Some Typical Properties of Large AND/OR Boolean Formulas

Hanno Lefmann
Universität Dortmund
FB Informatik, LS 2
D-44221 Dortmund
Germany
lefmann@ls2.informatik.uni-dortmund.de

and

Petr Savický[1]
Institute of Computer Science
Academy of Sciences of Czech Republic
Prague, Czech Republic
savicky@uivt.cas.cz

Abstract

In this paper typical properties of large random Boolean AND/OR formulas are investigated. Such formulas with n variables are viewed as rooted binary trees chosen from the uniform distribution of all rooted binary trees on m nodes, where n is fixed and m tends to infinity. The leaves are labeled by literals and the inner nodes by the connectives AND/OR, both uniformly at random. In extending the investigation to infinite trees, we obtain a close relation between the formula size complexity of any given Boolean function f and the probability of its occurrence under this distribution, i.e., the negative logarithm of this probability differs from the formula size complexity of f only by a polynomial factor.

1 Introduction

In this paper we are going to study the distribution of Boolean functions determined by large random AND/OR Boolean formulas with a given number n of variables. We consider such formulas to be rooted binary trees chosen from the uniform distribution on trees with m leaves, where m tends to infinity, labeled by connectives and variables. Each of the $m-1$ inner nodes has degree two and is labeled by AND or OR with probability $1/2$ and independently of the labeling of all the other nodes. Each leaf is labeled by a literal, i.e. a variable or its negation, chosen from the uniform distribution on the $2n$ literals and independently of the labeling of all the other nodes.

[1] This research was supported by GA CR, Grant No. 201/95/0976, and by Heinrich-Hertz-Stiftung while visiting Universität Dortmund, FB Informatik, LS II.

Although the formula is large, it appears that, with high probability, the function computed by the formula is in fact determined only by a small part of the formula. Using this, we establish a close relation between the formula size complexity of any Boolean function and its probability in the distribution described above.

The study of the uniform distribution on AND/OR formulas of size approaching infinity was suggested by Woods [12]. He used a variant of the model described above. Namely, he used random formulas based on trees chosen from the uniform distribution on all nonisomorphic rooted trees of a given size and arbitrary degree. Woods proved [13] the existence of the limit probabilities for all Boolean functions and the fact that all these probabilities are positive. Moreover, he asked [12], whether these probabilities are related to the formula size complexity of the Boolean functions. In this paper, we present a natural distribution on trees, for which an affirmative answer to the question of Woods may be proved.

The distribution on functions represented by large AND/OR Boolean formulas was studied also in [4]. The main question there concerns the distribution of the weight of the function represented by the random formula, i.e. the number of ones in its table. In particular, the following is proved there. If both the size m of the formula and the number n of variables tend to infinity, then, for any constants a, b with $0 \leq a < b \leq 1$, the probability of the event that the weight of the random function is in the interval $[a2^n, b2^n)$ converges to a positive limit. This defines a probability measure on the interval $\langle 0, 1\rangle$. In [4], some properties of this measure are investigated and a motivation for investigating this measure, from the point of view of reasoning with uncertain information, is discussed.

A related model for studying the relation between the probability of Boolean functions and their complexity was suggested by Friedman [2]. In his model, there is a sequence of probability distributions, where each of these distributions is defined on formulas of the same size and with the same tree structure. The first distribution is defined on some simple functions. Each of the following distributions is formed by combining random functions chosen from some previous distributions using Boolean connectives. Hence, the complexity of formulas increases in the sequence. Friedman [2] suggests to study the moments of such distributions. A better understanding of the behaviour of these moments might yield lower and upper bounds in complexity theory.

In particular, Friedman investigated the moments of distributions involved in the random k-SAT problem, which includes iterated conjunction of small random disjunctions. For these moments there is a formula involving coefficients with a geometric interpretation. All of the 1-SAT coefficients and some of the 2-SAT coefficients are described in [2].

A model based on a sequence of distributions on Boolean formulas of increasing size, such that in each of the distributions the tree structure of the formulas is fixed, was studied also in [8] and [10]. There it is proved that the studied sequence of distributions converges to the uniform distribution on all Boolean functions. Moreover, in [10], using a sharp bound on the rate of this convergence, a relation is proved between the formula size complexity of any Boolean function and the supremum of the probability of the occurence of this function over all distributions in the sequence. On the contrary to the present result, the relation does not appear in the limit distribution, since it is the uniform one.

The limit distribution on formulas described in the present paper has the property that disjoint subformulas of the random formula viewed as random variables are independent.

Boolean formulas of this kind were already studied by Boppana, Razborov, Valiant and others, and used to prove results on the formula size complexity of the majority function and of the representation of Ramsey graphs, see [1], [7], [9] and [11]. In all these results, the independency of subformulas is the basic tool. Moreover, in a general setting, there is a connection to the study of nonlinear dynamical systems defined on finite functions (see [5], [6]). The common point is again the combination of independent random functions via simple rules.

