Generalised and Quotient Models for Random And/Or Trees and Application to Satisfiability

Antoine Genitrini* and Cécile Mailler†

May 26, 2022

Abstract

This article is motivated by the following satisfiability question: pick uniformly at random an and/or Boolean expression of length n, built on a set of k_n Boolean variables. What is the probability that this expression is satisfiable? asymptotically when n tends to infinity?

The model of random Boolean expressions developed in the present paper is the model of Boolean Catalan trees, already extensively studied in the literature for a *constant* sequence $(k_n)_{n\geqslant 1}$. The fundamental breakthrough of this paper is to generalise the previous results for any (reasonable) sequence of integers $(k_n)_{n\geqslant 1}$, which enables us, in particular, to solve the above satisfiability question.

We also analyse the effect of introducing a natural equivalence relation on the set of Boolean expressions. This new *quotient* model happens to exhibit a very interesting threshold (or saturation) phenomena at $k_n = {}^{n}\!/_{\ln n}$.

Keywords: Boolean formulas/functions; Catalan trees; Equivalence relation; Probability distribution; Satisfiability; Analytic combinatorics.

1 Introduction

For several decades, satisfiability problems have been extensively studied by computer scientists and probabilists, as well as statistical physicists. In this paper, we focus on the probabilistic version of satisfiability problems: what is the probability that a *random* Boolean expression is satisfiable? The answer to this question obviously depends on the distribution considered on the set of Boolean expressions.

One of the most studied satisfiability problems is the 3–SAT problem. It consists in choosing uniformly at random an expression among conjunctions of n clauses, each clause being a disjunction of three literals - where literals are chosen among a set of k_n variables and their negations. What is the probability that such a random Boolean expression is satisfiable? when n tends to infinity?

This question is already partially answered – see for example [1]: the following phase transition is proven. If the ratio ${}^{k_n}\!/_{n}$ is small enough, then the random expression is satisfiable with probability tending to 1 when n tends to infinity, whereas if the ratio ${}^{k_n}\!/_{n}$ is large enough, then, this probability tends to 0. Refining this statement is the challenging aim of a large literature.

There are many other satisfiability problems. The K–SAT problem is for example the object of a recent breakthrough by Coja-Oghlan and Panagiotou [5] and Coja-Oghlan [4], who obtained the existence of a sharp threshold when K tends to infinity. The 2-XORSAT problem is studied by Daudé and Ravelomanana [6], using Analytic Combinatorics to exhibit and describe precisely a phase transition phenomenon.

The aim of the present paper is to define and study a new satisfiability model (i.e. a new distribution on the set of Boolean expressions) inspired by the literature on quantitative logics.

Quantitative logics, which origin might go back to the work of Woods [20], aims at answering this question: Which Boolean function does a *random* Boolean expression represent? Once again, the answer to this question deeply depends on the model of randomness chosen for Boolean expressions.

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

†Department of Mathematical Sciences, University of Bath, BA2 7AY Bath, UK. c.mailler@bath.ac.uk.

The Catalan tree model, first studied by Lefmann and Savický [15], is defined as follows: A Boolean tree is a binary plane rooted tree (i.e. a Catalan tree) whose internal nodes are labelled by the connectives and or or and whose leaves are labelled by k variables and their negations. Pick up uniformly at random a tree among Boolean trees of size n, and denote by $\mathbb{P}_{n,k}$ the distribution it induces on the set of Boolean functions. Lefmann and Savický first proved the existence of a limiting probability distribution $\mathbb{P}_k$ on Boolean functions when the size n of the random Boolean expression tends to infinity.

Since the seminal paper by Chauvin et al. [2], the Analytic Combinatorics' community aims at understanding better the Catalan tree distribution $\mathbb{P}_k$ (and similarly defined distributions) on the set of Boolean functions. In particular, Kozik [14] proves, in the Catalan tree model, an asymptotic (when k tends to infinity) relation between the probability of a given function and its *complexity* (i.e. the complexity of a Boolean function being the size of the smallest tree representing it). His powerful approach, the *pattern theory*, easily classifies and counts large expressions according to specific structural constraints. It will be generalised in the present paper.

Remark that in the Catalan tree model defined above, the size n of the Boolean expressions tend to infinity while the number k of literals labelling them is fixed. For technical reasons, k is then sent to infinity in order to obtain an asymptotic estimate of the probability of a given Boolean function. It means that the trees we consider have a lot of repetitions in their leaves: it is legitimate to ask if this bias the distribution induced on the set of Boolean functions. Genitrini and Kozik [12, 11] have proposed another model where random Boolean expressions are built on an infinite set of variables. This approach avoids the bias induced by letting n tend to infinity while k stays fixed.

Our paper extends the Catalan model in order both (1) to let n and k tend to infinity together and (2) to fit in the satisfiability context.

Following the extended abstract [13], we also look at the influence of a natural notion of equivalence on the set of Boolean expressions and functions. Roughly speaking, we say that two expressions or functions are equivalent if the second one can be obtain from the first one by renumbering the variables. As an example, the expressions (x_1 and x_2) and (x_{12} and x_3) are equivalent.

We will describe and study in parallel these two models (with an without equivalence classes) where the number of variables and the size of expressions jointly tend to infinity. Since the proofs will be very similar in both models, we will try general notations that fit both models. The model without equivalence classes will permit, as a corollary to answer the satisfiability problem in the context of Catalan Boolean expressions. It will be very interesting to see that, although the proofs are completely similar for both models, the probability distributions induced on the set of Boolean functions behave differently: the introduction of equivalence classes gives birth to an interesting and quite mysterious threshold phenomenon.

The paper is organised as follows. In Section 2 we define our two new models: the *generalised* model where the number of variables depends on the size of the considered trees and the *quotient* model where we introduce a natural equivalence relation on Boolean trees and functions. Section 3 is devoted to stating and discussing our three main results: the satisfiability question for random Catalan expressions; the link between the probability of a Boolean function (resp. a class of Boolean functions) and its *complexity*, both in the generalised and the quotient models. Section 4 and Section 5 contain the technical core of the paper: Section 4 is an analytic part focusing mainly on the difficulties arising from the introduction of the equivalence relation, while Section 5 concerns both models and discusses Kozik's pattern theory. Finally Section 6 contains the proofs of our main results.

2 Description of the two models

2.1 Contextual definitions

A **Boolean function** is a mapping from $\{0,1\}^{\mathbb{N}}$ into $\{0,1\}$. The two constant functions $(x_i)_{i\geqslant 1} \mapsto 1$ and $(x_i)_{i\geqslant 1} \mapsto 0$ are respectively called true and false.

An and/or **tree** is a binary plane tree whose leaves are labelled by literals, i.e. by elements of $\{x_i, \bar{x}_i\}_{i\in\mathbb{N}}$, and whose internal nodes are labelled by the connective and or the connective or, respectively denoted by $\wedge$ and $\vee$. We will say that x_i and $\bar{x}_i$ are two different literals but they are respectively the positive and the negative version of the same variable x_i. Every and/or tree is equivalent to a Boolean expression and thus represents a Boolean function: for example, the tree in Fig. 1 is equivalent to the expression $([x_1 \vee (\neg x_1 \vee x_2)] \vee x_3) \vee (x_4 \wedge x_1)$, where $\neg x = 1 - x$ for all $x \in \{0,1\}$, and represents the constant function true.

The **size** of an and/or tree is its number of leaves: remark that, for all $n \geqslant 1$, there is infinitely many and/or trees of size n. Finally we define the **tree-structure** of an and/or tree to be the and/or tree where the labels of the

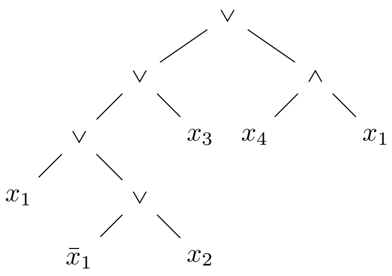

Figure 1: An and/or tree computing the constant function true.

leaves (but not of the internal nodes) have been removed.

Definition 1. *The* **complexity** *of a non constant Boolean function f, denoted by $L(f)$, is defined to be the size of its* **minimal trees***, i.e. the size of the smallest trees computing f. The complexity of* true *and* false *is defined to be 0.*

Although a Boolean function is defined on an infinite set of variables, it may actually depend only on a finite subset of *essential variables.*

Definition 2. *Given a Boolean function f, we say that the variable x is* **essential** *for f if, and only if, $f_{|x\leftarrow 0} \not\equiv f_{|x\leftarrow 1}$ (where $f_{|x\leftarrow\alpha}$ is the restriction of f to the subspace where $x = \alpha$). We denote by $E(f)$ the number of essential variables of f.*

Remark that the complexity and the number of essential variables of a Boolean function are related by the following inequalities: $E(f) \leqslant L(f) \leqslant 2^{E(f)+2}$ (see e.g. [7, p. 77–78] for the second inequality). Note that, asymptotically when $E(f)$ tends to infinity a tight asymptotic upper-bound is $2^{E(f)}/E(f)$, as proved by Lupanov [16] for the upper bound and Lutz [17] for the lower bound.

In the whole paper, our models propose a way to make n and k tend to infinity together:

Definition 3. *Let $(k_n)_{n\geqslant 1}$ be an increasing sequence of integers such that k_n tends to infinity when n tends to infinity.*

2.2 The generalised Catalan tree model

Let us recall the definition of the Catalan tree model defined and studied by Paris et al. [18], Lefmann & Savický [15], Chauvin et al. [2] and Kozik [14]. In those papers, the authors fix an integer $k \geqslant 1$ and consider the uniform distribution on and/or trees of size n whose leaf-labels are constrained to be in $\{x_1, \bar{x}_1, \dots, x_k, \bar{x}_k\}$. They study the induced distribution on the set of Boolean variables and prove that this distribution converges to a limit distribution $\mathfrak{p}_k$ when the size n of the trees tends to infinity. Given a Boolean function f, they then prove asymptotic theorems for $\mathfrak{p}_k(f)$ when k tends to infinity. In this approach, the order of the two limits (on n and then on k) is a priori important.

We define first the generalised Catalan tree model, that is a natural extension of the previous model.

The model (G) is defined as follows:

(1) consider the uniform distribution on and/or trees of size n which leaf-labels belong to $\{x_1, \bar{x}_1, \dots, x_{k_n}, \bar{x}_{k_n}\}$,

(2) denote by $\mathbb{P}_n$ the distribution it induces on the set of Boolean functions, and call this new distribution the **generalised Catalan tree distribution**.

Remark that there are A_n and/or trees of size n labelled with k_n variables, with

$$A_n = 2^{n-1}(2k_n)^n \cdot \mathtt{Cat}_n, \qquad \text{where } \mathtt{Cat}_n = \frac{1}{n}\binom{2n-2}{n-1}, \tag{1}$$

i.e. $\mathtt{Cat}_n$ is the number of binary plane trees having n leaves.

For all Boolean function f, we denote by $A_n(f)$ the number of and/or trees of size n labelled with k_n variables that compute f. Thus, by definition,

$$\mathbb{P}_n(f) = \frac{A_n(f)}{A_n}.$$

2.3 The quotient Catalan tree model

A second natural generalisation of the Catalan tree model is obtained by introducing equivalence classes of Boolean trees and functions. The idea is the following: the functions $(x_i)_{i\geqslant 1} \mapsto x_1 \wedge x_2$ and $(x_i)_{i\geqslant 1} \mapsto x_{38} \wedge \bar{x}_{12}$ can be seen as two realisations of the function *conjunction*.

Informally, two and/or trees are equivalent if the leaves of the first one can be relabelled (and negated) without collision in order to obtain the second tree. We define formally this equivalence relation as follows.

Definition 4. *Let A and B be two* and/or *trees. Trees A and B are* **equivalent** *if*

(i) their tree-structures are identical;

(ii) two leaves are labelled by the same variable in A if and only of they are labelled by the same variable in B;

(iii) two leaves are labelled by the same literal in A if and only of they are labelled by the same literal in B.

This equivalence relation on Boolean trees *induces straightforwardly an equivalence relation on Boolean functions.* Note that all functions of an equivalence class have the same complexity and the same number of essential variables. In the following, we will denote by $\langle f\rangle$ the equivalence class of the function f. We denote by $L\langle f\rangle = L(f)$ (resp. $E\langle f\rangle = E(f)$) the common complexity (resp. number of essential variables) of the elements of $\langle f\rangle$.

