Equivalence classes of random Boolean trees and application to the Catalan satisfiability problem*

Antoine Genitrini† and Cécile Mailler‡
Antoine.Genitrini@lip6.fr, Cecile.Mailler@uvsq.fr.

An and/or tree is a binary plane tree, with internal nodes labelled by connectives, and with leaves labelled by literals chosen in a fixed set of k variables and their negations. We introduce the first model of such Catalan trees, whose number of variables k_n is a function of n, its number of leaves. We describe the whole range of the probability distributions depending on the functions k_n, as soon as it tends jointly with n to infinity. As a by-product we obtain a study of the satisfiability problem in the context of Catalan trees.
Our study is mainly based on analytic combinatorics and extends the Kozik's *pattern theory*, first developed for the fixed-k Catalan tree model.
keywords: Random Boolean expressions; Boolean formulas; Boolean function; Probability distribution; Satisfiability; Analytic combinatorics.

1 Introduction

Since years, many scientists of different areas, e.g. computer scientists, mathematicians or statistical physicists, are studying satisfiability problems (like k–SAT problems) and some questions that arise around them: for example, phase transitions between satisfiable and unsatisfiable expressions or constraints satisfaction problems. The classical 3–SAT problem takes into consideration expressions of a specific form: conjunction of clauses that are themselves disjunctions of three literals. The literals are chosen among a set whose size is linked to the size of the expression. Then one question consists of deciding if a large random expression is satisfiable or not. Actually we know among other things, see [1] for example, that satisfiability is related to the ratio between the size of the expression and the number of allowed literals. There is a phase transition such that, if the ratio is smaller than a critical value, the random

*Partially supported by the A.N.R. project *BOOLE*, 09BLAN0011.
†Laboratoire d'Informatique de Paris 6, CNRS UMR 7606 and Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France.
‡Laboratoire de Mathématiques de Versailles; CNRS UMR 8100 and Université de Versailles Saint-Quentin-en-Yvelines, 45 avenue des États-Unis, 78035 Versailles, France.

expression is satisfiable with probability tending to 1, when the size of the expression tends to infinity. Otherwise, when the ratio is larger than the critical value, the probability tends to 0.

An interesting paper [3] about Boolean satisfiability problems deals with random 2–XORSAT expressions. Using generating functions, in the context of analytic combinatorics, the authors describe precisely the phase transition between satisfiable and unsatisfiable expressions.

Still dealing with Boolean expressions, but in a completely different direction, researchers have studied the complete probability distribution on Boolean functions induced by random Boolean expressions. The first approach, by Lefmann and Savický [12], consists in fixing a finite set of k variables, allowing the two logical connectives and and or and choosing uniformly at random a Boolean expression of *size* n in this logical system. Their model is usually called the *Catalan model*. Lefmann and Savický first proved the existence of a limiting probability distribution on Boolean functions when the size of the random Boolean expressions tends to infinity. Since the seminal paper by Chauvin et al. [2], almost all quantitative studies of such Boolean distributions are deeply related to analytic combinatorics: a survey by Gardy [6] provides a wide range of models with various numerical results. Later, Kozik [11] proved a strong relation between the limiting probability of a given function and its *complexity* (i.e. the minimal *size* of an expression representing the function). His approach lies in two separate steps: (i) first let the size of the Boolean expressions taken into consideration tend to infinity, and then (ii) let the number of variables used to label the expressions tend to infinity. His powerful machinery, the *pattern theory*, easily classifies and counts large expressions according to structural constraints. The main objection to this model is about the two consecutive limits that cannot be interchanged: in order to obtain quantitative results, a function must be fixed and thus we cannot consider functions whose complexity depends on n. Genitrini and Kozik have proposed another model [10, 9] that builds random Boolean expressions over an infinite set of variables. This approach avoids the bias induced by both successive limits. According to our knowledge, the single paper that relates the number of variables to the size is [7]: it finds a large family of functions of small complexity. However from this results we cannot derive any quantitative results of the probability of a small family of functions whose complexities depends on n. Moreover looking at satisfiability problems in this context seems to be very exciting.