The outline of the present paper is as follows. In Section 2, we investigate a decomposition of a tree into segments and prove some properties of the distribution of segments in a large random tree. For the analysis of this random tree we use similar ideas as in [4], but we need more accurate estimates. Using the decomposition of the tree into segments, we introduce a distribution on infinite trees, determined by a sequence of independent choices of segments from some distribution. It turns out that the distribution on Boolean functions determined by this distribution on infinite trees is equal to the limit of the distributions on functions determined by the uniform distribution on formulas of size m, if m tends to infinity. This characterization of the limit distribution is the crucial tool to derive in Section 3 lower and upper bounds for the probability $P(f)$, that a given function f occurs, in terms of the formula size complexity $L(f)$ of f. In particular, we will show that the negative logarithms of the lower and upper bound differ from the formula size complexity of the function at most by a polynomial factor:

Theorem 1.1 *There exist positive constants $c_1, c_2 > 0$, such that for every large enough positive integer n the following is valid:*

For every Boolean function f of n variables satisfying $L(f) \geq \Omega(n^3)$, it is

$$e^{-c_1 L(f) \log n} \leq P(f) \leq e^{-c_2 L(f)/n^3} .$$

Finally, in Section 4 we state some open problems.

2 Approximation by an Infinite Tree

First, we will investigate the tree structure of AND/OR formulas. For doing so, we need some definitions. The *size* of a tree is the number of its nodes. A binary tree consisting of two nonempty subtrees connected to the root will be called *1-separable* or only separable, if its two subtrees have different size. In such a tree, the unique maximum subtree will be called *tail*. If the tail is also separable, we say that the original tree is *2-separable*, and so on. Hence, a k-separable tree allows k steps of such a decomposition. The tail obtained in the i-th step will be called *i-th tail*, where the 0-th tail is the whole tree. Moreover, the whole tree with the k-th tail replaced by a new leaf, which is a *special leaf*, distinguishable from all the other leaves, will be called the *k-head* or, if k follows from the context, simply head. The special leaf is included in order to mark the position, where the tail was connected to. We shall also consider the decomposition of the k-head into k segments, where the *i-th segment* is the $(i-1)$-st tail with the i-th tail replaced by the special leaf denoting the original position of the i-th tail. We do not count the special leaves to the sizes of the

segments. Hence, the size of a k-separable tree is the sum of the sizes of its k segments plus the size of the k-th tail.

In particular, we shall prove in the following that, if m tends to infinity, then, with probability approaching 1, a random binary tree of size $2m-1$ is 1-separable and the corresponding tail has size at least $(2m-1-t(m))$, where $t(m) \ll m$ is any function tending to infinity with m.

Let $g(x) = \sum_{i=1}^{\infty} a_i x^i$ be the generating function for the nonempty rooted binary trees. That is, a_i counts the number of rooted binary trees of size i. A single node is the only binary tree with at most two vertices, hence $a_1 = 1$ and $a_2 = 0$. Note that $a_i = 0$ for every even i. Using the recursion $a_i = \sum_{j=1}^{i-2} a_j a_{i-j-1}$ for all $i \geq 3$, we obtain the identity $g(x) = x(1+g(x)^2)$. From this and the fact that $g(0) = 0$, we infer that

$$g(x) = \frac{1}{2x} \cdot \left(1 - \sqrt{1-4x^2}\right).$$

Using Taylor expansion, we obtain for $0 \leq x \leq 1/2$ that

$$g(x) = \frac{1}{2x} \cdot \left(1 - \sum_{i=0}^{\infty} \binom{1/2}{i} \cdot (-4x^2)^i\right) = \sum_{n=1}^{\infty} C(2\,n-1) \cdot x^{2n-1}, \tag{1}$$

where $C(2\,n-1)$ are the Catalan numbers,

$$C(2\,n-1) = \frac{1}{n} \cdot \binom{2\,n-2}{n-1},$$

counting the number of rooted binary trees of size $2n-1$ for every $n \geq 1$.

In every k-separable tree, each segment consists of its root and two children. One of them is the special leaf, the second is some nonempty binary tree. Since we do not count the special leaf to the size, the size of a segment is 1 plus the size of the nonempty subtree. Hence, the size of the segment is always even. As there are two possible positions for the special leaf, the number of segments of size $2r$ is $2 \cdot C(2\,r-1)$.

Lemma 2.1 *There exists an $\varepsilon > 0$ such that the following is true. Let k and $r_1, r_2, \ldots, r_k$ satisfy $r_i \geq 1$ for all $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} r_i \leq \varepsilon m$. Let H be a k-head with i-th segment of size $2r_i$ for $i = 1, 2, \ldots, k$. Then, the probability that a random tree of size $2m-1$ is k-separable and that its k-head is H equals*

$$\left(1 + O\left(\frac{1}{m}\sum_{i=1}^{k} r_i\right)\right) \cdot \prod_{i=1}^{k} 2^{-2r_i}. \tag{2}$$

Proof: The required probability P is equal to the ratio of the number of k-separable trees with the given head H over the number of all trees of size $2m-1$. Every k-separable tree with head H consists of H and a tail of size $2(m - \sum_{i=1}^{k} r_i) - 1$. By our assumptions on $\sum_{i=1}^{k} r_i$, composing H with any tail of this size yields a k-separable tree. Moreover, different tails lead to different trees. Hence, we have

$$P = \frac{C(2\,(m - \sum_{i=1}^{k} r_i) - 1)}{C(2\,m-1)}. \tag{3}$$

Using Stirling's formula $n! = \sqrt{2\pi n}(n/\mathrm{e})^n(1+O(1/n))$, we obtain for $n \to \infty$ that

$$C(2\,n-1) = \frac{1}{4\sqrt{\pi}} n^{-3/2} 2^{2n} \left(1 + O\left(\frac{1}{n}\right)\right). \tag{4}$$

This implies

$$\frac{C(2\,(m-\sum_{i=1}^{k} r_i)-1)}{C(2\,m-1)} = \left(1 - \frac{1}{m}\sum_{i=1}^{k} r_i\right)^{-3/2} \cdot \frac{1 + O\left(\frac{1}{m-\sum_{i=1}^{k} r_i}\right)}{1+O\left(\frac{1}{m}\right)} \cdot 2^{-2\sum_{i=1}^{k} r_i}\,.$$

Now, there is an $\varepsilon > 0$ depending on the constant in the O-term from (4) such that, if $\sum_{i=1}^{k} r_i \le \varepsilon m$, then

$$P = \left(1 + O\left(\frac{1}{m}\sum_{i=1}^{k} r_i\right)\right) \cdot \prod_{i=1}^{k} 2^{-2r_i}.$$

This proves the lemma. □

By the following result, for k not too large, the tree is k-separable with probability approaching 1 if m tends to infinity.