Definition 5. *Let $\langle f\rangle$ be a class of Boolean functions. The* **multiplicity** *of the class $\langle f\rangle$, is given by*

$$R\langle f\rangle = L\langle f\rangle - E\langle f\rangle.$$

It corresponds to the number of repetitions of variables in a minimal tree of a function from $\langle f\rangle$.

Recall that $(k_n)_{n\geqslant 1}$ is an increasing sequence of integers that tends to infinity when n tends to infinity. In the following, we only consider equivalence classes of trees having at least one element whose leaf-labels are in $\{x_1, \bar{x}_1, \ldots, x_{k_n}, \bar{x}_{k_n}\}$. It means that we restrict ourselves to trees of size n labelled by at most k_n different variables. Note that if $k_n \geqslant n$ for all $n \geqslant 1$, this is not a restriction because a tree of size n cannot contain more that n different leaf-labels.

The model (E) is defined as follows:

(1) consider the uniform distribution on classes of equivalence of trees of size n (labelled with at most k_n different variables),

(2) the distribution it induces on the set of equivalence classes of Boolean function is denoted by $\mathbb{P}_n$ and called the **quotient Catalan tree distribution**.

We denote by A_n the number of equivalence classes of trees of size n (in which at most k_n different variables appear as leaf-labels). Given a class of Boolean functions $\langle f\rangle$, we denote by $A_n\langle f\rangle$ the number of equivalence classes of trees of size n (labelled with at most k_n different variables) that compute a function of $\langle f\rangle$. We thus have

$$\mathbb{P}_n\langle f\rangle = \frac{A_n\langle f\rangle}{A_n}.$$

Proposition 1. *The number of classes of trees of size n satisfies:*

$$A_n = \mathtt{Cat}_n \cdot \sum_{p=1}^{k_n} \left\{ {n \atop p} \right\} 2^{2n-1-p},$$

where $\mathtt{Cat}_n$ is the number of (unlabelled) binary planar trees having n leaves (cf. Equation (1)*), and where $\left\{ {n \atop p} \right\}$ is the Stirling number of the second kind.*[1]

Proof. An equivalence class of and/or trees can be seen as

- a binary plane tree (factor $\mathtt{Cat}_n$)
- whose internal nodes are labelled by and and or connectives (factor 2^{n-1}),

[1] In Proposition 1, $\left\{ {n \atop p} \right\}$ is the number of partitions of n objects in p non-empty subsets (see e.g. [7, p. 735–737]).

- whose leaves are partitioned onto $1 \leqslant p \leqslant k_n$ parts (factor $\left\{{n \atop p}\right\}$),
- each of these parts being then partitioned onto two parts (one on them being possibly empty: factor 2^{n-p}).

□

Remark on notations: We have already used the notation A_n to define the model (G). We will keep the same notation for these two distinct objects because they will have the same role in the proofs. But formally, we have

$$A_n^{(\mathtt{G})} = \mathtt{Cat}_n \cdot 2^{2n-1} \cdot k_n^n \quad \text{and} \quad A_n^{(\mathtt{E})} = \mathtt{Cat}_n \cdot \sum_{p=1}^{k_n} \left\{ {n \atop p} \right\} 2^{2n-1-p}.$$

3 Main results and discussion

We have defined the two models we are interested in: the generalised and the quotient Catalan trees distributions. Both distributions are called $\mathbb{P}_n$ for simplicity's sake, but we will use $\mathbb{P}_n^{(\mathtt{G})}$ and $\mathbb{P}_n^{(\mathtt{E})}$ when the precision is needed. The aim of this paper is to study the behaviour of both distributions when the size n of the considered trees tends to infinity.

Let us remark that the distribution induced by (G) is based on an uniform distribution among trees of the same size. But the distribution induced by (E) lies on an uniform distribution among classes of trees of the same size. Obviously both induced distributions on Boolean functions are distinct.

Theorem 1 (Model (G))**.** *Let $(k_n)_{n\geqslant 1}$ be an increasing sequence of integers tending to infinity when n tends to infinity. For all Boolean functions f, there exists a positive constant $\alpha_f^{(\mathtt{G})}$ such that, asymptotically when n tends to infinity,*

$$\mathbb{P}_n(f) \sim \alpha_f^{(\mathtt{G})} \cdot \left(\frac{1}{k_n}\right)^{L(f)+1}.$$

This result has an interesting corollary concerning the Catalan-SAT problem: recall that a Boolean expression is said *satisfiable* if it does not represent the constant function false.

Corollary 1 (Catalan-SAT)**.** *Let $(k_n)_{n\geqslant 1}$ be an increasing sequence of integers tending to infinity when n tends to infinity. Pick up uniformly at random an* and/or *tree of size n with leaf-labels in $\{x_1, \bar{x}_1, \ldots, x_{k_n}, \bar{x}_{k_n}\}$. This random* and/or *tree is equivalent to a Boolean expression that is satisfiable with probability tending to 1 when n tends to infinity.*

Theorem 2 (Model (E))**.** *Let $(k_n)_{n\geqslant 1}$ be an increasing sequence of integers tending to infinity when n tends to infinity. There exists a sequence $(M_n)_{n\geqslant 1}$ such that $M_n \sim_{n\to\infty} \frac{n}{\ln n}$ and such that, for all fixed equivalence classes of Boolean functions $\langle f \rangle$, there exists a positive constant $\alpha_{\langle f \rangle}^{(\mathtt{E})}$ satisfying:*

(i) if, for all sufficiently large n, $k_n \leqslant M_n$, then, asymptotically when n tends to infinity,

$$\mathbb{P}_n\langle f \rangle \sim \alpha_{\langle f \rangle}^{(\mathtt{E})} \cdot \left(\frac{1}{k_{n+1}}\right)^{R\langle f \rangle+1};$$

(ii) if, for all sufficiently large n, $k_n \geqslant M_n$, then, asymptotically when n tends to infinity,

$$\mathbb{P}_n\langle f \rangle \sim \alpha_{\langle f \rangle}^{(\mathtt{E})} \cdot \left(\frac{\ln n}{n}\right)^{R\langle f \rangle+1}.$$

First note, that we could give some corollary about satisfiability for the second model (E) too. However, in the classical context of SAT problems, there are no quotient formulas. So we omit this by-product.

Let us discuss these results in view of the classical Catalan tree distribution studied by [2] and [14]: let us recall briefly its definition. Let $k \geqslant 1$ be an integer. We denote by $T_{n,k}$ the number of trees of size n, with leaf-labels in $\{x_1, \bar{x}_1, \ldots, x_k, \bar{x}_k\}$. Given a Boolean function f, we denote by $T_{n,k}(f)$ the number of such trees computing f. The Catalan distribution is thus defined by, for all Boolean functions f,

$$\mathfrak{p}_k(f) := \lim_{n\to+\infty} \frac{T_{n,k}(f)}{T_{n,k}}.$$

The existence of the above limit is proved in [15] or [2]. Kozik proved:

Theorem 3 (Kozik [14])**.** *Let k be a fixed positive integer. For all Boolean functions f, there exists a positive constant c_f such that*

$$\mathfrak{p}_k(f) \sim_{k\to\infty} c_f \cdot \left(\frac{1}{k}\right)^{L(f)+1}.$$

As one can see Theorems 1 and 3 are very similar, and we will see that their proofs are also very similar after having observed a simple but fundamental trick: one has to consider separately the tree-structure of an and/or tree and its leaf-labelling. It was not clear before this work how to generalise Kozik's proof in order to tackle the Catalan-SAT problem (cf. Corollary 1).

Introducing equivalence classes makes things different, and an interesting threshold effect appears (see Theorem 2). We still have no intuition for this threshold. Obviously we will see in the proof where it comes from.

In the classical Catalan tree model, each Boolean function is studied separately instead of being considered among its equivalence class. We can translate the result obtained by Kozik in terms of equivalence classes by summing over all Boolean functions belonging to a given equivalence class: note that there are $\binom{k}{E(f)}2^{E(f)}$ functions in the equivalence class of f. Therefore, the result of Kozik is equivalent to: for all classes $\langle f\rangle$, there exists a constant $c_{\langle f\rangle}$ such that, asymptotically when k tends to infinity,

$$\lim_{n\to+\infty} \mathfrak{p}_{n,k}\langle f\rangle \sim c_{\langle f\rangle}\left(\frac{1}{k}\right)^{L(f)-E(f)+1} = c_{\langle f\rangle}\left(\frac{1}{k}\right)^{R\langle f\rangle+1}.$$

The classical Catalan tree distribution can be seen as a degenerate case of our model where there exists a fixed integer k such that $k_n = k$ for all $n \geqslant 1$. Recall that we assume in the present paper that k_n tends to infinity when n tend to infinity: the case $k_n = k$ is thus not a particular case of our results, but only a degenerate one.

Once again, the proof of Theorem 2 relies on similar ideas as Kozik's proof of Theorem 3. To emphasise the similarities between the proof of our two main theorems (Theorems 1 and 2), we will develop their proofs together in Section 6.

4 Technical key point

As we already mentioned, the key idea of this paper is to separate the tree-structure of an and/or tree and its leaf-labelling. Recall that

$$A_n^{(\mathtt{G})} = 2^{n-1}\mathtt{Cat}_n \cdot (2k_n)^n \quad \text{and} \quad A_n^{(\mathtt{E})} = 2^{n-1}\mathtt{Cat}_n \cdot \sum_{p=1}^{k_n} \left\{ n \atop p \right\} 2^{n-p}.$$

For all $m, n \geqslant 1$, let us denote by

$$\mathtt{Lab}_{n,m} := \begin{cases} (2m)^n & \text{in model } (\mathtt{G}); \\ 2^n \cdot \displaystyle\sum_{p=1}^{m} \left\{ n \atop p \right\} 2^{-p} & \text{in model } (\mathtt{E}). \end{cases}$$

In both models, $\mathtt{Lab}_{n,m}$ corresponds to the number of ways to label the n leaves with m variables, thus

$$A_n = 2^{n-1}\mathtt{Cat}_n \cdot \mathtt{Lab}_{n,k_n}.$$

Finally, let us introduce the key quantity

$$\mathsf{rat}_n := \frac{\mathtt{Lab}_{n-1,k_n}}{\mathtt{Lab}_{n,k_n}}.$$

Note that in the model (G), the quantity $1/\mathsf{rat}_n = 2k_n$ corresponds to the number of the possible labellings of the $(n+1)^{\text{th}}$ leaf once the other leaves are already labelled. In the model (E), the leaf-labellings are not longer independent and this quantity $1/\mathsf{rat}_n$ is thus less explicit. A detailed analysis of this quantity is needed in the following. This section is devoted to its asymptotic analysis.

Proposition 2. *Let $(k_n)_{n\geqslant 1}$ be an increasing sequence of integer tending to infinity when n tends to infinity.*

(G) *For all integer p,*

$$\frac{\mathtt{Lab}_{n-p,k_n}}{\mathtt{Lab}_{n,k_n}} = \frac{1}{(2k_n)^p}.$$

(E) *There exists a sequence $(M_n)_{n\geqslant 1}$ with $M_n \sim_{n\to\infty} \frac{n}{\ln n}$ and such that, for all integer p, asymptotically when n tends to infinity,*

$$\frac{\mathtt{Lab}_{n-p,k_n}}{\mathtt{Lab}_{n,k_n}} = \begin{cases} \frac{1+o(1)}{(2k_n)^p} & \text{if } k_n \leqslant M_n \text{ for large enough } n; \\ (1+o(1))\ \left(\frac{\ln n}{2n}\right)^p & \text{if } k_n \geqslant M_n \text{ for large enough } n. \end{cases}$$

In particular, taking $p = 1$ gives

Proposition 3. *Let $(k_n)_{n\geqslant 1}$ be an increasing sequence of integer tending to infinity when n tends to infinity.*

(G) $\mathsf{rat}_n = \dfrac{1}{2k_n}$.