Our paper extends the Catalan model in order to fit in the satisfiability context. By using an equivalence relation on Boolean expressions, we manage to let both the number of variables and the size of expressions tend jointly to infinity. The number of variables is a function of the size of the expressions and thus we deal with satisfiability in the context of Catalan expressions. Furthermore by extending the techniques of Kozik, we describe in details the probability distribution on functions and exhibit some threshold for the latter distribution: as soon as the number of variables is *large enough* compared to the size of the expressions, the general behaviour of the induced probability on Boolean functions does not change anymore by adding more variables.

The paper is organized as follows. Section 2 introduces our new model based on an equivalence relation of Boolean expressions. Then, Section 3 states our three main results: (1) the satisfiability question for random Catalan expressions; (2) the link between the probability of a class of functions and their common complexity; (3) the behaviour of the probability related to the dynamic between the number of variables and the size of the expressions. Section 4 is devoted to the technical core of the paper. Finally Section 5 applies our approach to and/or trees and proves the main results.

2 Probability distributions on equivalence classes of Boolean functions

2.1 Contextual definitions

A Boolean function is an mapping from $\{0,1\}^{\mathbb{N}}$ into $\{0,1\}$. The two constant functions $(x_i)_{i\geq 1} \mapsto 1$ and $(x_i)_{i\geq 1} \mapsto 0$ 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 Figure 1 is equivalent to the expression $([x_1 \vee (\bar{x}_1 \vee x_2)] \vee x_3) \vee (x_4 \wedge x_1)$ and represents the function f such that, for all $(x_i) \in \{0,1\}^{\mathbb{N}}$, $f((x_i)_{i\geq 1}) = ([x_1 \vee (\neg x_1 \vee x_2)] \vee x_3) \vee (x_4 \wedge x_1) = 1$, where $\neg x = 1 - x$ for all $x \in \{0,1\}$.

The **size** of an and/or tree is its number of leaves: remark that, for all $n \geq 1$, there is infinitely many and/or trees of size n.

The **complexity** of a non constant Boolean function f, denoted by $L(f)$, is defined as the size of its **minimal trees** , i.e. the smallest trees computing f. The complexity of true and false is 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*: 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: $E(f) \leq L(f) \leq 2^{E(f)+2}$ (see e.g. [4, p. 77–78] for the second inequality).

2.2 Equivalence relations

Analytic combinatorics (cf. [4]) is based on the notion of combinatorial classes. A *combinatorial class* is a denumerable (or finite) set of objects on which a size notion is defined such that each object has a non-negative size and the set of objects of any given size is finite. Thus our class of and/or trees is not a combinatorial class since there is infinitely many of trees of a given size. To use analytic combinatorics, we define an equivalence relation on Boolean trees. In the rest of the paper, we define a **tree-structure** to be an and/or tree whose leaf-labels have been removed (but internal nodes remain labelled).

Informally two trees are equivalent if the leaves of first one can be relabelled (and negated) without collision in order to obtain the second tree.

Definition 1. *Let A and B be two* and/or *trees. Trees A and B are* **equivalent** *if (1) their tree-structures are identical, if (2) two leaves are labelled by the same variable in A if and only of they are labelled by the same variable in B, and if (3) 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.

2.3 Probability distribution

Let $(k_n)_{n\geq 1}$ **be an increasing sequence that tends to infinity** when n tends to infinity. In the following, we only consider trees such that: for all $n \geq 1$, **the set of variables that appear as leaf-labels** (negated or not) **of a tree of size n has cardinality at most k_n**. Note that if $k_n \geq n$ for all $n \geq 1$, this hypothesis is not a restriction. Therefore, we will assume that $k_n \leq n$.

Definition 2. *We denote by t_n the number of equivalence classes of trees of size n in which at most k_n different variables appear as leaf-labels. We define the ordinary generating function $T(z)$ as $T(z) = \sum_n t_n z^n$.*

Proposition 3. *The number of classes of trees of size n satisfies:*

$$t_n = C_n \cdot \sum_{p=1}^{k_n} \left\{ {n \atop p} \right\} 2^{2n-1-p},$$

where C_n is the number of unlabelled binary trees[1] of size n and $\left\{ {n \atop p} \right\}$ is the Stirling number of the second kind[2].