Lemma 2.2 *Let k and r, possibly depending on m, be such that $kr = o(m)$ and $k = o(r^{1/2})$. Then, the probability P that the tree is k-separable and each of the k corresponding segments has size at most $2r$, equals*

$$P = 1 - O\left(\frac{k}{r^{1/2}}\right) + O\left(\frac{kr}{m}\right)\,.$$

Proof: In order to prove the lemma, we shall compute the sum of the probabilities from Lemma 2.1 for all k-heads H with the segments of size at most $2r$. As a first approximation, let us consider the limits, when m approaches infinity. By Lemma 2.1 used for $k = 1$, the limit of the probability of the occurrence of an individual segment of size $2j$ is 2^{-2j}. There are $2C(2j-1)$ segments of size $2j$. Using (1) for $x = 1/2$, i.e., $g(1/2) = 1$, we obtain

$$\sum_{j=1}^{\infty} \frac{2C(2\,j-1)}{2^{2j}} = 1\,. \tag{5}$$

To express the probability required in the lemma, we will estimate the sum of the first r terms of this series. By (4) we have

$$\frac{2C(2\,j-1)}{2^{2j}} = O\left(\frac{1}{j^{3/2}}\right)$$

and hence

$$\sum_{j=r+1}^{\infty} \frac{C(2\,j-1)}{2^{2j}} = O\left(\int_r^{\infty} \frac{1}{x^{3/2}}\mathrm{d}x\right) = O\left(\frac{1}{r^{1/2}}\right). \tag{6}$$

By Lemma 2.1, the desired probability P is given by

$$\begin{aligned} P &= \left(1 + O\left(\frac{kr}{m}\right)\right) \cdot \sum_{1 \le r_1, \ldots, r_k \le r} \prod_{i=1}^{k} \frac{2C(2\,r_i - 1)}{2^{2r_i}} \\ &= \left(1 + O\left(\frac{kr}{m}\right)\right) \cdot \left(\sum_{j=1}^{r} \frac{2C(2\,j-1)}{2^{2j}}\right)^k \\ &= \left(1 + O\left(\frac{kr}{m}\right)\right) \cdot \left(1 - O\left(\frac{1}{r^{1/2}}\right)\right)^k \qquad \text{by (6).} \end{aligned}$$

As $kr = o(m)$ and $k = o(r^{1/2})$, we infer that

$$P = 1 - O\left(\frac{k}{r^{1/2}}\right) + O\left(\frac{kr}{m}\right) \ ,$$

which yields the desired result. □

By Lemma 2.1, for m approaching infinity the probability of the occurrence of each individual segment of size $2r$ converges to 2^{-2r}. By (5), these limits determine a well-defined distribution on segments. For the random segments from this distribution consider the usual random labeling of the inner nodes and the leaves, except of the special leaf, by connectives AND/OR and literals containing the Boolean variables $x_1, x_2, \ldots, x_n$. From now on, let n be reserved for the number of these variables and assume that n is fixed. The resulting distribution on labeled segments will be denoted by D_1. Moreover, let D_k be the distribution on labeled k-heads formed by choosing k segments independently from D_1 and connecting them via the special leaves in the straightforward manner.

For a labeled head occurring either in the distribution D_k or as the k-head of a k-separable tree, we define the function computed by the head as follows. Assume that the special leaf is labeled by a new variable. Hence, we have a labeling of the head by $n+1$ variables. Now, the function computed by the labeled head is the partial Boolean function (i.e., not defined everywhere in its possible domain) of n original variables, defined for those inputs, for which the value of the formula of $n+1$ variables does not depend on the new variable. If the head of a k-separable tree computes a total function (defined everywhere), we say that the head is *closed*.

For every total Boolean function f, the probability $P_k(f)$ of its occurrence in the distributions D_k is nondecreasing with k, since adding a new segment may change the function represented by the head only if the function determined by the previous segments is not a total function. Hence, the probabilities $P_k(f)$ have some limits $P(f) = \lim_{k \to \infty} P_k(f)$. As every function is represented by some labeling of a sufficiently large closed head, these limits are positive.

Let D be the distribution on labeled infinite trees, which are formed by connecting an infinite sequence of labeled segments chosen independently from D_1. Note that a tree chosen from D contains exactly one infinite path. We say that a tree chosen from the distribution D *computes a function* f, if for some integer k, the first k labeled segments in the tree form a closed k-head computing f. It is easy to see that this happens with probability one. It is also easy to see that the probability that a tree from D computes f is equal to $P(f)$.

The size of a Boolean formula is defined as the number of occurrences of variables contained in the formula. Hence, a formula of size m is based on a tree of size $2m-1$.