(E) *There exists a sequence $(M_n)_{n\geqslant 1}$ with $M_n \sim_{n\to\infty} \frac{n}{\ln n}$ and such that, asymptotically when n tends to infinity,*

$$\mathsf{rat}_n = \begin{cases} \frac{1+o(1)}{2k_n} & \text{if } k_n \leqslant M_n \text{ for large enough } n; \\ (1+o(1))\ \frac{\ln n}{2n} & \text{if } k_n \geqslant M_n \text{ for large enough } n. \end{cases}$$

Remark that, with this definition of rat_n, Theorems 1 and 2 can be rephrased as: for all Boolean functions f, there exists constants

$$\mathbb{P}_n^{(\mathtt{G})}(f) \sim \lambda_f \cdot \mathsf{rat}_n^{L(f)+1},$$

and

$$\mathbb{P}_n^{(\mathtt{E})}\langle f\rangle \sim \lambda_{\langle f\rangle} \cdot \mathsf{rat}_n^{R\langle f\rangle+1}.$$

The proof of Proposition 3 (G) is obvious and the rest of this section is devoted to the more technical proof of Proposition 3 (E).

The following proposition, which can be seen as some particular case of Bonferroni inequalities allows to exhibit bounds on $\mathtt{Lab}_{n,k_n}$.

Proposition 4 (cf. for example [19])**.** *For all $n \geqslant 1$, for all $p \in \{1,\dots,n\}$,*

$$\frac{p^n}{p!} - \frac{(p-1)^n}{(p-1)!} \leqslant \left\{ {n \atop p} \right\} \leqslant \frac{p^n}{p!}.$$

In view of these inequalities and of the expression of $\mathtt{Lab}_{n,k_n}$, both the following sequences naturally appear:

Lemma 1. *Let n be a positive integer.*

(i) The following sequence is unimodal:

$$\left(a_p^{(n)}\right)_{p\in\{1,\dots,n\}} = \left(\frac{p^n}{p!}2^{-p}\right)_{p\in\{1,\dots,n\}},$$

i.e. there exists an integer M_n such that $\left(a_p^{(n)}\right)_p$ is strictly increasing on $\{1,\dots,M_n\}$ and strictly decreasing on $\{M_n+1,\dots,n\}$.

(ii) Moreover, the sequence $(M_n)_n$ is increasing and asymptotically satisfies:

$$M_n \sim_{n\to\infty} \frac{n}{\ln n}.$$

Proof. (i) Let us prove that the sequence $\left(a_p^{(n)}\right)_{1\leqslant p\leqslant n}$ is log-concave, i.e. that the sequence $\left(\frac{a_{p+1}^{(n)}}{a_p^{(n)}}\right)_{1\leqslant p\leqslant n-1}$ is decreasing. Let p be an integer in $\{1,\dots,n-1\}$. By Definition of $a_p^{(n)}$:

$$\frac{a_{p+1}^{(n)}}{a_p^{(n)}} = \left(\frac{p+1}{p}\right)^n \cdot \frac{1}{2(p+1)},$$

and consequently, for all $n\geqslant 0$,

$$\frac{a_{p+1}^{(n)}}{a_p^{(n)}} > 1 \iff n\ln\left(\frac{p+1}{p}\right) - \ln(2p+2) > 0.$$

The function $\phi_n\ :\ p\mapsto n\ln\left(\frac{p+1}{p}\right)-\ln(2p+2)$ is strictly decreasing. Note that both $\phi_n(1)$ and $\phi_n(n-1)$ are tending to infinity when n tends to infinity. Then, for all n large enough, there exists a unique M_n such that $\left(a_p^{(n)}\right)_p$ is strictly increasing on $\{1,\dots,M_n\}$ and strictly decreasing on $\{M_n+1,\dots,n\}$. Let us suppose n large enough for the rest of the proof.

(ii) Let us denote by x_n the single solution of equation:

$$\left(\frac{x+1}{x}\right)^n \cdot \frac{1}{2(x+1)} = 1, \qquad \text{when it exists.} \tag{2}$$

First remark that the sequence $(x_n)_{n\geqslant 1}$ is increasing. We indeed know: $\phi_n(x_n)=0$ and $\phi_{n+1}(x_{n+1})=0$, which implies that $\phi_n(x_{n+1}) = -\ln\left(1+\frac{1}{x_{n+1}}\right) < 0$. Therefore, since for each n, the function ϕ_n is decreasing, we have that $x_{n+1}\geqslant x_n$, for all large enough n. Therefore, the sequence $(M_n)_{n\geqslant 1}$ is asymptotically increasing.

Since, asymptotically when n tends to infinity,

$$\left(\frac{\frac{n}{\ln n}+1}{\frac{n}{\ln n}}\right)^n \cdot \frac{1}{2(\frac{n}{\ln n}+1)} \sim \frac{\ln n}{2},$$

we have that $n/\ln n \leqslant x_n$ and therefore, x_n tends to infinity. Thus, Equation (2) evaluated in x_n is equivalent to

$$n\ln\left(1+\frac{1}{x_n}\right) = \ln 2 + \ln(x_n+1), \tag{3}$$

which implies $x_n\ln x_n \sim n$, when n tends to infinity. We easily deduce from this asymptotic relation that $\ln x_n \sim \ln n$ and that $x_n \sim \frac{n}{\ln n}$ when n tends to infinity. Since $M_n = \lfloor x_n \rfloor$, we conclude that $M_n \sim {}^{n}\!/_{\ln n}$, when n tends to infinity. □

We are now ready to understand the asymptotic behaviour of $\mathtt{Lab}_{n,k_n}/2^n$: *roughly speaking, asymptotically, the sum* $\mathtt{Lab}_{n,k_n}/2^n$ *does essentially only depend on the terms around* M_n.

Lemma 2. *Let $(u_n)_{n\geqslant 1}$ be an increasing sequence such that $u_n\leqslant n$ for all integer $n\geqslant 1$ and u_n tends to infinity when n tends to infinity.*

(i) If, for all large enough n, $u_n\leqslant M_n$, then, for all sequences $(\delta_n)_{n\geqslant 1}$ such that $\delta_n = o(u_n)$ and $\frac{u_n\sqrt{\ln u_n}}{\sqrt{n}} = o(\delta_n)$, we have, asymptotically when n tends to infinity,

$$\frac{\mathtt{Lab}_{n,u_n}}{2^n} = (1+o(1)) \sum_{p=u_n-\delta_n}^{u_n} \frac{p^n}{p!}2^{-p}. \tag{4}$$

(ii) If, for large enough n, $u_n\geqslant M_n$, then, for all sequences $(\delta_n)_{n\geqslant 1}$ such that $\delta_n = o(u_n)$ and $\frac{u_n\sqrt{\ln u_n}}{\sqrt{n}} = o(\delta_n)$, for all sequences $(\eta_n)_{n\geqslant 1}$ such that $\eta_n = o(M_n)$, $\lim_{n\to+\infty}\frac{\eta_n^2}{M_n} = +\infty$ and $\sqrt{M_n\ln(u_n-M_n)} = o(\eta_n)$, we have, asymptotically when n tends to $+\infty$,

$$\frac{\mathtt{Lab}_{n,u_n}}{2^n} = (1+o(1)) \sum_{p=M_n-\delta_n}^{\min\{M_n+\eta_n,u_n\}} \frac{p^n}{p!}2^{-p}. \tag{5}$$

Proof of Lemma 2 (i). Via Proposition 4, we can bound $\frac{\mathtt{Lab}_{n,u_n}}{2^n}$: for all $n \geqslant 1$,

$$\frac{1}{2} \cdot \sum_{p=1}^{u_n-1} \frac{p^n}{p!\, 2^p} + \frac{u_n^n}{u_n!\, 2^{u_n}} \leqslant \frac{\mathtt{Lab}_{n,u_n}}{2^n} \leqslant \sum_{p=1}^{u_n} \frac{p^n}{p!\, 2^p}. \tag{6}$$

Let us assume that $u_n \leqslant M_n$ for all large enough n, and let us prove that the two bounds of Equations (6) are of the same asymptotic order when n tends to infinity.

Denote, for all integer $m \geqslant 1$, $S_m = \sum_{p=1}^{m} a_p^{(n)}$. Thus Equations (6) implies

$$\frac{S_{u_n}}{2} \leqslant \frac{\mathtt{Lab}_{n,u_n}}{2^n} \leqslant S_{u_n}.$$

Let us split the sum S_{u_n} into two parts: the last δ_n summands, and the rest.

$$S_{u_n} = S_{u_n-\delta_n-1} + \sum_{p=u_n-\delta_n}^{u_n} a_p^{(n)}.$$

By assumption, $\delta_n = o(u_n)$ and we therefore can choose n large enough such that $u_n > \delta_n$. Let us prove that $S_{u_n-\delta_n-1}$ is negligible in front of a_{u_n}, and thus in front of $\sum_{p=u_n-\delta_n}^{u_n} a_p^{(n)}$. Recall that $\left(a_p^{(n)}\right)_{p\geqslant 1}$ is increasing on $\{1,\ldots,M_n\}$, which implies

$$S_{u_n-\delta_n-1} \leqslant u_n \cdot a_{u_n-\delta_n}.$$

For all large enough n, via Stirling formula, we deduce:

$$\begin{aligned}\frac{a_{u_n-\delta_n}}{a_{u_n}} &= 2^{\delta_n} \left(\frac{u_n-\delta_n}{u_n}\right)^n \frac{u_n!}{(u_n-\delta_n)!} = \left(\frac{2u_n}{\mathtt{e}}\right)^{\delta_n} \left(\frac{u_n-\delta_n}{u_n}\right)^{n-u_n+\delta_n-\frac{1}{2}} (1+o(1)) \\ &= \exp\left[\delta_n \ln\left(\frac{2u_n}{\mathtt{e}}\right) + \left(n-u_n+\delta_n-\frac{1}{2}\right)\ln\left(1-\frac{\delta_n}{u_n}\right) + o(1)\right].\end{aligned}$$

Since $\delta_n = o(u_n)$, we get $\ln\left(1-\frac{\delta_n}{u_n}\right) = -\frac{\delta_n}{u_n} - \frac{\delta_n^2}{2u_n^2} + o\left(\frac{\delta_n^2}{u_n^2}\right)$. Moreover, $u_n \leqslant M_n$ thus,

$$\frac{a_{u_n-\delta_n}}{a_{u_n}} = \exp\left[\delta_n \ln 2 + \delta_n \ln u_n - \frac{n\delta_n}{u_n} - \frac{n\delta_n^2}{2u_n^2} + o\left(\frac{n\delta_n^2}{u_n^2}\right)\right].$$

Therefore, by using $u_n \leqslant M_n$, and Equation (3), we deduce $\frac{n}{M_n} \geqslant \ln 2 + \ln M_n$,

$$\begin{aligned}\frac{a_{u_n-\delta_n}}{a_{u_n}} &\leqslant \exp\left[\delta_n \ln 2 + \delta_n \ln M_n - \frac{n\delta_n}{M_n} - \frac{n\delta_n^2}{2u_n^2} + o\left(\frac{n\delta_n^2}{u_n^2}\right)\right] \\ &\leqslant \exp\left[-\frac{n\delta_n^2}{2u_n^2} + o\left(\frac{n\delta_n^2}{u_n^2}\right)\right].\end{aligned}$$

From the assumption $\frac{u_n\sqrt{\ln u_n}}{\sqrt{n}} = o(\delta_n)$, we deduce $\ln u_n = o\left(\frac{n\delta_n^2}{u_n^2}\right)$, thus we can conclude

$$\frac{S_{u_n-\delta_n-1}}{a_{u_n}} \leqslant u_n \frac{a_{u_n-\delta_n}}{a_{u_n}} \leqslant \exp\left[\ln u_n - \frac{n\delta_n^2}{2u_n^2} + o\left(\frac{n\delta_n^2}{u_n^2}\right)\right] = o(1).$$

And consequently, we get $S_{u_n} \sim_{n\to\infty} \sum_{p=u_n-\delta_n}^{u_n} a_p^{(n)}$. □

Proof of Lemma 2, (ii). Assume that $u_n \geqslant M_n$ for all large enough n. Let us split the sums of the lower and upper bounds of Equations (6) into three parts: the first from index 1 to $M_n - \delta_n - 1$, the second from index $M_n - \delta_n$ to $M_n + \eta_n$, and the third from index $M_n + \eta_n + 1$ to u_n. Remark that, if $u_n \leqslant M_n + \eta_n$, then the third part is empty and the second one is truncated:

$$S_{u_n} = S_{M_n-\delta_n-1} + \sum_{p=M_n-\delta_n}^{M_n+\eta_n} a_p^{(n)} + \sum_{p=M_n+\eta_n+1}^{u_n} a_p^{(n)}.$$

By arguments similar to those developed in the proof of assertion *(i)*, we can prove that $S_{M_n-\delta_n-1}$ is negligible in front of $a^{(n)}_{M_n}$, and thus in front of $\sum_{p=M_n-\delta_n}^{M_n+\eta_n} a_p^{(n)}$. Therefore, if $u_n \leqslant M_n+\eta_n$, assertion *(ii)* is proved. Let us now assume that $u_n \geqslant M_n+\eta_n+1$: to end the proof, we prove that $\sum_{p=M_N+\eta_n+1}^{u_n} a_p^{(n)}$ is negligible in front of $a^{(n)}_{M_n}$, and thus in front of $\sum_{p=M_n-\delta_n}^{M_n+\eta_n} a_p^{(n)}$.