Proof. Once the tree-structure of the binary tree is chosen (factor $2^{n-1}C_n$), we partition the set of leaves into p parts such that two leaves that belong to the same part are labelled by the same variable. It gives the contribution $\left\{ {n \atop p} \right\}$. Then, we choose to label each leaf by a positive or negative literal: contribution 2^n. The equivalence relation states that a tree and the one obtained from it by replacing the positive literals corresponding to a fixed variable by its negation (and conversely) are equivalent. Thus, for each class we multi-count the number of trees: correction 2^{-p}. □ □

Given a set $\mathcal{S}$ of equivalence classes of trees and S_n the number of elements of $\mathcal{S}$ of size n, we define the **ratio** of $\mathcal{S}$ by $\mu_n(\mathcal{S}) = S_n/t_n$. For a given Boolean function f, we denote by $T_n\langle f\rangle$ the number of equivalence classes of trees of size n that compute a function of $\langle f\rangle$, and we define the **probability** of $\langle f\rangle$ as the ratio of $T_n\langle f\rangle$:

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

One goal of this paper consists in studying the behaviour of the probabilities $(\mathbb{P}_n\langle f\rangle)_{f\in\mathcal{F}}$ when the size n of the trees tends to infinity.

3 Results

We state here our main result: the behaviour of $\mathbb{P}_n\langle f\rangle$ for all fixed functions $f \in \mathcal{F}$ in the framework of and/or trees. Note that f is a single function, not a family of functions. Neither f nor $\langle f\rangle$ depend on n, but we look at their representations with trees of size n over at most k_n variables.

The main idea of this part is that *a typical tree computing a Boolean function f is a minimal tree of f into which a large tree has been plugged, that does not change the function computed by the minimal tree.*

[1] In Proposition 3, C_n is the $(n-1)$st Catalan number (see e.g. [4, p. 6–7]).
[2] In Proposition 3, $\left\{ {n \atop p} \right\}$ is the number of partitions of n objects in p non-empty subsets (see e.g. [4, p. 735–737]).

This informal but fundamental idea is central in the recent results about quantitative logics (e.g. logic of implication [5]). The proofs in the distinct studies are different, only because of technical incidences induced by the connectives under study. Thus, we are convinced that our results hold in other logics.

Definition 4. *Let f be a Boolean functions. We denote by $L\langle f\rangle = L(f)$ (resp. $E\langle f\rangle = E(f)$) the complexity (resp. number of essential variables) of class $\langle f\rangle$. The* **multiplicity** *of the class $\langle f\rangle$, denoted by $R\langle f\rangle$, is the result $L\langle f\rangle - E\langle f\rangle$: it corresponds to the number of repetitions of variables in a minimal tree of a function from $\langle f\rangle$.*

We recall that a Boolean expression is said *satisfiable* if does not represent the constant function false.

Theorem 5. *Let $(k_n)_{n\geq 1}$ be an increasing sequence of integers tending to $+\infty$ when n tends to $+\infty$. A random Catalan expression is satisfiable with probability tending to 1, when the size of the expression tends to infinity.*

Theorem 6. *Let $(k_n)_{n\geq 1}$ be an increasing sequence of integers tending to $+\infty$ when n tends to $+\infty$. There exists a sequence $(M_n)_{n\geq 1}$ such that $M_n \sim \frac{n}{\ln n}$ (when n tends to $+\infty$) and such that, for all fixed equivalence classes of Boolean functions $\langle f\rangle$, there exists a positive constant $\lambda_{\langle f\rangle}$ satisfying*

(i) if, for all sufficiently large n, $k_n \leq M_n$, then, asymptotically when n tends to $+\infty$,
$$\mathbb{P}_n\langle f\rangle \sim \lambda_{\langle f\rangle} \cdot \left(\frac{1}{k_{n+1}}\right)^{R\langle f\rangle+1};$$

(ii) if, for all sufficiently large n, $k_n \geq M_n$, then, asymptotically when n tends to $+\infty$,
$$\mathbb{P}_n\langle f\rangle \sim \lambda_{\langle f\rangle} \cdot \left(\frac{\ln n}{n}\right)^{R\langle f\rangle+1}.$$

Informally, $\lambda_{\langle f\rangle}$ is related to the number of places where some large trees can be plugged in minimal trees. By taking the complexity of both extremal constant functions true and false to 0, the theorem fits to their equivalence classes too.