Theorem 2.3 *Let f be a Boolean function of n variables and let m tend to infinity. Then, the probability, that a random AND/OR formula chosen from the uniform distribution on formulas of size m computes f, converges to $P(f)$.*

Proof: In order to prove the theorem, it suffices to prove that for every total Boolean function f, the difference between the probability of computing f by a random formula chosen from the uniform distribution on formulas of size m and $P_k(f)$ converges to zero, if k and m both tend to infinity in a controlled way.

Fix some ε with $0 < \varepsilon < 1/3$ and set $k = \lceil m^{1/3-\varepsilon} \rceil$. Moreover, we will use an auxiliary parameter $r = \lceil m^{2/3} \rceil$. Denote by $\Pr_1(A)$ the probability of some event A in the uniform distribution on formulas of size m, while $\Pr_2(A)$ will be the probability of A in the distribution D_k. Let $\mathcal{H}(k,r)$ denote the set of all k-heads containing only segments of size at most $2r$. For any k-head $H \in \mathcal{H}(k,r)$, by $\Pr_1(H)$ we mean the probability of the event "the tree is k-separable and its k-head is equal to H". Similarly, $\Pr_2(H)$ is the probability of the event "the k-head H occurs". Moreover, by $\mathcal{H}(k,r)$, used either in $\Pr_1(.)$ or in $\Pr_2(.)$, we mean the disjunction of the events corresponding to all $H \in \mathcal{H}(k,r)$.

In order to compare the probabilities $\Pr_1(f)$ and $\Pr_2(f)$ of the event that the random formula in the corresponding distribution computes the given total function f, we will first compare the conditional probabilities $\Pr_1(f \mid H)$ and $\Pr_2(f \mid H)$. The assignments of connectives and literals, which guarantee the total function f in the second of these two conditional events, are exactly the same as the assignments of the k-head H in the first event that make H closed and give the function f. Hence, $\Pr_2(f \mid H) = \Pr_1(f \wedge (H \text{ is closed}) \mid H)$. Therefore,

$$\begin{aligned} 0 &\leq \Pr_1(f \mid H) - \Pr_2(f \mid H) = \Pr_1(f \wedge (H \text{ is not closed}) \mid H) \\ &\leq \Pr_1(H \text{ is not closed} \mid H)\,. \end{aligned}$$

In order to compute this probability, consider a fixed input x. Each of the k nodes on the path from the root to the k-th tail computes either the AND or OR of two subformulas. One of them contains the tail and the second belongs completely to the head. The assignment of connectives and literals is symmetric with respect to the values 0 and 1. Hence, the subformula, which belongs to the head, computes on input x both the values 0 and 1 with probability $1/2$. Hence, with probability $1/2$, the value of the subformula, which contains the tail, has no influence on the value computed in the node. Since this happens for all k nodes in the path independently, the probability, that the tail is needed to compute the value of the function in the root, is at most $(1/2)^k$.

As there are 2^n different inputs, the probability that for at least one of them, the value of the function is not determined by the head is at most 2^{n-k}. By combining the arguments, we obtain

$$|\Pr_1(f \mid H) - \Pr_2(f \mid H)| \leq 2^{n-k} = O(2^{-m^{1/3-\varepsilon}}), \tag{7}$$

as n is fixed.

For our choice of k and r, we have $kr/m = O(m^{-\varepsilon})$. Hence, by Lemma 2.1, we have $\Pr_1(H) = \Pr_2(H) \cdot (1 + O(m^{-\varepsilon}))$ for every k-head $H \in \mathcal{H}(k,r)$. Moreover, $\Pr_1(\neg\mathcal{H}(k,r)) = O(m^{-\varepsilon})$ by Lemma 2.2. The leftmost expression in (6) is equal to the probability that the random segment chosen from D_1 has size at least $2r+2$. Hence, by the same argument as in the proof of Lemma 2.2, we have $\Pr_2(\neg\mathcal{H}(k,r)) = O(k/r^{1/2}) = O(m^{-\varepsilon})$.

Using (7), we summarize as follows

$$\begin{aligned} & \Pr_1(f) - \Pr_2(f) \\ = & \sum_{H \in \mathcal{H}(k,r)} [\Pr_1(f \mid H) \cdot \Pr_1(H) - \Pr_2(f \mid H) \cdot \Pr_2(H)] + O(m^{-\varepsilon}) \\ = & \sum_{H \in \mathcal{H}(k,r)} \Pr_2(H) \cdot [\Pr_1(f \mid H) - \Pr_2(f \mid H) + \Pr_1(f \mid H) O(m^{-\varepsilon})] + O(m^{-\varepsilon}) \\ = & \; O(m^{-\varepsilon}) \,, \end{aligned}$$

since $\sum_{H \in \mathcal{H}(k,r)} \Pr_2(H) \leq 1$. This yields the desired result. □

3 Bounding the Limit Probabilities

For any Boolean function f, the *formula size complexity* $L(f)$ is the minimum size of a formula representing f. The following lower bound on the probability $P(f)$ of the occurrence of f is a direct consequence of our construction of the distribution D.

