled. By sampling in $\mathcal{B}^\bullet$ instead than in $\mathcal{B}$, with the same parameter we obtain a tree of size from 1 to 74 with probability near to 0.501. Then by taking a larger parameter, the peaked distributions do only a little bit evolve. However, for the flat distributions, with probability almost 0.5 we will get larger and larger trees: for $x = 0.249999$ we sample trees of sizes from 1 to 3134 and with the parameter $x = 0.2499999$ we obtain with half probability trees from sizes in $[1, 570351]$.

For the different experiments, we have sampled a hundred of binary trees of size $10,000 \pm 1\%$, $100,000 \pm 10\%$ and $100,000 \pm 1\%$. The ratio n/ε is the same for the two first sets of trees and is smaller for the next one. In Figures 3, 4 and 5, we represent the normalized complexity (i.e. the complexity multiplied by n/ε) with respect to the size of the generated trees.

The trees represented in red points have been sampled with a singular sampler. In the first plot (tree size about 10.000), the average normalized complexity (represented by the red line) is approximately equal to 2.1072. For trees of size about $100,000$ the normalized mean value obtained by the singular sampler is approximately 2.1464 (for $\varepsilon = 0.1$) and 2.1330 (for $\varepsilon = 0.01$). In the context of the pointed sampling (blue crosses and dashed blue line representing the average values), with anticipated rejection, both normalized mean values are respectively about 1.1710 and 1.1055 and 1.0673.

By studying both plots of Figures 3, 4 and 5, we remark that the complexity from a sample to another could be relatively large. However by generating few trees (one hundred) the mean values are close to the theoretical asymptotic values.

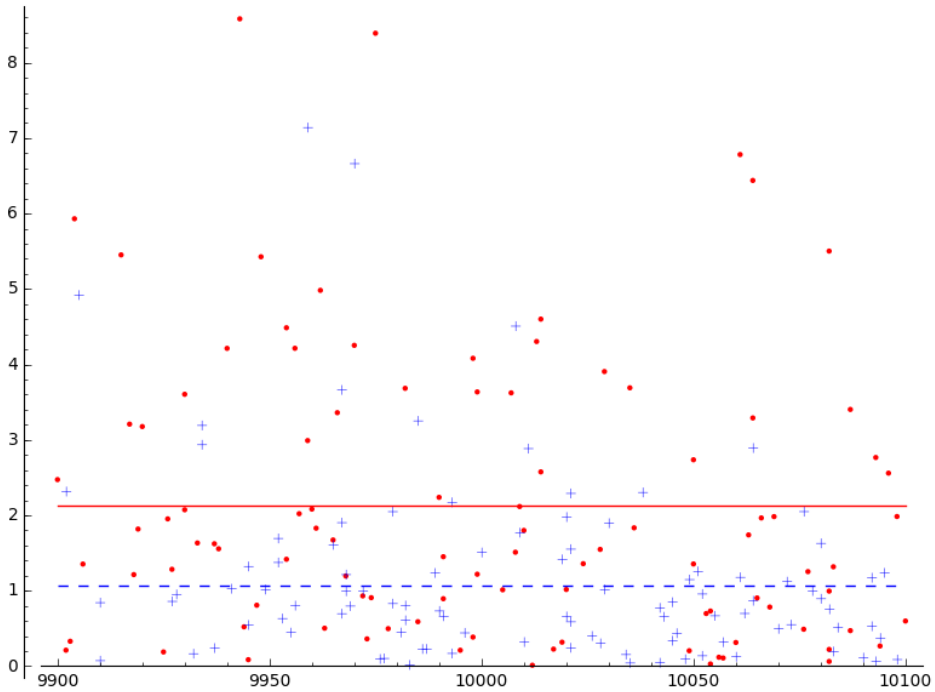

Figure 3: *Plot of sampled tree sizes* $(10,000 \pm 1\%)$ *and their corresponding normalized complexity.*

6 Conclusion

The focus of this paper is on Δ-singular functions, but the results can be broadened to a larger class of functions. Indeed the proofs do not need Δ-singularity: the approximation stated in Equation (1) is sufficient to obtain all the results.