In the *finite* context [2, 11], 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 $\langle f\rangle$, asymptotically when k tends to infinity,

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

Of course, interchanging the two limits is not a priori possible. However, the *finite* context looks like a degenerate case of our model where there exists an fixed integer k such that $k_n = k$ for all $n \geq 1$. But let us remark that we assume in the present paper that k_n tends to $+\infty$ when n tend to infinity: the case $k_n = k$ is thus not a particular case of our results.
Concerning the infinite context [10, 9] $k_n = +\infty$, we already noticed that the cases such that k_n is larger than n are equivalent to the model $k_n = n$, even if $k_n = +\infty$. Therefore, this infinite context is actually the extreme case $k_n = n$ of our model, and is thus fully treated in the present paper. In this specific setting, the Stirling numbers introduced in Proposition 3 induce Bell numbers, that naturally appear in [10, 9].

4 Technical key points

In this section, we state the technical core of our results, and we demonstrate how a threshold does appear according to the dependence k_n in n.

4.1 Threshold induced by k_n's behaviour

Definition 7. *Let us define the following quantity:* $B_{n,k_n} = \sum_{p=1}^{k_n} \left\{ {n \atop p} \right\} 2^{-p}$. *The number* B_{n,k_n} *quantitatively represents the labelling constraints of leaf-labelling by variables (cf. Proposition 3).*

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

Proposition 8 (for example [13])**.** *For all* $n \geq 1$*, for all* $p \in \{1, \ldots, n\}$*,*

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

In view of these inequalities and of the expression of B_{n,k_n} (cf. Definition 7), both following sequences naturally appear:

Lemma 9. *Let* n *be a positive integer.*

(i) The sequence $(a_p^{(n)})_{p\in\{1,\ldots,n\}} = \left(\frac{p^n}{p!}2^{-p}\right)_p$ *is unimodal, i.e. there exists an integer* M_n *such that* $(a_p)_p$ *is strictly increasing on* $\{1, \ldots, M_n\}$ *and strictly decreasing on* $\{M_n+1, \ldots, n\}$*.*

(ii) Moreover, the sequence $(M_n)_n$ *is increasing and asymptotically satisfies:* $M_n \sim n/\ln n$*.*

We are now ready, to understand the asymptotic behaviour of B_{n,k_n}: *roughly speaking, asymptotically,* B_{n,k_n} *does essentially only depend on the terms around* M_n *in the sum.*

Lemma 10. *Let* $(u_n)_{n\geq 1}$ *be an increasing sequence such that* $u_n \leq n$ *for all integer* $n \geq 1$ *and* u_n *tends to* $+\infty$ *when* n *tends to* $+\infty$*.*

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

$$B_{n,u_n} = \Theta\left(\sum_{p=u_n-\delta_n}^{u_n} \frac{p^n}{p!}2^{-p}\right). \tag{1}$$

(ii) If, for large enough n*,* $u_n \geq M_n$*, then, for all sequences* $(\delta_n)_{n\geq 1}$ *such that* $\delta_n = o(u_n)$ *and* $\frac{u_n\sqrt{\ln u_n}}{n} = o(\delta_n)$*, for all sequences* $(\eta_n)_{n\geq 1}$ *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* $+\infty$*,*

$$B_{n,u_n} = \Theta\left(\sum_{p=M_n-\delta_n}^{\min\{M_n+\eta_n,u_n\}} \frac{p^n}{p!}2^{-p}\right). \tag{2}$$

Definition 11. *Let the fraction* rat_n *be the quantitative evolution of the leaf-labelling constraints from trees of size* $n-1$ *to size* n*:* $\mathsf{rat}_n = B_{n-1,k_n}/B_{n,k_n}$.
Its asymptotic behaviour is quantified by the two following Lemmas 12 and 13.