In view of Lemma 1, we have

$$\sum_{p=M_n+\eta_n+1}^{u_n} a_p^{(n)} \leqslant (u_n - M_n - \eta_n) \cdot a^{(n)}_{M_n+\eta_n}.$$

Via Stirling formula,

$$\begin{aligned}\frac{a^{(n)}_{M_n+\eta_n}}{a^{(n)}_{M_n}} &= 2^{-\eta_n}\left(\frac{M_n+\eta_n}{M_n}\right)^n \frac{M_n!}{(M_n+\eta_n)!} = \left(\frac{2(M_n+\eta_n)}{\mathsf{e}}\right)^{-\eta_n}\left(\frac{M_n+\eta_n}{M_n}\right)^{n-M_n-\frac{1}{2}}(1+o(1))\\ &= \exp\left[-\eta_n \ln\left(\frac{2(M_n+\eta_n)}{\mathsf{e}}\right) + \left(n - M_n - \frac{1}{2}\right)\ln\left(1+\frac{\eta_n}{M_n}\right) + o(1)\right].\end{aligned}$$

Since $\ln\left(1+\frac{\eta_n}{M_n}\right) \leqslant \frac{\eta_n}{M_n}$, we get:

$$\begin{aligned}\frac{a^{(n)}_{M_n+\eta_n}}{a^{(n)}_{M_n}} &\leqslant \exp\left[-\eta_n \ln 2 + \eta_n - \eta_n \ln(M_n+\eta_n) + \frac{\eta_n}{M_n}(n - M_n - \frac{1}{2}) + o(1)\right]\\ &= \exp\left[-\eta_n \ln 2 - \eta_n \ln(M_n+\eta_n) + \frac{n\eta_n}{M_n} + o(1)\right].\end{aligned}$$

Our assumption states $\frac{\eta_n}{M_n} = o(1)$, thus

$$\begin{aligned}\frac{a^{(n)}_{M_n+\eta_n}}{a^{(n)}_{M_n}} &\leqslant \exp\left[-\eta_n \ln 2 - \eta_n \ln M_n - \eta_n \ln\left(1+\frac{\eta_n}{M_n}\right) + \frac{n\eta_n}{M_n} + o(1)\right]\\ &= \exp\left[-\eta_n \ln 2 - \eta_n \ln M_n - \frac{\eta_n^2}{M_n} + \frac{n\eta_n}{M_n} + \mathcal{O}\left(\frac{\eta_n^3}{M_n^2}\right)\right]\end{aligned}$$

Since $M_n = \lfloor x_n \rfloor$, we have

$$n \ln\left(1+\frac{1}{x_n}\right) = n\left(\frac{1}{M_n} - \frac{1}{2M_n^2} + \mathcal{O}\left(\frac{1}{M_n^3}\right)\right),$$

therefore

$$\ln 2 + \ln(x_n+1) = \ln 2 + \ln M_n + \mathcal{O}\left(\frac{1}{M_n}\right).$$

Equation (3) implies:

$$\begin{aligned}\frac{n}{M_n} &= \ln 2 + \ln M_n + \frac{n}{2M_n^2} + \mathcal{O}\left(\frac{n}{M_n^3}\right) + \mathcal{O}\left(\frac{1}{M_n}\right)\\ &= \ln 2 + \ln M_n + \frac{n}{2M_n^2} + \mathcal{O}\left(\frac{n}{M_n^3}\right),\end{aligned}$$

because $\frac{1}{M_n} = o(\frac{n}{M_n^3})$. Thus, we conclude

$$\frac{a^{(n)}_{M_n+\eta_n}}{a^{(n)}_{M_n}} \leqslant \exp\left[-\frac{\eta_n^2}{M_n} + \mathcal{O}\left(\frac{\eta_n^3}{M_n^2}\right) + \mathcal{O}\left(\frac{n\eta_n}{M_n^3}\right)\right] = \exp\left[-\frac{\eta_n^2}{M_n} + o\left(\frac{\eta_n^2}{M_n}\right)\right],$$

because, from assumption: $\sqrt{M_n \ln(u_n - M_n)} = o(\eta_n)$, we deuce $\sqrt{M_n} = o(\eta_n)$. Finally we get

$$\frac{\sum_{p=M_n+\eta_n+1}^{u_n} a_p^{(n)}}{a^{(n)}_{M_n}} \leqslant (u_n - M_n - \eta_n)\frac{a^{(n)}_{M_n+\eta_n}}{a^{(n)}_{M_n}} \leqslant \exp\left[\ln(u_n - M_n) - \frac{\eta_n^2}{M_n} + o\left(\frac{\eta_n^2}{M_n}\right)\right] = o(1),$$

since, by assumption, $\sqrt{M_n \ln(u_n - M_n)} = o(\eta_n)$. Therefore, asymptotically when n tends to infinity,

$$S_{u_n} \sim \sum_{p=M_n-\delta_n}^{M_n+\eta_n} a_p^{(n)},$$

which concludes the proof. □

We are now ready for the proof of Proposition 2: let us decompose this proof in the two following Lemmas 3 and 4:

Lemma 3. *Let $(k_n)_{n\geqslant 1}$ be a sequence of integerssuch that $k_n \leqslant M_n$ for large enough n, then, for all integer p, asymptotically when n tends to infinity,*

$$\frac{\mathtt{Lab}_{n-p,k_n}}{\mathtt{Lab}_{n,k_n}} = (1+o(1)) \left(\frac{1}{(2k_n)^p}\right).$$

Proof. **(i) Let us first assume that $\boldsymbol{k_n \leqslant M_{n-p}}$.** Let $(\delta_n)_{n\geqslant 1}$ an integer-valued sequence such that $\delta_n = o(k_n)$ and $\frac{k_n\sqrt{\ln k_n}}{\sqrt{n}} = o(\delta_n)$ when n tends to infinity. Lemma 2 applied to $u_n = k_n$ gives, asymptotically when n tends to infinity,

$$\frac{\mathtt{Lab}_{n,k_n}}{2^n} = (1+o(1)) \sum_{i=k_n-\delta_n}^{k_n} a_i^{(n)}.$$

Moreover, since $k_n \leqslant M_{n-p}$, and since the sequence $(\delta_n)_{n\geqslant 1}$ satisfies $\delta_n = o(k_n)$ and $\frac{k_n\sqrt{\ln k_n}}{\sqrt{n-p}} = o(\delta_n)$, applying Lemma 2 to the sequence $u_n = k_n$ gives us, asymptotically when n tends to infinity,

$$\frac{\mathtt{Lab}_{n-p,k_n}}{2^{n-p}} = (1+o(1)) \sum_{i=k_n-\delta_n}^{k_n} a_i^{(n-p)}.$$

Therefore,

$$\frac{\mathtt{Lab}_{n-p,k_n}}{\mathtt{Lab}_{n,k_n}} = (2^{-p}+o(1)) \frac{\sum_{i=k_n-\delta_n}^{k_n} a_i^{(n-p)}}{\sum_{i=k_n-\delta_n}^{k_n} a_i^{(n)}}.$$

We have

$$(k_n-\delta_n)^p \sum_{i=k_n-\delta_n}^{k_n} a_i^{(n-p)} \quad \leqslant \quad \sum_{i=k_n-\delta_n}^{k_n} i^p a_i^{(n-p)} = \sum_{i=k_n-\delta_n}^{k_n} a_i^{(n)} = \sum_{i=k_n-\delta_n}^{k_n} i^p a_p^{(n-p)} \leqslant k_n^p \sum_{p=k_n-\delta_n}^{k_n} a_p^{(n-p)},$$

which implies

$$\frac{\mathtt{Lab}_{n-p,k_n}}{\mathtt{Lab}_{n,k_n}} \sim \frac{1}{(2k_n)^p} \quad \text{when } n \to +\infty.$$

(ii) Now assume that $\boldsymbol{M_{n-p} < k_n \leqslant M_n}$. Let $(\delta_n)_{n\geqslant 1}$ be an integer-valued sequence such that $\delta_n = o(k_n)$ and $\frac{k_n\sqrt{\ln k_n}}{\sqrt{n-p}} = o(\delta_n)$. Let $(\eta_n)_{n\geqslant 1}$ be an integer-valued sequence such that $\eta_n = o(M_{n-p})$, and $\sqrt{M_{n-p}\ln(k_n - M_{n-p})} = o(\eta_n)$. Applying Lemma 2 *(ii)* to the sequence $u_n = k_n$, we obtain

$$\frac{\mathtt{Lab}_{n-p,k_n}}{2^{n-p}} = (1+o(1)) \sum_{i=M_{n-p}-\delta_n}^{\min\{M_{n-p}+\eta_n,k_n\}} a_i^{(n-p)}.$$

Moreover, since $\delta_n = o(k_n)$ and $\frac{k_n\sqrt{\ln k_n}}{\sqrt{n}} = o(\delta_n)$, via Lemma 2 *(i)*,applied to the sequence $u_n = k_n$,

$$\frac{\mathtt{Lab}_{n,k_n}}{2^n} = (1+o(1)) \sum_{i=k_n-\delta_n}^{k_n} a_i^{(n)}.$$

Let us remark, as above, that

$$(k_n-\delta_n)^p \sum_{i=k_n-\delta_n}^{k_n} a_i^{(n-p)} \leqslant \frac{\mathtt{Lab}_{n,k_n}}{2^n} \leqslant k_n^p \sum_{i=k_n-\delta_n}^{k_n} a_i^{(n-p)}.$$

Moreover, since $k_n > M_{n-p}$, using similar arguments as those developed to prove Lemma 2 *(i)*,

$$\sum_{i=k_n-\delta_n}^{k_n} a_i^{(n-p)} \sim \sum_{i=k_n-\delta_n}^{\min\{k_n,M_{n-p}+\eta_n\}} a_i^{(n-p)} \sim \frac{\texttt{Lab}_{n-p,k_n}}{2^{n-p}}.$$

Therefore, since $\delta_n = o(k_n)$, we get

$$\frac{\texttt{Lab}_{n-p,k_n}}{\texttt{Lab}_{n,k_n}} = (1+o(1))\,\frac{1}{(2k_n)^p},$$

which concludes the proof. □

Lemma 4. *Let $(k_n)_{n\geqslant 1}$ be a sequence of integers that tends to infinity when n tends to infinity. Let us assume that $k_n \geqslant M_n$ for large enough n, then, for all integer p, asymptotically when n tends to infinity,*

$$\frac{\texttt{Lab}_{n-p,k_n}}{\texttt{Lab}_{n,k_n}} = (1+o(1))\,\left(\frac{\ln n}{2n}\right)^p.$$

Proof. By assumption, $k_n \geqslant M_n$, which implies $k_n \geqslant M_{n-p}$. Let $(\delta_n)_{n\geqslant 1}$ be a sequence of integers such that $\delta_n = o(M_{n-p})$ and $\frac{M_n\sqrt{\ln M_n}}{\sqrt{n}} = o(\delta_{n+p})$. Let $(\eta_n)_{n\geqslant 1}$ be another sequence of integers such that $\eta_n = o(M_{n-p})$, and $\sqrt{M_n \ln(k_n - M_n)} = o(\eta_{n+p})$. We thus can apply Lemma 2 *(ii)* to $u_n = k_n$ and conclude that, asymptotically when n tends to infinity,

$$\frac{\texttt{Lab}_{n-p,k_n}}{2^{n-p}} = (1+o(1)) \sum_{i=M_{n-p}-\delta_n}^{\min\{M_{n-p}+\eta_n,k_n\}} a_i^{(n-p)}.$$