Furthermore, the proofs sketches can be broadened to accept functions that are even less constrained, whose coefficients satisfy the following asymptotic behaviour:

$$C_n \underset{n\to\infty}{\sim} \frac{c_0}{\Gamma(\alpha)} \rho^{-n} n^{\alpha-1} \log(n)^{\beta-1}.$$

We can for example mention circular graphs that satisfy the latter constraint.

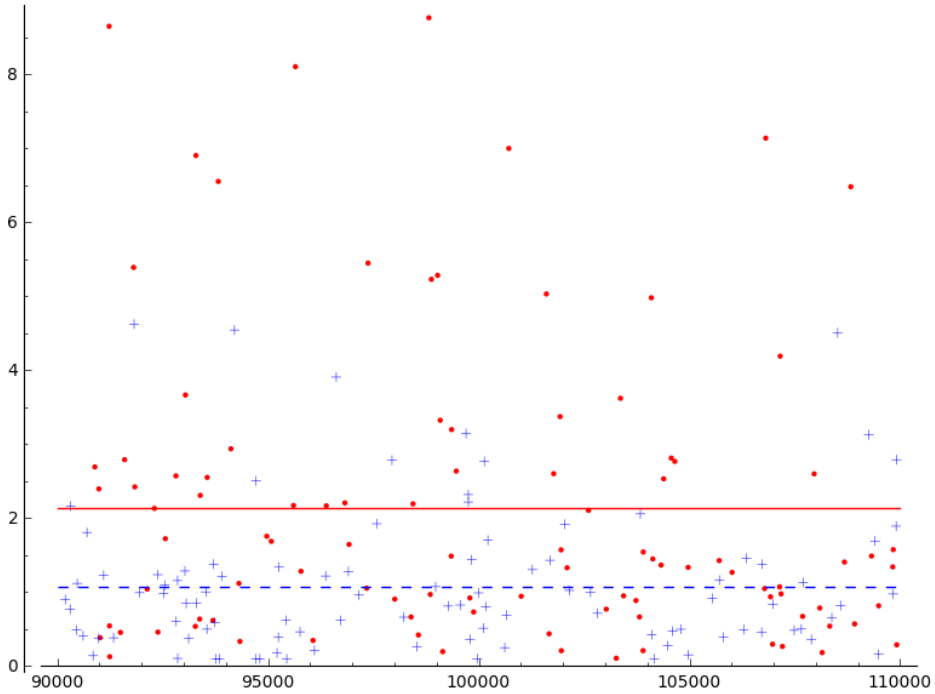

Figure 4: *Plot of sampled tree sizes* $(100,000 \pm 10\%)$ *and their corresponding normalized complexity.*

References

[BBJ13] A. Bacher, O. Bodini, and A. Jacquot. Exact-size sampling for motzkin trees in linear time via boltzmann samplers and holonomic specification. In *10th SIAM Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 52–61, 2013.

[BBJ14] A. Bacher, O. Bodini, and A. Jacquot. Efficient random sampling of binary and unary-binary trees via holonomic equations. *ArXiv e-prints*, 2014.

[BFP10] O. Bodini, E. Fusy, and C. Pivoteau. Random sampling of plane partitions. *Combinatorics, Probability & Computing*, 19(2):201–226, 2010.

[BLR15] O. Bodini, J. Lumbroso, and N. Rolin. Analytic samplers and the combinatorial rejection method. In *Proceedings of the Twelfth Workshop on Analytic Algorithmics and Combinatorics, ANALCO*, pages 40–50, 2015.

[BN07] F. Bassino and C. Nicaud. Enumeration and random generation of accessible automata. *Theor. Comput. Sci.*, 381(1-3):86–104, 2007.

[BNW08] F. Bassino, C. Nicaud, and P. Weil. Random generation of finitely generated subgroups of a free group. *IJAC*, 18(2):375–405, 2008.

[BP10] O. Bodini and Y. Ponty. Multi-dimensional boltzmann sampling of languages. In *21st International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, pages 49–64, Vienna, Austria, July 2010.

[BRS12] O. Bodini, O. Roussel, and M. Soria. Boltzmann samplers for first-order differential specifications. *Discrete Applied Mathematics*, 160(18):2563–2572, 2012.