Let us now deduce the following results on the behaviour of B_{n,k_n}, when n tends to infinity.

Lemma 12. *Let* $(k_n)_{n\geq 1}$ *be a sequence of integers that tends to* $+\infty$ *when* n *tends to* $+\infty$. *Let us assume that* $k_n \leq M_n$ *for large enough* n*, then, asymptotically when* n *tends to infinity,*

$$\frac{B_{n,k_{n+1}}}{B_{n+1,k_{n+1}}} = \Theta\left(\frac{1}{k_{n+1}}\right).$$

Lemma 13. *Let* $(k_n)_{n\geq 1}$ *be a sequence of integers that tends to* $+\infty$ *when* n *tends to* $+\infty$. *Let us assume that* $k_n \geq M_n$ *for large enough* n*, then, asymptotically when* n *tends to infinity,*

$$\frac{B_{n,k_{n+1}}}{B_{n+1,k_{n+1}}} = \Theta\left(\frac{\ln n}{n}\right).$$

4.2 Adjustment of Kozik's pattern language theory

In 2008, Kozik [11] 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 and/or trees with a finite number of variables.

We recall the definitions of patterns, illustrate them on an example and then extend Kozik's paper results to our new model.

Definition 14. *(i) A* **pattern** *is a binary tree with internal nodes labelled by* $\wedge$ *or* $\vee$ *and with external nodes labelled by* $\bullet$ *or* $\square$. *Leaves labelled by* $\bullet$ *are called* **pattern leaves** *and leaves labelled by* $\square$ *are called* **placeholders**. *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 placeholder 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 placeholders.

Definition 15. *We define the composition of two pattern languages* $L[P]$ *as the pattern language of trees which are obtained by replacing every placeholder of a tree from* L *by a tree from* P.

Definition 16. *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* $\{(x,y) : |x| \leq \tau+\varepsilon, |y| \leq m(\tau)+\varepsilon\}$ *for some positive* ε.

Definition 17. *Let* L *be a pattern language,* $\mathcal{M}$ *be a family of trees and* Γ *a subset of* $\{x_i\}_{i\geq 1}$*, whose cardinality does not depend on* n. *Given an element of* $L[\mathcal{M}]$*,*

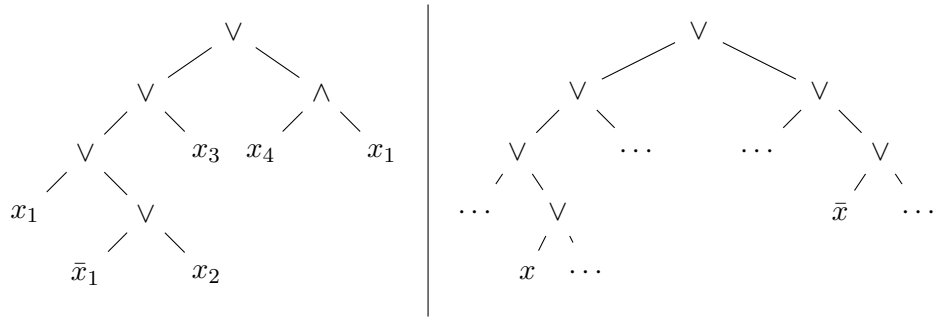

Figure 1: *Left: a tree that computes the function* true. *Right: a simple tautology.*

(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 18. *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.*

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 $1/8$.

We can, for example, define the unambiguous pattern language N by induction as follows: $N = \bullet | N \vee N | N \wedge \square$, meaning that a pattern from N is either a single pattern leaf, or a tree rooted by $\vee$ whose two subtrees are patterns from N, or a tree rooted by $\wedge$ whose left subtree is a pattern from N and whose right subtree is a placeholder. 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 subcritical for $\mathcal{I}$.

On the left-hand side of Fig. 1, we have depicted a Boolean tree that computes the constant function true. It has 5 N-pattern leaves, 1 N-repetition and 2 $(N, \{x_2\})$-restrictions.
The following key lemma is a generalization of the corresponding lemma of Kozik [11, Lemma 3.8].