Moreover, since the sequence $(\delta_n)_{n\geqslant 1}$ verifies $\delta_n = o(M_{n-p}) = o(M_n)$ and $\frac{M_n\sqrt{\ln M_n}}{\sqrt{n}} = o(\delta_{n+p}) = o(\delta_n)$, and since the sequence $(\eta_n)_{n\geqslant 1}$ verifies $\eta_n = o(M_{n-p}) = o(M_n)$, and $\sqrt{M_n \ln(k_n - M_n)} = o(\eta_{n+p}) = o(\eta_n)$, we have,

$$\frac{\texttt{Lab}_{n,k_n}}{2^n} = (1+o(1)) \sum_{i=M_n-\delta_n}^{\min\{M_n+\eta_n,k_n\}} a_i^{(n)}.$$

Let us note that

$$(M_n-\delta_n)^p \sum_{i=M_n-\delta_n}^{\min\{M_n+\eta_n,k_n\}} a_i^{(n-p)} \leqslant \frac{\texttt{Lab}_{n,k_n}}{2^n} \leqslant (M_n+\eta_n)^p \sum_{i=M_n-\delta_n}^{\min\{M_n+\eta_n,k_n\}} a_i^{(n-p)}.$$

Since $k_n \geqslant M_n \geqslant M_{n-p}$, via similar arguments to those developed for the proof of Lemma 2 *(ii)*, we get

$$\sum_{i=M_n-\delta_n}^{\min\{M_n+\eta_n,k_n\}} a_i^{(n-p)} \sim \sum_{i=M_n-\delta_n}^{\min\{M_{n-p}+\eta_n,k_n\}} a_i^{(n-p)}.$$

We thus have to compare

$$S_n = \sum_{i=M_n-\delta_n}^{\min\{M_{n-p}+\eta_n,k_n\}} a_i^{(n-p)}$$

and

$$T_n = \sum_{i=M_{n-p}-\delta_n}^{\min\{M_{n-p}+\eta_n,k_n\}} a_i^{(n-p)},$$

and to prove that those two sums are equivalent when n tends to infinity. Decompose S_n as follows:

$$S_n = T_n + \sum_{i=\min\{M_{n-p}+\eta_n,k_n\}}^{\min\{M_n+\eta_n,k_n\}} a_i^{(n-p)} - \sum_{i=M_{n-p}-\delta_n}^{M_n-\delta_n} a_i^{(n-p)}.$$

Arguments from the proof of Lemma 2 *(ii)* imply that the second summand is negligible in front of the first. Let us assume that the third term is non-zero, i.e. $M_n - \delta_n > M_{n-p} - \delta_n$ (note that if this term is zero then $S_n \sim T_n$ is already proved). Via Lemma 1, since $\frac{M_n}{M_{n-p}} = 1 + o(\frac{1}{M_n})$, we have

$$\sum_{i=M_{n-p}-\delta_n}^{M_n-\delta_n} a_i^{(n-p)} \leqslant (M_n - \delta_n - M_{n-p} + \delta_n) a_{M_{n-p}-\delta_n}^{(n-p)} = o(1)\ a_{M_{n-p}-\delta_n}^{(n-p)} = o\left(a_{M_{n-p}}^{(n-p)}\right),$$

in view of Lemma 2 *(i)*. Therefore, since $a_{M_{n-p}}^{(n-p)} \leqslant T_n$, we have $S_n \sim T_n$ when n tends to infinity, which implies, since $\eta_n = o(M_n)$ and $\delta_n = o(M_n)$,

$$\frac{\mathtt{Lab}_{n-p,k_n}}{\mathtt{Lab}_{n,k_n}} = (1 + o(1))\ \frac{1}{(2M_n)^p} = (1 + o(1))\ \left(\frac{\ln n}{2n}\right)^p. \qquad \square$$

Finally, this fundamental technical part allows us to use Kozik's key ideas in order to describe the probability distribution induced on Boolean functions, in our two new models.

5 Adjustment of Kozik's pattern language theory

In 2008, Kozik [14] introduced a quite effective way to study Boolean trees: he defined a notion of pattern that permits to easily classify and count large trees according to some constraints on their structures. Kozik applied this pattern theory to study the classical Catalan tree distribution. We recall the definitions of patterns, illustrate them on examples and then extend Kozik's paper results in order to use them in our new models. This part will extensively use Analytic Combinatorics (generating functions, symbolic methods, singularity analysis): we refer the reader to Flajolet & Sedgewick's book [7] for an introduction to these methods.

Definition 6. *(i) A* **pattern** *is a binary tree with internal nodes labelled by $\wedge$ or $\vee$ and with external nodes labelled by $\bullet$ or $\Box$. Leaves labelled by $\bullet$ are called* **pattern leaves** *and leaves labelled by $\Box$ are called* **place-holders**. *A* **pattern language** *is a set of patterns.*

(ii) Given a pattern language L and a family of trees $\mathcal{M}$, we denote by $L[\mathcal{M}]$ the family of all trees obtained by replacing every place-holder in an element from L by a tree from $\mathcal{M}$.

(iii) We say that L is **unambiguous** *if, and only if, for any family $\mathcal{M}$ of trees, any tree of $L[\mathcal{M}]$ can be built from a unique pattern from L into which trees from $\mathcal{M}$ have been plugged.*

The generating function of a pattern language L is $\ell(x, y) = \sum_{d,p} L(d, p)x^d y^p$, where $L(d, p)$ is the number of elements of L with d pattern leaves and p place-holders.

Definition 7. *We define the* **composition** *of two pattern languages $L[P]$ to be the pattern language of trees which are obtained by replacing every place-holder of a tree from L by a tree from P.*

Given an integer i and a pattern L, the pattern $L^{(i)}$ is defined by the following recursion: $L^{(1)} = L$ and $L^{(i+1)} = L^{(i)}[L]$.

Definition 8. *A pattern language L is* **sub-critical** *for a family $\mathcal{M}$ if the generating function $m(z)$ of $\mathcal{M}$ has a square-root singularity τ, and if $\ell(x, y)$ is analytic in some set $\{(x, y) : |x| \leqslant \tau + \varepsilon, |y| \leqslant m(\tau) + \varepsilon\}$ for some positive ε.*

Definition 9. *Let L be a unambiguous pattern language, $\mathcal{M}$ be a family of trees and Γ a subset of $\{x_i\}_{i\geqslant 1}$, which cardinality does not depend on n. Given an element of $L[\mathcal{M}]$,*

*(i) the number of its L-***repetitions** *is the number of its L-pattern leaves minus the number of different variables that appear in the labelling of its L-pattern leaves.*

*(ii) the number of its (L, Γ)-***restrictions** *is the number of its L-pattern leaves that are labelled by variables from Γ, plus the number of its L-repetitions.*

Definition 10. *Let $\mathcal{I}$ be the family of the trees with internal nodes labelled by a connective and leaves without labelling, i.e. the family of tree-structures.*

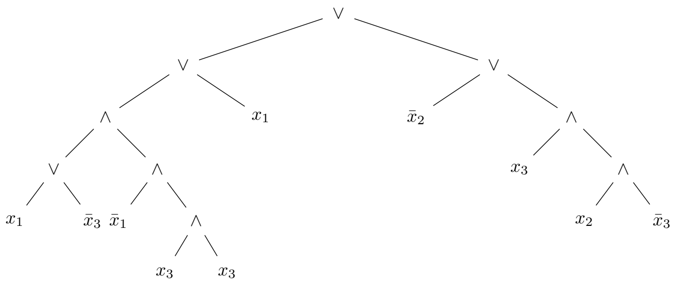

Figure 2: The tree computes the function $x_1 \vee \neg x_2$.

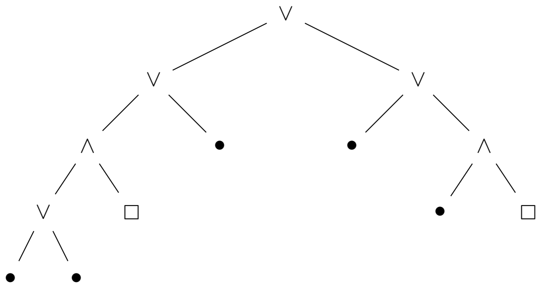

Figure 3: The pattern is an element of the pattern language N.

The generating function of $\mathcal{I}$ satisfies $I(z) = z + 2I(z)^2$, that implies $I(z) = (1 - \sqrt{1-8z})/4$ and thus its dominant singularity is $1/8$. Let I_n be the n-th coefficient of $I(z)$.

We can, for example, define the unambiguous pattern language N by induction as follows: $N = \bullet | N \vee N | N \wedge \Box$, meaning that a pattern from N is either a single pattern leaf, or a tree rooted by $\vee$ which two sub-trees are patterns from N, or a tree rooted by $\wedge$ which left sub-tree is a pattern from N and which right sub-tree is a place-holder. An element of N is represented in Fig 3. Its generating function verifies $n(x,y) = x + n(x,y)^2 + yn(x,y)$ and is equal to $n(x,y) = \frac{1}{2}(1 - y - \sqrt{(1-y)^2 - 4x})$. It is thus sub-critical for $\mathcal{I}$.

The tree depicted in Fig. 2 is built from the pattern of Fig. 3. It has 5 N-pattern leaves, 2 N-repetitions and 4 $(N, \{x_1, x_2\})$-restrictions. It is also built from the pattern of Fig. 4 and has 2 $N[N]$-pattern leaves, and 2 $(N[N], \{x_1, x_2\})$-restrictions.

The following key lemma is a generalization of the corresponding lemma of Kozik [14, Lemma 3.8].

Lemma 5. *Let L be an unambiguous pattern, sub-critical for the tree-structures family $\mathcal{I}$. Let r be a fixed positive integer.*

(G) *Let $A_n^{[r]}$ (resp. $A_n^{[\geqslant r]}$) be the number of labelled (with at most k_n variables) trees of $L[\mathcal{I}]$ of size n and with r L-repetitions (resp. at least r L-repetitions).*

(E) *Let $A_n^{[r]}$ (resp. $A_n^{[\geqslant r]}$) be the number of* equivalence classes *of labelled (with at most k_n variables) trees of $L[\mathcal{I}]$ of size n and with r L-repetitions (resp. at least r L-repetitions).*

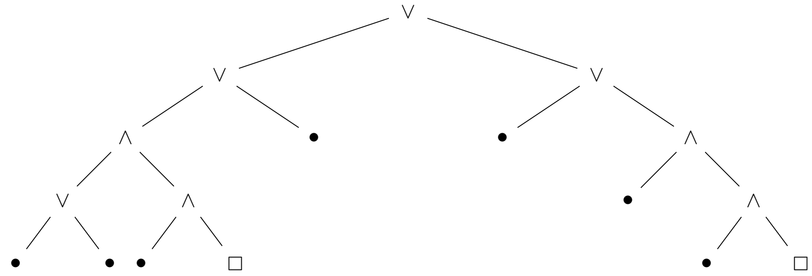

Figure 4: The pattern is an element of the pattern language $N[N]$.