Theorem 3.1 *Let f be a Boolean function of n variables. Then*

$$P(f) \geq \frac{1}{4} \cdot \left(\frac{1}{8n}\right)^{L(f)+1} .$$

Proof: Consider any formula ϕ of size $L(f)$ representing f and the two closed 2-heads $\phi \vee ((x_i \wedge \neg x_i) \wedge y)$ and $\phi \wedge ((x_i \vee \neg x_i) \vee y)$, where $1 \leq i \leq n$ and y denotes the position of the special leaf. Each of these two heads consists of two segments, one of size $2L(f)$ and the second of size 4. Hence, the tree structure of the two segments together has the probability $(1/4)^{L(f)+2}$ to occur. Since there are $L(f)+2$ leaves and 3 inner nodes in each of the heads, the labeling of each head has the probability $1/(2n)^{L(f)+2} \cdot (1/2)^3$ to occur. There are n possibilities how to choose i and, due to the commutativity of the connectives, 8 equivalent variants of each of the heads having the same probability. Altogether, we have described $16n$ different closed 2-heads computing f, each with probability

$$\frac{1}{8} \cdot \left(\frac{1}{8n}\right)^{L(f)+2} .$$

This proves the theorem. □

Pick a random Boolean formula according to the distribution D and an arbitrary input x. By considerations made already in the proof of Theorem 2.3, the first k segments of

the formula are sufficient to determine the value of the whole infinite formula in x with probability $1-2^{-k}$. Thus, with probability at least $1-2^{n-k}$, the first k segments determine the value of the whole formula for all inputs. This is close to 1 if $k \gg n$. Moreover, estimate (6) together with the definition of D_1 implies that the probability that the size of a segment is bigger than r, is at most $O(1/r^{1/2})$. Hence, with high probability, only small parts (e.g. of polynomial size in n) of the formula are really needed to compute the function. It follows that, with high probability, the computed function has small complexity. This may be used to prove an upper bound on the probability $P(f)$ of the occurence of a function f, if the formula size $L(f)$ of f is large. However, this upper bound would be only one over a polynomial in $L(f)$. In the following, an upper bound of the magnitude one over an exponential of $L(f)$ will be given. To prove this, we demonstrate a way of pruning the infinite formula from the distribution D. The resulting equivalent pruned formula is finite with probability 1 and, moreover, has size l with probability exponentially small in l.

The pruning is controlled by assigning a set of conditions to each inner node of the tree. If the set of conditions assigned to a node is contradictory, the node will be deleted. The conditions are simply some requirements to the values of single variables and they are computed as follows. The root is assigned the empty set of conditions. Assume, an inner node v of the formula is assigned a set ρ of conditions. If both successors of v are inner nodes, the set ρ of conditions is assigned to both of them without any change. If only one of the successors is an inner node, say the left one, let x_i be the variable used in the literal in the right successor. Then, the left successor is assigned the set $\rho \cup \{x_i = a\}$ of conditions , where a is the value of x_i which does not force the AND or OR in the node v to a constant. In the remainder of this section, we always assume that the nodes of the formula from D are assigned to the pruning conditions computed by these rules.

By the construction of the pruning conditions, it is easy to see that the following is true. If some inner node v is assigned a set ρ of conditions and, for some input x, the input variables do not satisfy some of the conditions, then the value computed for x in the node v has no influence to the value computed for x by the whole formula. If ρ contains both $x_i = 0$ and $x_i = 1$ for some i, then the node v has no influence for any input and, hence, it may be deleted. The deletion is performed by replacing the node by any constant and when all such replacements are finished, the formula is transformed to a formula without constants using standard simplification rules.

Let ϕ be a random Boolean formula labeled by the sets of conditions as described. Denote by $\|\phi\|$ the number of inner nodes in the formula ϕ assigned to a consistent set of conditions. After the transformation described above, the number of inner nodes in the new equivalent formula is at most $\|\phi\|$ and hence its formula size is at most $\|\phi\| + 1$.

In order to give a small upper bound on the probability that $\|\phi\|$ is large, where ϕ is chosen from the distribution D, we give an estimate of the expected value

$$\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|}\right] ,$$

where ε is an appropriate positive real number. To this end, we represent the distribution D as a simple growing process on trees and first consider the expectation of a similar quantity in some finite parts of ϕ.

Consider the following two types of nodes, *c-nodes* and *n-nodes* (connecting and normal, respectively). We start with one c-node. In each step, each node is either expanded into two successors or stopped. A c-node is always expanded into one c-node and one n-node. With probability 1/2, the new c-node is either the left successor or the right successor of the old one. An n-node is either expanded or stopped, each with probability 1/2. If expanded, both successors are again n-nodes. It is easy to see that the resulting tree consists of an infinite sequence of independently chosen (possibly infinite) segments connected via c-nodes. Moreover, the event that a given individual segment occurs means that, during the process, in each node of the segment, the result of the random choice matches the given structure. In each node, this happens with probability 1/2 and the random bits used in the random choices are independent. Hence, the probability of a given segment of size $2r$ is $(1/2)^{2r}$. Thus, the tree structure of a single segment is generated according to the distribution D_1. In particular, each segment is finite with probability 1. By assigning the random labeling by connectives and literals to all segments as before, we obtain a formula from D. We shall refer to this process as to the *basic growing process*.

The tree is generated level by level. If all the nodes in level $j+1$ for some j are created, it is known for every node in level j, whether it is an inner node or a leaf. At this time, the random labeling of these nodes by connectives and the literals is chosen. Now, since all nodes in level j have their labels, the sets of conditions for all the nodes in level j may be computed. Let v be an inner node of level j. Note that the distribution of the subtree below v including the labeling and the sets of conditions depend on the rest of the tree only via the labeling of the node v. Hence, we can consider v as a starting node of a separate process. The initialization of the process is given by the type of node v (n-node or c-node) and by the set ρ of conditions in v. Note that v is known to be an inner node, so, if v is an n-node, the process starts by expanding v deterministicaly. We shall refer to this process as to the *generalized growing process*. The basic growing process generating formulas from D is the special case, when the starting node is a c-node and ρ is the empty set of conditions.