Lemma 19. *Let* L *be an unambiguous pattern, and* $\mathcal{T}$ *the families of* and/or *trees. Let* r *be a fixed positive integer. Let* $T_n^{[r]}$ *(resp.* $T_n^{[\geq r]}$*) be the number of labelled (with at most* k_n *variables) trees of* $L[\mathcal{T}]$ *of size* n *and with* r *L-repetitions (resp. at least* r *L-repetitions). We assume that* L *is sub-critical for the family* $\mathcal{I}$ *of the unlabelled-leaves trees. Then, asymptotically when* n *tends to infinity,*

$$\frac{T_n^{[r]}}{t_n} = \mathcal{O}\left(\mathsf{rat}_n^r\right) \qquad \textit{and} \qquad \frac{T_n^{[\geq r]}}{t_n} = \mathcal{O}\left(\mathsf{rat}_n^r\right).$$

Proof. The number of labelled trees of $L[\mathcal{T}]$ of size n and with at least r L-repetitions is given by:

$$t_n^{[\geq r]} = \sum_{d=r+1}^{n} I_n(d)\mathrm{Lab}(n, k_n, d, r),$$

where $I_n(d)$ is the number of tree-structures with d L-pattern leaves and the number $\mathrm{Lab}(n, k_n, d, r)$ 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:

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

The factor 2^n corresponds to the polarity of each leaf (the variable labelling it is either negated or not); 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 $B_{n-r-j+1,k_n}$ corresponds to the choices of the variable assigned to each class of leaves. Therefore,

$$t_n^{[\geq r]} \leq 2^n \cdot B_{n-r,k_n} \sum_{j=1}^{r} \left\{ {r+j \atop j} \right\} \sum_{d=r+j}^{n} I_n(d) \binom{d}{r+j}.$$

Let $\ell(x, y)$ be the generating function of the pattern L. Then, for all $p \geq 0$,

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

Thus,

$$\frac{t_n^{[\geq r]}}{t_{n,k_n}} \leq \frac{B_{n-r,k_n}}{B_{n,k_n}} \sum_{j=1}^{r} \left\{ {r+j \atop j} \right\} \frac{[z^n] z^{r+j} \frac{\partial^{r+j} \ell}{\partial x^{r+j}}(z, I(z))}{[z^n] I(z)}.$$

Since $z^{r+j} \frac{\partial^{r+j} \ell}{\partial x^{r+j}}(z, I(z))$ and $I(z)$ have the same 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:

$$\frac{t_n^{[r]}}{t_n} \leq \frac{t_n^{[\geq r]}}{t_n} = \mathcal{O}\left(\frac{B_{n-r,k_n}}{B_{n,k_n}} \right) = \mathcal{O}\left(\mathsf{rat}_n^r \right).$$

□ □

5 Behaviour of the probability distribution

Once we have adapted the pattern theory to our model and proved the central Lemma 19, we are ready to quantitatively study our model. A first step consists to understand the asymptotic behaviour of $\mathbb{P}_n\langle \mathsf{true} \rangle$. It is indeed natural to focus on this "simple" function before considering a general class $\langle f \rangle$; and moreover, 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 equivalence class. We prove in this section the main Theorem 6 for both classes $\langle \mathsf{true} \rangle$ and $\langle \mathsf{false} \rangle$ of complexity zero, using the duality of both connectives $\wedge$ and $\vee$ and both positive and negative literals. Theorem 5 is then obtained as a by-product of Theorem 6. The main ideas of the proof for a general equivalence class are given in Section 5.2.

5.1 Tautologies

Recall that a **tautology** is a tree that represents the Boolean function true. Let $\mathcal{A}$ be the family of tautologies. In this part, we prove that the probability of $\langle \mathsf{true} \rangle$ is asymptotically equal to the ratio of a simple subset of tautologies.