Then, asymptotically when n tends to infinity, in both models,

$$\frac{A_n^{[r]}}{A_n} = \mathcal{O}\left(\mathtt{rat}_n^r\right) \qquad \text{and} \qquad \frac{A_n^{[\geqslant r]}}{A_n} = \mathcal{O}\left(\mathtt{rat}_n^r\right).$$

Proof. First recall that $A_n = I_n \cdot \mathtt{Lab}_{n,k_n}$ in both models.
Model (G). The number of labelled trees of $L[\mathcal{I}]$ of size n and with at least r L-repetitions is given by:

$$A_n^{[\geqslant r]} = \sum_{d=r+1}^{n} I_n(d) \cdot \mathtt{Lab}(n, k_n, d, r),$$

where $I_n(d)$ is the number of tree-structures with d L-pattern leaves (among the n number of leaves) and $\mathtt{Lab}(n, k_n, d, r)$ corresponds to the number of leaf-labellings of these trees giving at least r L-repetitions. The following enumeration contains some multi-counting and we therefore get an upper bound:

$$\mathtt{Lab}(n, k_n, d, r) \leqslant 2^n \cdot \sum_{j=1}^{r} \binom{d}{r+j} \left\{ {r+j \atop j} \right\} k_n(k_n-1)\cdots(k_n-j+1) k_n^{n-r-j}.$$

The factor 2^n corresponds to the polarity of each leaf (whether the literal is positive or negative); the index j stands for the number of different variables involved in the r repetitions; the binomial factor corresponds to the choices of the pattern leaves that are involved in the r repetitions; the Stirling number corresponds to the partition of the $r+j$ leaves into j parts; the factor $k_n(k_n-1)\cdots(k_n-j+1)$ stand for the choice of the repeated variables, from left to right; finally, the factor k_n^{n-r-j} corresponds to the choices of the variables assigned to all remaining leaves. We have

$$\mathtt{Lab}(n, k_n, d, r) \leqslant 2^n k_n^{n-r} \cdot \sum_{j=1}^{r} \binom{d}{r+j} \left\{ {r+j \atop j} \right\},$$

in other terms,

$$\mathtt{Lab}(n, k_n, d, r) \leqslant 2^r \mathtt{Lab}_{n-r,k_n} \cdot \sum_{j=1}^{r} \binom{d}{r+j} \left\{ {r+j \atop j} \right\},$$

since $\mathtt{Lab}_{n,m} = (2m)^n$ (in model (G)), and

$$A_n^{[\geqslant r]} \leqslant 2^r \cdot \mathtt{Lab}_{n-r,k_n} \sum_{j=1}^{r} \left\{ {r+j \atop j} \right\} \sum_{d=r+j}^{n} I_n(d) \binom{d}{r+j}. \tag{7}$$

Let $\ell(x, y)$ be the generating function of the pattern L. Note that $\frac{x^p}{p!}\partial_1^p \ell$ corresponds to pointing p distinct pattern leaves (without order) in the L-patterns (where ∂_1 stands for the derivative according to the first coordinate). Then, for all $p \geqslant 0$,

$$\frac{z^p}{p!}\partial_1^p \ell(z, I(z)) = \sum_{n=1}^{\infty} \sum_{d=p}^{\infty} I_n(d) \binom{d}{p} z^n.$$

Thus,

$$\frac{A_n^{[\geqslant r]}}{A_n} \leqslant \frac{2^r \mathtt{Lab}_{n-r,k_n}}{\mathtt{Lab}_{n,k_n}} \sum_{j=1}^{r} \left\{ {r+j \atop j} \right\} \frac{[z^n] z^{r+j} \partial_1^{r+j} \ell(z, I(z))}{[z^n] I(z)}.$$

Since $\partial_1^{r+j}\ell(z, I(z))$ and $I(z)$ have the same dominant singularity because of the sub-criticality of the pattern L according to $\mathcal{I}$, the previous sum tends to a constant (because r is fixed) when n tends to infinity and so we conclude, using Propositions 2 and 3:

$$\frac{A_n^{[r]}}{A_n} \leqslant \frac{A_n^{[\geqslant r]}}{A_n} = \mathcal{O}\left(\frac{\mathtt{Lab}_{n-r,k_n}}{\mathtt{Lab}_{n,k_n}}\right) = \mathcal{O}\left(\mathtt{rat}_n^r\right).$$

Model (E). The number of equivalence classes of labelled trees of $L[\mathcal{I}]$ of size n and with at least r L-repetitions is given by:

$$A_n^{[\geqslant r]} = \sum_{d=r+1}^{n} I_n(d) \cdot \mathtt{Lab}(n, k_n, d, r),$$

where $I_n(d)$ is the number of tree-structures with d L-pattern leaves and $\mathtt{Lab}(n, k_n, d, r)$ corresponds to the number of leaf-labellings of these trees giving at least r L-repetitions. The following enumeration contains some multi-counting and we therefore get an upper bound:

$$\mathtt{Lab}(n, k_n, d, r) \leqslant 2^n \cdot \sum_{j=1}^{r} \binom{d}{r+j} \left\{ {r+j \atop j} \right\} \frac{\mathtt{Lab}_{n-r,k_n}}{2^{n-r}}.$$

The factor 2^n corresponds to the polarity of each leaf (whether the literal is positive or negative); the index j stands for the number of different variables involved in the r repetitions; the binomial factor corresponds to the choices of the pattern leaves that are involved in the r repetitions; the Stirling number corresponds to the partition of $r + j$ leaves into j parts; finally, the factor $\mathtt{Lab}_{n-r,k_n}$ corresponds to the rest of the partition. Therefore,

$$A_n^{[\geqslant r]} \leqslant 2^r \cdot \mathtt{Lab}_{n-r,k_n} \sum_{j=1}^{r} \left\{ {r+j \atop j} \right\} \sum_{d=r+j}^{n} I_n(d) \binom{d}{r+j}.$$

Applying the same reasoning as for model (G) starting from Equation (7) permits to conclude the proof. □

We have finally adapted Kozik's theory in order to apply it in the new contexts. Since we have extended the pattern theory, we are able to use in the following the same key-ideas to describe the probability distributions we are interested in.

6 Behaviour of the probability distribution

Once we have adapted the pattern theory to our model and proved the central Lemma 5, we are ready to prove our main results, namely Theorems 1 and 2. A first step consists to understand the asymptotic behaviour of $\mathbb{P}_n^{(\mathtt{G})}(\mathsf{true})$ and $\mathbb{P}_n^{(\mathtt{E})}\langle\mathsf{true}\rangle$.

It is natural to focus on this "simple" function before considering a general class $\langle f \rangle$; and it happens to be essential for the continuation of the study. In addition, the methods used to study tautologies (mainly pattern theory) will also be the core of the proof for a general function (model (G)) or a general equivalence class (model (E)).

First, let us introduce some measure in the context of Boolean expressions. Given a family $\mathcal{G}$ of and/or trees (resp. equivalence classes of and/or trees), we define its **ratio** $\mu_n(\mathcal{G})$ as follows: let G_n be the number of elements of $\mathcal{G}$ of size n,

$$\mu_n(\mathcal{G}) := \frac{G_n}{A_n}.$$

6.1 Tautologies

First note that true is the unique element of its equivalence class $\langle\mathsf{true}\rangle$.

A **tautology** is an and/or tree that represents the Boolean function true. By symmetry, the functions true and false have the same probability in both models. Let $\mathcal{T}$ be the family of tautologies. In this part, we prove that the probability of true is asymptotically equal to the ratio of a simple subset of tautologies.

Definition 11 (cf. Fig. 5)**.** *A* **simple tautology** *is an* and/or *tree that contains two leaves labelled by a variable x and its negation $\bar{x}$ and such that all internal nodes from the root to both these leaves are labelled by $\vee$-connectives. We denote by $\mathcal{S}$ the family of simple tautologies.*

Proposition 5. *The ratio of simple tautologies verifies*

$$\mu_n(\mathcal{S}) \sim \frac{3}{2} \cdot \mathsf{rat}_n, \ \textit{when } n \textit{ tends to infinity.}$$

Moreover, asymptotically when n tends to infinity, almost all tautologies are simple tautologies, meaning that

$$\mu_n(\mathcal{T}) \sim \mu_n(\mathcal{S}), \ \textit{when } n \textit{ tends to infinity.}$$

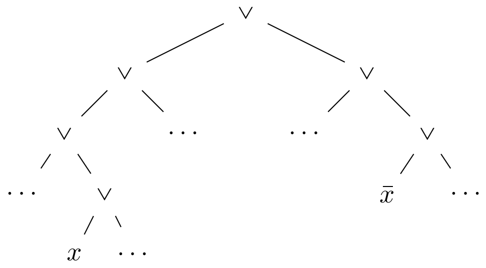

Figure 5: A simple tautology.

Proof. The proof is divided in two steps. The first one is dedicated to the computation of the ratio $\mu_n(\mathcal{S})$. The second part of the proof shows that almost all tautologies are simple tautologies.

Let us consider the non-ambiguous pattern language $M = \bullet | M \vee M | \square \wedge \square$. Remark that a tree such that two M-pattern leaves are labelled by a variable and its negation, is a simple tautology. The generating function of M is $m(x, y) = \frac{1}{2}(1 - \sqrt{1 - 4(x + y^2)})$. It is sub-critical for $\mathcal{I}$.

The generating function $\tilde{I}(z) = \frac{1}{2}{}^{\partial^2}\!/_{\partial x^2}(m(xz, I(z))_{|x=1}$ enumerates and/or trees with two marked distinct leaves linked to the root by or-nodes. Therefore, $DC_n = \tilde{I}_n \cdot \mathtt{Lab}_{n-1,k_n}$ is the number of simple tautologies where simple tautologies realized by a unique pair of leaves are counted once, those that are realized by two pairs of leaves are counted twice, and so on. We have

$$\frac{DC_n}{A_n} = \frac{\tilde{I}_n \cdot \mathtt{Lab}_{n-1,k_n}}{I_n \cdot \mathtt{Lab}_{n,k_n}},$$

and using a consequence of [7, Theorem VII.8] (cf. a detailed proof in [11]):

$$\lim_{n\to\infty} \frac{\tilde{I}_n}{I_n} = \lim_{z \to \frac{1}{8}} \frac{\tilde{I}'(z)}{I'(z)}.$$

Note that

$$\tilde{I}(z) = \frac{z^2}{(1 - 4(z + I(z)^2))^{3/2}},$$

and thus,

$$\frac{\tilde{I}'(z)}{I(z)} = \frac{2z}{(1 - 4(z + I(z)^2))^{3/2}} + \frac{(1 + 2I'(z)I(z))}{I'(z)} \; \frac{6z^2}{(1 - 4(z + I(z)^2))^{5/2}}.$$

Note that, when $z \to 1/8$, $I'(z) \to +\infty$. Moreover, $I(1/8) = 1/4$. Thus,

$$\frac{\tilde{I}'(z)}{I(z)} \sim \frac{{}^{3}\!/_{8^2}}{(1 - 4(1/8 + 1/16))^{5/2}} = \frac{3}{2} \quad \text{when } z \to \frac{1}{8}.$$

Thus, we get the upper bound $3/2 \cdot \mathsf{rat}_n$ for the ratio of simple tautologies: it remains to deal with the double-counting in order to compute a lower bound.

In DC_n, simple tautologies realized by a unique pair of leaves are counted once, those that are realized by two pairs of leaves are counted twice, and so on. Let us denote by ST_n^i the number of simple tautologies counted at least i times in DC_n: we have $DC_n = \sum_{i \geqslant 1} ST_n^{(i)}$.

Our aim is to remove from DC_n the tautologies that have been over-counted. Therefore, we count simple tautologies realized by three M-pattern leaves labelled by $\alpha/\alpha/\bar{\alpha}$ where α is a literal, and the tautologies realized by four M-pattern leaves labelled by $\alpha/\bar{\alpha}/\beta/\bar{\beta}$ where α and β are two different literals. Let us denote by

$$I_3(z) = \frac{1}{3!} \frac{\partial^3}{\partial x^3} m(xz, I(z))_{|x=1}$$

the generating function of tree-structures in which three M-pattern leaves have been pointed and

$$I_4(z) = \frac{1}{4!} \frac{\partial^4}{\partial x^4} m(xz, I(z))_{|x=1}$$

the generating function of tree-structures in which four M-pattern leaves have been pointed. Then, let

$$DC_n^{(3)} = 3 \cdot \mathtt{Lab}_{n-2,k_n}[z^n]I_3(z) \quad \text{and} \quad DC_n^{(4)} = 3 \cdot \mathtt{Lab}_{n-2,k_n}[z^n]I_4(z).$$

The integer $DC_n^{(3)}$ (resp. $DC_n^{(4)}$) counts (possibly with multiplicity) the trees in which three (resp. four) M-pattern leaves have been pointed, one of them labelled by a literal and the two others by its negation (resp. two of them labelled by two literals associated to two different variables and the two others by their negations). Remark that a tree having six M-pattern leaves labelled by $\alpha/\alpha/\bar{\alpha}/\beta/\beta/\bar{\beta}$ is counted twice by $DC_n^{(3)}$ and four times by $DC_n^{(4)}$.