For a random formula ϕ generated by the generalized growing process and any integer d, let $\|\phi\|_d$ mean the number of inner nodes of depth at most d in ϕ, which are assigned a consistent set of conditions. If ϕ is generated by the growing process started with a set of conditions ρ, then, by symmetry, the distribution of $\|\phi\|_d$ is the same for all other consistent starting sets of conditions with the same number of elements. Hence, only the number of elements of ρ is taken into account.

In the following, let $\varepsilon > 0$ be a real number, which will be specified later.

Definition 3.2 Let $\alpha(d,k) = \mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_d}\right]$, where ϕ is generated by the generalized growing process starting at an n-node with a consistent k-element set ρ of conditions. Recall that the starting n-node is always expanded by definition of the generalized growing process. Analogously, let $\beta(d,k) = \mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_d}\right]$, where ϕ is generated from a c-node, with a consistent k-element set ρ of conditions.

In the following two lemmas, we give recurrence relations for $\alpha(d,k)$ and $\beta(d,k)$.

Lemma 3.3 *For every real ε and for all integers $d \geq 0$ and $k = 0, 1, \ldots, n$*

$$\begin{aligned} \alpha(0,k) &= 1+\varepsilon \\ \alpha(d+1,k) &= (1+\varepsilon)\left(\frac{1}{4}\alpha(d,k)^2 + \frac{k}{4n}\alpha(d,k) + \right. \\ &\qquad \left. + \frac{1}{2}\left(1-\frac{k}{n}\right)\alpha(d,k+1) + \frac{k}{4n} + \frac{1}{4}\right) . \end{aligned}$$

Proof: Assume that the starting node v is an n-node with a consistent k-element set ρ of conditions. Hence, $\|\phi\|_0 = 1$ and $\alpha(0,k) = 1+\varepsilon$.

In order to prove the second identity, we consider some cases according to the labeling of the successors v_1 and v_2 of the node v. Then, the expected value of $(1+\varepsilon)^{\|\phi\|_{d+1}}$ is computed by using the expansion to the conditional expectations according to these cases. Let ϕ_1 and ϕ_2 be the subformulas below v_1 and v_2. Note that $\|\phi\|_{d+1} = 1 + \|\phi_1\|_d + \|\phi_2\|_d$ for every nonnegative integer d.

In *case 1*, both successors v_1, v_2 are expanded and, hence, become inner nodes. This happens with probability 1/4. Under this condition, the continuation of the process consists of two independent processes, starting in v_1 and v_2, both with an initial k-element set ρ of conditions. Therefore,

$$\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_{d+1}} \mid \text{case 1}\right] \cdot \Pr(\text{case 1}) = (1+\varepsilon)\alpha(d,k)^2 \cdot \frac{1}{4} .$$

In *case 2*, v_1 is expanded and v_2 is stopped. This occurs with probability 1/4. Under the condition that this happens, the set of conditions in v_1 is created by adding the new condition due to v_2 to the set ρ. We distinguish three subcases according to the relation between ρ and the new condition.

In *case 2a*, which happens with probability $k/(2n)$ if case 2 occurs, the new condition is already included in ρ. Then,

$$\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_{d+1}} \mid \text{case 2a}\right] \cdot \Pr(\text{case 2a}) = (1+\varepsilon)\alpha(d,k) \cdot \frac{k}{8n} .$$

In *case 2b*, with probability $k/(2n)$ if case 2 occurs, the new condition is contradictory to some of the conditions in ρ. In this situation, there are no undeleted inner nodes below v_1. Hence,

$$\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_{d+1}} \mid \text{case 2b}\right] \cdot \Pr(\text{case 2b}) = (1+\varepsilon) \cdot \frac{k}{8n} .$$

In *case 2c*, with probability $1-k/n$ if case 2 occurs, the new condition involves a variable not used in ρ and hence it is independent of ρ. In this case, the initial set of conditions in v_1 has $k+1$ elements. Hence,

$$\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_{d+1}} \mid \text{case 2c}\right] \cdot \Pr(\text{case 2c}) = (1+\varepsilon)\alpha(d,k+1) \cdot \frac{1}{4}\left(1-\frac{k}{n}\right) .$$

Case 3, when v_1 is stopped and v_2 is expanded, is symmetric to case 2.

In *case 4*, with probability 1/4, both v_1, v_2 are stopped. In this situation we have

$$\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_{d+1}} \mid \text{case 4}\right] \cdot \Pr(\text{case 4}) = (1+\varepsilon)\cdot\frac{1}{4}\ .$$

By collecting the contributions of all four cases, we obtain the lemma. □

Lemma 3.4 *For every real number $\varepsilon > 0$, and for all nonnegative integers d and k,*

$$\begin{aligned}\beta(0,k) &= 1+\varepsilon \\ \beta(d+1,k) &= (1+\varepsilon)\left(\frac{1}{2}\alpha(d,k)\beta(d,k) + \frac{k}{4n}\beta(d,k) + \right. \\ &\qquad \left. + \frac{1}{2}\left(1-\frac{k}{n}\right)\beta(d,k+1) + \frac{k}{4n}\right)\ .\end{aligned}$$

Proof: Assume that the starting node v is a c-node with a k-element set ρ of conditions. The starting node is expanded to a c-node and an n-node. It is not necessary to distinguish whether the c-node is the left successor or the right successor, since it has no influence on the distribution of the number of undeleted nodes. In the next step, the new c-node is always expanded. There are two cases, which we consider separately, according to the behaviour of the n-node. The n-node is either expanded (case 1′) or stopped (case 2′), each with probability 1/2.