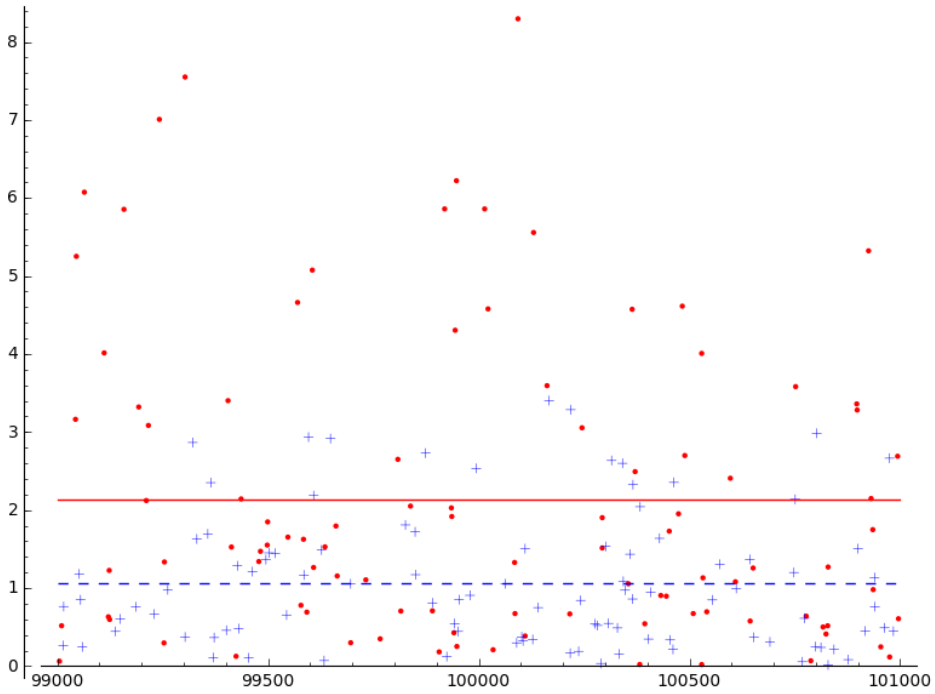

Figure 5: *Plots of sampled tree sizes* $(100,000 \pm 1\%)$ *and their corresponding normalized complexity.*

[CD09] B. Canou and A. Darrasse. Fast and sound random generation for automated testing and benchmarking in objective caml. In *Proceedings of the 2009 ACM SIGPLAN Workshop on ML*, ML '09, pages 61–70, New York, NY, USA, 2009. ACM.

[DFLS04] P. Duchon, P. Flajolet, G. Louchard, and G. Schaeffer. Boltzmann samplers for the random generation of combinatorial structures. *Combinatorics, Probability & Computing*, 13(4-5):577–625, 2004.

[Duc11] P. Duchon. Random generation of combinatorial structures: Boltzmann samplers and beyond. In *Winter Simulation Conference*, pages 120–132, 2011.

[FFP07] P. Flajolet, É. Fusy, and C. Pivoteau. Boltzmann sampling of unlabeled structures. In *Proceedings of the Fourth Workshop on Analytic Algorithmics and Combinatorics, ANALCO*, pages 201–211, 2007.

[FS09] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge Univ. Press, 2009.

[FZC94] P. Flajolet, P. Zimmermann, and B. Van Cutsem. A calculus for the random generation of labelled combinatorial structures. *Theor. Comput. Sci.*, 132(2):1–35, 1994.

[NW78] A. Nijenhuis and H. S. Wilf. *Combinatorial algorithms for computers and calculators*. Computer science and applied mathematics. Academic Press, New York, 1978.

[PSS12] C. Pivoteau, B. Salvy, and M. Soria. Algorithms for combinatorial structures: Well-founded systems and Newton iterations. *J. of Combinatorial Theory, Series A*, 119:1711–1773, 2012.

[Rém85] J. L. Rémy. Un procédé itératif de dénombrement d'arbres binaires et son application a leur génération aléatoire. *ITA*, 19(2):179–195, 1985.