For all integer i, a simple tautology counted at least i times by DC_n is counted at least $(i-1)$ times by $DC_n^{(3)} + DC_n^{(4)}$. Therefore,

$$ST_n \geqslant DC_n - (DC_n^{(3)} + DC_n^{(4)}).$$

In view of Lemma 5,

$$\frac{DC_n^{(3)}}{T_n} \leqslant c_3 \cdot \frac{\mathtt{Lab}_{n-2,k_n}}{\mathtt{Lab}_{n,k_n}} \quad \text{and} \quad \frac{DC_n^{(4)}}{T_n} \leqslant c_4 \cdot \frac{\mathtt{Lab}_{n-2,k_n}}{\mathtt{Lab}_{n,k_n}},$$

where c_3 and c_4 are positive constants. Then, asymptotically when n tends to infinity, in view of Propositions 2 and 3: $\mu_n(\mathcal{F}) = \mu_n(DC) + o\left(\mathsf{rat}_n\right) \sim 3/2 \cdot \mathsf{rat}_n$.

Let us now turn to the second part of the proof: asymptotically, almost all tautologies are simple tautologies. Let us consider the pattern $N = \bullet|N \vee N|N \wedge \square$. This pattern is unambiguous, its generating function satisfies $n(x,y) = x + n(x,y)^2 + y \cdot n(x,y)$ and is thus equal to $\frac{1}{2}(1 - y - \sqrt{(1-y)^2 - 4x})$. Consequently, N is sub-critical for the family $\mathcal{I}$ of tree-structures.

A tautology has at least one $N[N]$-repetition. Otherwise, we can assign all its N-pattern leaves to false and, the whole tree computes false: impossible for a tautology.

Consider a tautology t with exactly one $N[N]$-repetition. this repetition must be a $x|\bar{x}$ repetition and must occur among the N-pattern leaves, using the same kind of argument than above.

Then, let us assume that there is an $\wedge$-node denoted by ν between the N-pattern leaf x and the root of the tree. This node ν has a left sub-tree t_1 and a right sub-tree t_2. Necessarily the leaf x appears in t_1. Then, one can assign all the N-pattern leaves of t_2 (which are $N[N]$-pattern leaves of t) to false, since there is no more repetition among the $N[N]$-pattern leaves of t. Also assign all the $N[N]$-pattern leaves of t minus the sub-tree rooted at ν to false. Then, we can see that t computes false: impossible. We have thus shown that t is a simple tautology.

In a nutshell, tautologies with exactly one $N[N]$-repetition are simple tautologies, a tautology must have at least one $N[N]$-repetition and, thanks to Lemma 5, tautologies with more than one $N[N]$-repetitions have a ratio of order $o\left(\mathsf{rat}_n\right)$, which is negligible in front of the ratio of simple tautologies. □

The latter proposition gives us for free the proof for the satisfiability problem. In fact, both dualities between the two connectives and positive and negative literals transform expressions computing true to expressions computing false, which implies $\mathbb{P}_n^{(\mathtt{G})}(\mathsf{false}) = 3/2 \cdot \mathsf{rat}_n$. Moreover, the only expressions that are not satisfiable compute the function false and $\mathbb{P}_n^{(\mathtt{G})}(\mathsf{false}) = 3/2 \cdot \mathsf{rat}_n$ tends to 0 as n tends to infinity, which proves Corollary 1.

6.2 Proofs of Theorems 1 and 2

This last section is devoted to the general result, i.e. to the study of the behaviour of $\mathbb{P}_n^{(\mathtt{G})}(f)$ and $\mathbb{P}_n^{(\mathtt{E})}\langle f\rangle$ for all non constant Boolean function f. The main idea of this part is that, roughly speaking, *a typical tree computing a Boolean function f is a minimal tree of f into which a single large tree has been plugged.*

In the following, f (resp. $\langle f\rangle$) is fixed,we denote by $r = L(f)$ its complexity, and by Γ_f the set of the essential variables of f. We also fix t to be an and/or tree computing f.

Moreover, we will need the folowing patterns:

$$N = \bullet|N \vee N|N \wedge \square,$$

$$P = \bullet|P \vee \square|P \wedge P,$$

and (see Definition 7 where the composition of patterns is defined)

$$R = N^{(r+1)}[N \oplus P] \quad \text{and} \quad \bar{R} = N^{(r+1)}[(N \oplus P)^2],$$

where the language $N \oplus P$ is defined such that the $N \oplus P$-pattern leaves of a tree are its N-pattern leaves plus its P pattern leaves. It is proved in [14] that this pattern language is indeed non-ambiguous and sub-critical for $\mathcal{I}$ if N and P are non-ambiguous and sub-critical for $\mathcal{I}$.

We have already noticed that assigning all N-pattern leaves of a Boolean tree to false make the whole tree calculate false. The pattern P has the dual property that: assigning all the P-patterns leaves of a tree to true make the whole tree calculate true. This is why these two patterns are so useful in the proof of our main result.

Proposition 6. *A tree t computing f (define $r := L(f)$) with at least one leaf on the $(r+2)^{th}$ level of the R-pattern must have at least $r+1$ (R, Γ_f)-restrictions.*

Proof. Let us assume that t computes f, and has at least one leaf on the $(r+2)^{\text{th}}$ level of the R pattern but has less than r R-repetitions. Let i be the smallest integer (smaller than $r+2$) such that the number of $(N^{(i)}, \Gamma_f)$-restrictions is equal to the number of $(N^{(i-1)}, \Gamma_f)$-restrictions.

There must be either a repetition or an essential variable in the first level: if there is none, then we can assign all the N pattern leaves to false and this operation does not changes the represented function. This function is then the constant function false, which is impossible; so $i \leqslant r+1$.

First case: Let us assume that there are strictly less than r $(N^{(i)}, \Gamma_f)$-restrictions. There is no repetition and no essential variable in the pattern leaves at level i. Therefore, we can assign them all to false and make the place-holders of the level $i-1$ compute false. Let us replace those place-holders by false in the tree. Furthermore, replace by false all the non-essential remaining variables. And simplify the obtained tree to simplify all the constant leaves false and true. We obtain a tree $t^\star$, which still computes f, and whose leaves are all former $N^{(i-1)}$ pattern leaves of t labelled by essential variables. The tree $t^\star$ therefore contains strictly less than r leaves, which is impossible since the complexity of f is r.

Second Case: Let us assume that t has exactly r $(N^{(i)}, \Gamma_f)$-restrictions. Since $i \leqslant r+1$, there is no restriction in the place-holders of the level $r+2$. Therefore, we can replace the place-holders by wild-cards $\star$, which means that those wild-cards can be evaluated to true or false independently from each other and without changing the function computed by t. We can also replace the remaining leaves labelled by non-essential and non-repeated variables by such wild-cards.

We simplify those wild-cards. Such a simplification has to delete at least one non-wild-card leaf. If we deleted a non-repeated essential variable, then the tree $t^\star$ does not depend on this essential variable and computes f: this is impossible. Thus, we deleted a repetition: $t^\star$ has strictly less than $R(f)$ repetitions and computes f. It is impossible. □ □

Remark that in Lemma 5, we only count repetitions and not restrictions as it was done in the original lemma by Kozik. Though, we will need to consider essential variables and the following lemma permits to handle them. An **expansion** of a tree t is a tree obtained by replacing a sub-tree s of t by $s \diamond t_e$ (or $t_e \diamond s$) where $\diamond \in \{\wedge, \vee\}$.

Lemma 6. *Let L be an unambiguous pattern, sub-critical for $\mathcal{I}$. Let f be a fixed Boolean function, Γ_f the set of its essential variables, and $\mathcal{M}_f$ the set of minimal trees computing f. Let $\mathcal{E}$ be the family of trees obtained by expanding once a tree of $\mathcal{M}_f$ by trees having exactly p (L, Γ_f)-restrictions. Then, there exists a constant $\alpha^{(\mathtt{G})} > 0$ (resp. $\alpha^{(\mathtt{E})} > 0$) such that*

$$\mu_n(\mathcal{E}) \sim \alpha^{(\mathtt{G})} \cdot \mathsf{rat}_n^{L(f)+p} \text{ in model } (\mathtt{G}),$$

resp.

$$\mu_n(\mathcal{E}) \sim \alpha^{(\mathtt{E})} \cdot \mathsf{rat}_n^{R\langle f \rangle + p} \text{ in model } (\mathtt{E}).$$

Proof. Let E_n be the number of (resp. equivalence classes of) trees of size n in $\mathcal{E}$. We will denote by i the number of leaves that are involved in the p (L, Γ_f)-restrictions of the expansion tree: $p+1 \leqslant i \leqslant 2p$. Let γ_f be the cardinal of Γ_f.

In the model $(\mathtt{G})$, for all large enough n,

$$\mu_n(\mathcal{E}) = \frac{E_n}{A_n} \leqslant \mathtt{cst}_f \sum_{i=p+1}^{2p} [z^{n-L(f)}] \frac{\partial^i}{i! \partial x^i} \left(\ell(xz, I(z))\right)_{|x=1} \frac{(2\gamma_f)^p (2(k_n - \gamma_f))^{n-L(f)-p}}{I_n (2k_n)^n},$$

where $\mathtt{cst}_f = 2L(f) \cdot |\mathcal{M}_f|$ is an upper bound for the different places in a minimal tree of f where an expansion can be plugged in. Since L is sub-critical for $\mathcal{I}$, there exists a positive constant α such that

$$\sum_{i=p+1}^{2p} \frac{[z^{n-L(f)}] \partial^i / i! \partial x^i \left(\ell(xz, I(z))\right)_{|x=1}}{I_n} \sim \alpha \cdot \frac{I_{n-L(f)}}{I_n} \sim \alpha \left(\frac{1}{8}\right)^{L(f)} > 0$$

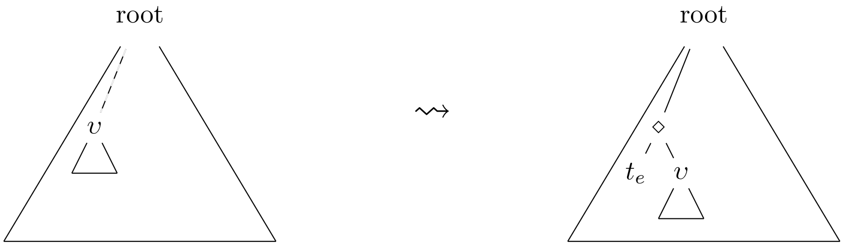

Figure 6: An expansion at node υ. Note that the expansion tree t_e could have been on the right size of the $\diamond$-connective instead of its left side.

asymptotically when n tends to infinity. Therefore, in view of Section 4, we have

$$\mu_n(\mathcal{E}) \sim \alpha \cdot \mathsf{rat}_n^{L(f)+p}.$$

In the model (E), we have, with the same reasoning:

$$\mu_n\langle\mathcal{E}\rangle = \frac{E_n}{A_n} \leqslant \mathtt{cst}_f \sum_{i=p+1}^{2p} [z^{n-L(f)}] \frac{\partial^i}{i!\partial x^i} \left(\ell(xz, I(z))\right)_{|x=1} \frac{2^{p+R\langle f\rangle} \cdot \mathtt{Lab}_{n-p-R\langle f\rangle,k_n}}{I_n \cdot \mathtt{Lab}_{n,k_n}},$$

from which we state the same conclusion as for the model (G). □

Consider the family $\mathcal{E}$ of trees obtained by replacing a sub-tree s by $s \wedge t_e$ where t_e is a simple tautology into a minimal tree of f. Since a simple tautology has at least one S-repetition, thanks to Lemma 6, there exists two positive constants $\alpha^{(\mathtt{G})}$ and $\alpha^{(\mathtt{E})}$ such that

$$\mu_n^{(\mathtt{G})}(\mathcal{E}) \sim \alpha^{(\mathtt{G})} \cdot \mathsf{rat}_n^{L(f)+1} \text{ in model } (\mathtt{G}),$$

and

$$\mu_n^{(\mathtt{E})}\langle\mathcal{E}\rangle \sim \alpha^{(\mathtt{E})} \cdot \mathsf{rat}_n^{R\langle f\rangle+1} \text{ in model } (\mathtt{E}).$$

Thanks to Lemma 5, we know that terms computing f with more than $R(f) + 2$ repetitions are negligible in front of the above family. Therefore, since trees with no leaf on the $(r + 2)^{\text{th}}$ level are negligible, we have proved weaker versions of Theorems 1 and 2, where the equivalent for the probabilities is replaced by an upper and a lower bounds of the same order. The rest of the proofs consists in sharpening both bounds.