In case 1′, the conditional expectation of $(1+\varepsilon)^{\|\phi\|_{d+1}}$ is $(1+\varepsilon)\alpha(d,k)\beta(d,k)$, since the processes starting in the n-node and in the new c-node are independent.

The case 2′ will be splitted into subcases according to the influence of the condition due to the literal in the n-node to the set ρ' of conditions in the new c-node. With probability $k/(2n)$ we have $\rho' = \rho$, which gives the conditional expectation $(1+\varepsilon)\beta(d,k)$. With the same probability, ρ' is contradictory, which yields the conditional expectation $1+\varepsilon$, since only node v is not deleted. With probability $(1-k/n)$, it is $|\rho'| = |\rho|+1$, giving the conditional expectation $(1+\varepsilon)\beta(d,k+1)$. Now, by combining these conditional expectations according to the probabilities of the corresponding conditions as in Lemma 3.3, we obtain Lemma 3.4. □

Theorem 3.5 *There exists a constant $c > 1$ such that for every positive integer n the following is valid:*

For every Boolean function f of n variables, it is

$$P(f) \le (1+O(1/n))c^{-L(f)/n^3}.$$

Proof: In order to prove the theorem, we first show upper bounds on $\alpha(d,k)$ and $\beta(d,k)$, if ε is 1 over some polynomial in n.

With foresight, set $u_k = 5(n-k+1)$ and $v_k = 5(n-k+2)^2$ for every $k = 0,\ldots,n$. We will find a range of ε, in which the following inequalities hold

$$\alpha(d,k) \le 1 + u_k\varepsilon \tag{8}$$

$$\beta(d,k) \le 1 + v_k\varepsilon\ . \tag{9}$$

for $k = 0, 1, \ldots, n$ and every nonnegative integer d.

Both these inequalities are satisfied for $d = 0$, any $k = 0, 1, \ldots, n$ and any $\varepsilon > 0$. By induction on d, we will show in the following that these two inequalities hold for all $d \geq 0$, provided ε is small enough. The calculations will give a sufficient condition on ε, which guarantees the induction step.

If (8) is satisfied for some $d = d_0$, the inequality

$$(1+\varepsilon)\left(\frac{1}{4}(1+u_k\varepsilon)^2 + \frac{k}{4n}(1+u_k\varepsilon) + \frac{1}{2}\left(1-\frac{k}{n}\right)(1+u_{k+1}\varepsilon) + \frac{k}{4n} + \frac{1}{4}\right) \leq 1 + u_k\varepsilon \quad (10)$$

implies (8) for $d = d_0 + 1$, since, by Lemma 3.3, the LHS of (10) is an upper bound on $\alpha(d_0+1, k)$. By expanding the products and collecting the terms ε^i with the same exponent, the difference between the RHS and the LHS of (10) can be transformed to

$$\begin{aligned} & \left[\left(\frac{1}{2} - \frac{k}{4n}\right)u_k - \frac{1}{2}\left(1-\frac{k}{n}\right)u_{k+1} - 1\right]\varepsilon + O(n^2\varepsilon^2) \geq \\ \geq \quad & \left(\frac{1}{2} - \frac{k}{4n}\right)(u_k - u_{k+1} - 4)\varepsilon + O(n^2\varepsilon^2) \geq \frac{1}{4}\varepsilon + O(n^2\varepsilon^2)\,. \end{aligned}$$

We conclude that there is a positive constant δ_1 such that for every ε with $0 < \varepsilon \leq \delta_1/n^2$ inequality (10) is satisfied and so, for all $d \geq 0$, inequality (8) holds.

To extend inequality (9) to all $d \geq 0$, a stronger condition on ε is needed in the induction step, but the arguments are similar. The corresponding inequality, which guarantees (9) for $d = d_0 + 1$ – if satisfied for $d = d_0$ – is, by Lemma 3.4,

$$(1+\varepsilon)\left(\frac{1}{2}(1+u_k\varepsilon)(1+v_k\varepsilon) + \frac{k}{4n}(1+v_k\varepsilon) + \frac{1}{2}\left(1-\frac{k}{n}\right)(1+v_{k+1}\varepsilon) + \frac{k}{4n}\right) \leq 1 + v_k\varepsilon\,. \quad (11)$$

Taking the difference of the RHS of this inequality and its LHS, we obtain

$$\begin{aligned} & \left[\left(\frac{1}{2} - \frac{k}{4n}\right)v_k - \frac{1}{2}\left(1-\frac{k}{n}\right)v_{k+1} - \frac{1}{2}u_k - 1\right]\varepsilon + O(n^3\varepsilon^2) \geq \\ \geq \quad & \left(\frac{1}{2} - \frac{k}{4n}\right)(v_k - v_{k+1} - 2u_k - 4)\varepsilon + O(n^3\varepsilon^2) \geq \frac{1}{4}\varepsilon + O(n^3\varepsilon^2)\,. \end{aligned}$$

Hence, there exists a positive constant $\delta_2 \leq \delta_1$ such that for every ε, $0 < \varepsilon \leq \delta_2/n^3$, inequality (11) is satisfied. It follows by induction that, for every such ε, both inequalities, (8) and (9), hold for every $d \geq 0$. For the remainder of the proof let $\varepsilon = \delta_2/n^3$.