Definition 20 (cf. right-hand side of Fig. 1)**.** *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 leaves are labelled by $\vee$-connectives. We denote by ST the family of simple tautologies.*

Proposition 21. *The ratio of simple tautologies verifies*

$$\mu_n(ST) = \frac{ST_n}{t_n} \sim \frac{3}{4}\mathsf{rat}_n, \text{ when } n \text{ tends to infinity.}$$

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

The latter proposition gives us for free the proof of Theorem 5. In fact, both dualities between the two connectives and positive and negative literals transform expressions computing true to expressions computing false, which implies $\mathbb{P}_n\langle \mathsf{false} \rangle = 3/4 \cdot \mathsf{rat}_n$. Moreover, the only expressions that are not satisfiable compute the function false and $\mathbb{P}_n\langle \mathsf{false} \rangle = 3/4 \cdot \mathsf{rat}_n$ tends to 0 as n tends to infinity, which proves Theorem 5.

5.2 Probability of a general class of functions

With similar arguments than those used for tautologies, it is possible to prove that the probability of the class of projections (i.e. $(x_i)_{i\geq 1} \mapsto x_j$) is equivalent to $5/8 \cdot \mathsf{rat}_n$, when n tends to $+\infty$.

This last section is devoted to the general result, i.e. to the study of the behaviour of $\mathbb{P}_n\langle f \rangle$ for all fixed $f \in \mathcal{F}$. The main idea of this part is that, roughly speaking, *a typical tree computing a Boolean function in $\langle f \rangle$ is a minimal tree of $\langle f \rangle$ into which a single large tree has been plugged.*

sketch. Our aim is to describe the asymptotic behaviour of $\mathbb{P}_n\langle f \rangle$, for a given class of Boolean functions $\langle f \rangle$.

- We first define several notions of *expansions* of a tree: the idea is to replace in a tree, a subtree S by $T \wedge S$, where T is chosen such that the expanded tree still computes the same function.
- The ratio of minimal trees of $\langle f \rangle$ expanded once is of the order of $\mathsf{rat}_n^{R(f)+1}$.
- The ratio of trees computing a function from $\langle f \rangle$ is asymptotically equal to the ratio of minimal trees expanded once.

The most technical part of the proof is the last one, because we need a precise upper bound of $\mathbb{P}_n\langle f \rangle$. But the ideas are more or less the same as those developed for the class $\langle \mathsf{true} \rangle$. □ □

6 Conclusion

We focussed on the logical context of and/or connectives because of the richness of this logical system: normal forms, functional completeness. However the implicational logical system (e.g. [5, 9]) could also be studied in this new context and we deeply believe the general behaviour to be identical. Indeed, the key idea is that *each repetition induces a factor* rat_n, and this remains true in all those models – although pattern theory does not adapt to every model, e.g. models with *implication*. Extending our results to these models would give nice unifications of the known results of the literature: papers [11, 5, 9] and [8].

The numerous results of the last decade in quantitative logics are now linked, through this new model, to satisfiability problems. Our Catalan model of expressions behaves differently than k–SAT or 2–XORSAT problems: asymptotically, almost all expressions are satisfiable, regardless of the ratio between the number of variables and the size of expressions. We can thus conclude that the behaviour of a SAT problem heavily depends on the considered subfamily of Boolean expressions.

Acknoledgements: We are grateful to Pierre Lescanne, whose remark at CLA'12 has allowed us to go beyond our initial idea and consider the more general framework presented here. We also want to thank Brigitte Chauvin and Danièle Gardy who proof-read a previous version of this paper and gave us precious advises to improve it.

References

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

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

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

[4] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge U.P., 2009.

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

[6] D. Gardy. Random Boolean expressions. In *Colloquium on Computational Logic and Applications*, volume AF, pages 1–36. DMTCS, 2006.

[7] A. Genitrini and B. Gittenberger. No Shannon effect on probability distributions on Boolean functions induced by random expressions. In *21st Meeting Analysis of Algorithms*, pages 303–316, 2010.

[8] A. Genitrini, B. Gittenberger, V. Kraus, and C. Mailler. Probabilities of Boolean functions given by random implicational formulas. *Electronic Journal of Combinatorics*, 19(2):P37, 20 pages (electronic), 2012.

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

[10] A. Genitrini, J. Kozik, and M. Zaionc. Intuitionistic vs. classical tautologies, quantitative comparison. In *TYPES*, pages 100–109, 2007.

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

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

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