The key point of the proof of Theorems 1 and 2 is that a typical tree computing a function f is a minimal tree of this function which has been expanded once. In the following, we will only consider two different expansions:

Definition 12 (cf. Figure 6)**.** *Recall that an* **expansion** *of a tree t is a tree obtained by replacing a sub-tree s of t by $s \diamond t_e$ (or $t_e \diamond s$) where $\diamond \in \{\wedge, \vee\}$.*

An expansion is a **T-expansion** *if the expansion tree t_e is a simple tautology and the connective $\diamond$ is $\wedge$ (or a simple contradiction and the connective $\diamond$ is $\vee$).*

An expansion is a **X-expansion** *if the expansion tree t_e has a leaf linked to the root by a $\wedge$-path (resp. a $\vee$-path) and the $\diamond$ connective is a $\vee$ (resp. $\wedge$).*

Corollary 2. *The ratio of the (resp. equivalence class of) minimal trees of f expanded once satisfies that there exists two positive constant λ_f and $\lambda_{\langle f\rangle}$ such that asymptotically when n tends to infinity:*

$$\mu_n^{(\mathtt{G})}(E[\mathcal{M}_f]) = \lambda_f \cdot \mathsf{rat}_n^{L(f)+1} + o\left(\mathsf{rat}_n^{L(f)+1}\right),$$

$$\mu_n^{(\mathtt{E})}\langle E[\mathcal{M}_f]\rangle = \lambda_{\langle f\rangle} \cdot \mathsf{rat}_n^{R\langle f\rangle+1} + o\left(\mathsf{rat}_n^{R\langle f\rangle+1}\right).$$

This corollary is a direct consequence of Lemma 6.

Lemma 7. *Let f be a fixed Boolean function and $\mathcal{M}_f$ the set of minimal trees of f.*

$$\mathbb{P}_n^{(\mathtt{G})}(f) \sim \mu_n^{(\mathtt{G})}(E[\mathcal{M}_f]) \textit{ when } n \to +\infty,$$

and

$$\mathbb{P}_n^{(\mathtt{E})}\langle f\rangle \sim \mu_n^{(\mathtt{E})}\langle E[\mathcal{M}_f]\rangle \textit{ when } n \to +\infty.$$

Proof. Let t be a tree computing f. Such a tree must have at least $R(f)+1$ $\bar{R}$-repetitions. Moreover, thanks to Lemma 5, trees with at least $R(f)+2$ $\bar{R}$-repetitions are negligible. We will show that a tree with exactly $R(f)+1$ $\bar{R}$-repetitions is in fact a minimal tree expanded once.

The term t must also have $R(f)+1$ R-repetitions and therefore, there is no additional repetition when we consider the $(r+3)^{\text{th}}$ level of the $\bar{R}$-pattern.

Let i be the first level such that the number of $(N^{(i)},\Gamma_f)$-restrictions is equal to the number of $N^{(i-1)}$-restrictions. Since there must be a restriction on the first level, $i \leqslant r+1$.

First Case: Assume that an essential variable α appears on the pattern leaves of the $(r+3)^{\text{th}}$ level. Therefore, t has at most $L(f)$ $(N^{(i)},\Gamma_f)$-restrictions. Let us replace the place-holders of the $(i-1)^{\text{th}}$ level by false and assign all the remaining non-essential variables to false. Simplify the tree to obtain a new and/or tree denoted by $t^\star$. The leaves of this tree are former $N^{(i-1)}$-pattern leaves of t, labelled by essential variables and $t^\star$ still computes f. But the variable α is essential for f: thus it must still appear in the leaves of $t^\star$, and by deleting its occurrence in the leaves of the $(r+3)^{\text{th}}$ level, we deleted one repetition. Therefore, $t^\star$ has at most $L(f)-1$ leaves which is impossible!

Second Case: There is no essential variable among the the pattern leaves of the $(r+3)^{\text{th}}$ level. Since there is also no repetition at this level, we can replace the place-holders of the level $(r+3)$ to wild-cards. We also replace the remaining non essential and non-repeated variables by wild-cards. We then simplify the wild-cards and obtained a simplified tree $t^\star$, computing f, with no wild-cards and which leaves are former leaves of the trees t, essential or repeated. During the simplification process, we have deleted at least one of these leaves and therefore $t^\star$ has at most $L(f)$ leaves: it is a minimal tree of f.

Let us consider the following fact: The lowest common ancestor of all the wild-cards in t has been suppressed during the simplification process. Assume that this fact is false: then two wild-cards have been simplified independently during the simplification process, and thus, at least two essential or repeated variables have been deleted. The tree $t^\star$ has thus at most $L(f)-1$ leaves and computes f, which is impossible since $L(f)$ is the complexity of f. Let us denote by t_e the sub-tree rooted at v the lowest common ancestor of the wild-cards. Thus a typical tree computing f is a minimal tree of f in which we have plugged a specific expansion tree t_e. □

Lemma 8. *Let t be a typical tree computing f. The expansion tree t_e is either a simple tautology (or simple contradiction), or an x-expansion - i.e. a tree with one $\wedge$-leaf (resp. $\vee$-leaf) labelled by an essential variable of f.*

Proof. As shown in the former lemma, a typical tree computing f is a minimal tree of f on which has been plugged an expansion tree t_e.

First Case: Let us assume that t_e has no $(N\oplus P)$-repetition and no essential variable among its $(N\oplus P)$-pattern leaves. Then, we can replace t_e by a wild-card and simplify this wild-card. This simplification suppresses at least one other leaf of the tree: the obtained tree is then smaller than the original minimal tree, and still computes f. It is impossible.

Second Case: Let us assume that t_e has at least two $((N\oplus P)^2,\Gamma_f)$-restrictions. Thanks to Lemma 6, this family of expanded trees is negligible.

Third Case: Let us assume that t_e has exactly one $((N\oplus P)^2,\Gamma_f)$-restriction. Then it must be a $(N\oplus P,\Gamma_f)$-restriction (see First Case).

- if it is a repetition, than one can show that it must be a simple tautology or a simple contradiction.
- if it is an essential variable, one can show that it must be an X-expansion.

□ □

7 Conclusion

In this paper, we have generalised the Catalan tree distribution on Boolean functions following two directions:

- letting the number of variables and the size of the Boolean trees tend to infinity together. It has allowed us to answer a fundamental satisfiability problem;
- the natural equivalence relation on Boolean trees and functions that we have introduced exhibits a very interesting threshold/saturation phenomenon for which we have no intuitive explanation up to now.

It is interesting to see that these two models can be analysed with very similar methods, namely, the ones used in the literature to study the classical Catalan tree model: Analytic Combinatorics and Kozik's pattern theory. The key idea that permitted to generalise those methods to our two new models was to dissociate the shapes of the trees and their leaf-labelling.

We strongly believe that our methods could be generalised further, for example to other logical systems (as the implication model, see e.g. [9, 11]), or to non-binary or non-planar uniform trees (see [10]). Our confidence rely on the fact that those models, in the $(k_n)_{n\geqslant 1}$ constant case, can be analysed with analytic combinatorics and pattern theory (or tools based on the same key ideas) as well, and we have shown here how to generalise those methods to a more general sequence $(k_n)_{n\geqslant 1}$.

A more challenging generalisation would be to consider different probability distributions on binary plane trees. For example, in view of [8, 3] we conjecture that the random binary search tree of size n, labelled with $(k_n)_{n\geqslant 1}$ variables defines a very interesting satisfiability problem, with a phase transition *à la* K–SAT. It would be very interesting (but, we expect, non trivial) to prove such a conjecture. Even more challenging would be to ask what effect the introduction of the equivalence relation has on this phase transition?

Acknowledgements: The authors are very grateful to Pierre Lescanne for fruitful discussions about this project, to Brigitte Chauvin and Danièle Gardy for proof-reading an early version of this manuscript and to the anonymous referees for their very insightful comments. The second author also wishes to thank EPSRC for support through the grant EP/K016075/1.

References

[1] Achlioptas, D., Moore, C.: Random k-SAT: Two moments suffice to cross a sharp threshold. SIAM Journal of Computing **36**(3), 740–762 (2006)

[2] Chauvin, B., Flajolet, P., Gardy, D., Gittenberger, B.: And/Or trees revisited. Combinatorics, Probability and Computing **13**(4–5), 475–497 (2004)

[3] Chauvin, B., Gardy, D., Mailler, C.: A sprouting tree model for random boolean functions. Random Structures and Algorithms (2014). DOI 10.1002/rsa.20567. (to appear)

[4] Coja-Oghlan, A.: The asymptotic k-SAT threshold. In: 46th Symposium on Theory of Computing, STOC, pp. 804–813 (2014)

[5] Coja-Oghlan, A., Panagiotou, K.: Going after the k-SAT threshold. In: 45th Symposium on Theory of Computing, STOC, pp. 705–714 (2013)

[6] Daudé, H., Ravelomanana, V.: Random 2-XORSAT phase transition. Algorithmica **59**(1), 48–65 (2011)

[7] Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge U.P. (2009)

[8] Fournier, H., Gardy, D., Genitrini, A.: Balanced And/Or trees and linear threshold functions. In: 6th SIAM Workshop on Analytic Algorithmics and Combinatorics (ANALCO), pp. 51–57. New York, USA (2009)

[9] Fournier, H., Gardy, D., Genitrini, A., Gittenberger, B.: The fraction of large random trees representing a given boolean function in implicational logic. Random Structures and Algorithms **40**(3), 317–349 (2012). DOI 10.1002/rsa.20379

[10] Genitrini, A., Gittenberger, B., Kraus, V., Mailler, C.: Associative and commutative tree representations for boolean functions. Theoretical Computer Science **570**, 70–101 (2015). DOI 10.1016/j.tcs.2014.12.025. URL `http://dx.doi.org/10.1016/j.tcs.2014.12.025`

[11] Genitrini, A., Kozik, J.: In the full propositional logic, 5/8 of classical tautologies are intuitionistically valid. Ann. of Pure and Applied Logic **163**(7), 875–887 (2012). DOI 10.1016/j.apal.2011.09.011

[12] Genitrini, A., Kozik, J., Zaionc, M.: Intuitionistic vs. classical tautologies, quantitative comparison. In: TYPES, pp. 100–109 (2007). DOI 10.1007/978-3-540-68103-8_7

[13] Genitrini, A., Mailler, C.: Equivalence classes of random boolean trees and application to the catalan satisfiability problem. In: Springer-Verlag (ed.) Latin American Theoretical INformatics, pp. 466–477. Motevideo, Uruguay (2014)

[14] Kozik, J.: Subcritical pattern languages for And/Or trees. In: Fifth Colloquium on Mathematics and Computer Science. DMTCS Proceedings (2008)

[15] Lefmann, H., Savický, P.: Some typical properties of large And/Or Boolean formulas. Random Structures and Algorithms **10**, 337–351 (1997)

[16] Lupanov, O.B.: A method of circuit synthesis. Izvesitya VUZ, Radiofiz **1**, 120–140 (1958). (in Russian)

[17] Lutz, J.H.: Almost everywhere high nonuniform complexity. Journal of Computer and System Sciences **44**(2), 220–258 (1992)

[18] Paris, J.B., Vencovská, A., Wilmers, G.M.: A natural prior probability distribution derived from the propositional calculus. Ann. of Pure and Applied Logic **70**, 243–285 (1994)

[19] Sibuya, M.: Log-concavity of Stirling numbers and unimodality of Stirling distributions. Ann. of the Institute of Statistical Mathematics **40**(4), 693–714 (1988)

[20] Woods, A.R.: Coloring rules for finite trees, and probabilities of monadic second order sequences. Random Structures and Algorithms **10**, 453–485 (1997)