Let ϕ be a formula from D. By definition of $\beta(d,k)$, it follows that $\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_d}\right] = \beta(d,0)$. Since $\|\phi\|_d \geq \|\phi\|_{d-1}$, the limit

$$\lim_{d\to\infty} \mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_d}\right]$$

exists and, by (9), is bounded from above by $1 + 5(n+2)^2\varepsilon$.

In order to obtain an upper bound on $\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|}\right]$, we use the following well-known result:

Lemma 3.6 (see [3]) *If $\sum_{d=0}^{\infty} \mathrm{E}[|Y_d|]$ is convergent for a sequence of random variables Y_d, then*

$$\mathrm{E}\left[\sum_{d=0}^{\infty} Y_d\right] = \sum_{d=0}^{\infty} \mathrm{E}[Y_d] \; .$$

Let $Y_0 = (1+\varepsilon)^{\|\phi\|_0}$ and $Y_d = (1+\varepsilon)^{\|\phi\|_d} - (1+\varepsilon)^{\|\phi\|_{d-1}}$ for every $d \geq 1$. Since $\|\phi\|_d \geq \|\phi\|_{d-1}$, we have $Y_d \geq 0$, and hence

$$\sum_{d=0}^{\infty} \mathrm{E}[|Y_d|] = \sum_{d=0}^{\infty} \mathrm{E}[Y_d] = \lim_{d\to\infty} \mathrm{E}\left[(1+\varepsilon)^{\|\phi\|_d}\right] \leq 1 + 5(n+2)^2\varepsilon = 1 + O(1/n) \; .$$

This shows that the assumption of Lemma 3.6 is satisfied, and it follows that

$$\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|}\right] = \mathrm{E}\left[\sum_{d=0}^{\infty} Y_d\right] = \sum_{d=0}^{\infty} \mathrm{E}[Y_d] \leq 1 + O(1/n) \; .$$

Using this, we will finish the proof of Theorem 3.5 as follows. If ϕ computes f, then $L(f) \leq \|\phi\| + 1$. Hence, by Markov's inequality,

$$\Pr(\phi \text{ computes } f) \leq \Pr\left((1+\varepsilon)^{\|\phi\|} \geq (1+\varepsilon)^{L(f)-1}\right)$$

$$\leq \frac{\mathrm{E}\left[(1+\varepsilon)^{\|\phi\|}\right]}{(1+\varepsilon)^{L(f)-1}} \leq (1+O(1/n))\left(1+\frac{\delta_2}{n^3}\right)^{-L(f)} \leq (1+O(1/n))\cdot c^{-L(f)/n^3}$$

for any absolute constant c with $1 < c < \mathrm{e}^{\delta_2}$ and n large enough. □

As an immediate consequence of Theorems 3.1 and 3.5, we obtain Theorem 1.1.

4 Open Problems

Our results show a close relation between the probability of Boolean functions of n variables in the distribution D and their formula size complexity provided the complexity is $\Omega(n^3)$. The situation for Boolean functions of complexity $o(n^3)$ is unknown. Also, closing the gap between the lower and the upper bound in Theorem 1.1 is an open problem. Another open problem of particular interest is to compute estimates on the probability of explicit functions in the limit distribution, exact enough to have consequences for the complexity of these functions.

Acknowledgement The authors would like to thank to Jan Krajíček for simplifying the limiting argument in the proof of Theorem 3.5.

References

[1] R. B. Boppana: Amplification of Probabilistic Boolean Formulas, *Proc. 26th Annual IEEE Symp. on Foundations of Computer Science,* 1985, 20-29.

[2] J. Friedman: Probabilistic Spaces of Boolean Functions of a Given Complexity: Generalities and Random k-SAT Coefficients, Research Report CS-TR-387-92, Princeton University, 1992.

[3] A. N. Kolmogorov: *Foundations of the Theory of Probability,* Chelsea Publishing Company, New York, 1950.

[4] J. B. Paris, A. Vencovská, G. M. Wilmers: A Natural Prior Probability Distribution Derived from the Propositional Calculus, *Annals of Pure and Applied Logic* 70, 1994, 243-285.

[5] Y. Rabani, Y. Rabinovich and A. Sinclair: A Computational View of Population Genetics, *Proc. 27th Annual ACM Symp. on the Theory of Computing,* 1995, 83-92.

[6] Y. Rabinovich, A. Sinclair and A. Wigderson: Quadratic Dynamical Systems, *Proc. 33rd Annual IEEE Symp. on Foundations of Computer Science,* 1992, 304-313.

[7] A. A. Razborov: Bounded-depth Formulae over $\{\wedge, \oplus\}$ and some Combinatorial Problems. In *Complexity of Algorithms and Applied Mathematical Logic* (in Russian), Ser. *Voprosy Kibernetiky* (Problems in Cybernetics), ed.: S. I. Adian, Moscow, 1988, 149-166.

[8] P. Savický: Bent Functions and Random Boolean Formulas, *Discrete Mathematics* 147 (1995), 211-234.

[9] P. Savický: Improved Boolean Formulas for the Ramsey Graphs, *Random Structures & Algorithms 6,* 1995, 407-415.

[10] P. Savický: Complexity and Probability of some Boolean Formulas, TR 679, 1996, Institute of CS, Prague, http://www.uivt.cas.cz.

[11] L. G. Valiant: Short Monotone Formulae for the Majority Function, *J. Algorithms 5,* 1984, 363-366.

[12] A. Woods: personal communication.

[13] A. Woods: Colouring rules for finite trees and probabilities of monadic second order sentences, preprint.