And/or tree probabilities of Boolean functions

Danièle Gardy[1] and Alan Woods[2]

[1] *PRISM, CNRS UMR 8144 and Université de Versailles Saint-Quentin, 78035 Versailles Cedex, France.*
`Daniele.Gardy@prism.uvsq.fr`
[2] *School of Mathematics and Statistics, University of Western Australia, Crawley W.A. 6009, Australia.*
`woods@maths.uwa.edu.au`

We consider two probability distributions on Boolean functions defined in terms of their representations by and/or trees (or formulas). The relationships between them, and connections with the complexity of the function, are studied. New and improved bounds on these probabilities are given for a wide class of functions, with special attention being paid to the constant function $True$ and read-once functions in a fixed number of variables.

Keywords: And/Or tree, Boolean formula, tautology, tree enumeration

1 Introduction

An and/or *formula* is a Boolean formula formed from literals (variables and their negations) using binary $\wedge$ and $\vee$ connectives (and brackets). An example is

$$((\bar{x}_1 \vee x_2) \wedge \bar{x}_3) \vee (x_1 \wedge \bar{x}_3) .$$

Corresponding to the formula is a binary planar (Catalan) tree with its leaves labelled by literals and its internal nodes labelled by connectives. (In the above example the root is labelled $\vee$.) Assigning truth values at the leaves and thinking of the internal nodes as logic gates, such an and/or *tree* computes at its root the truth value of the Boolean function defined by the formula. In the example above, the function is $\bar{x}_3$, or more precisely, the same function as is defined by this much simpler formula.

We will use the terms and/or formula and and/or tree synonymously. n will denote the number of variables $x_1, \ldots, x_n$ from which the variables in the formula are to be drawn. The *size* m is the number of occurrences of literals (i.e., the number of leaves). As the tree is binary, the number of connectives (internal nodes) is $m-1$ and the total number of nodes is $2m-1$. The *complexity* $L(f)$ of a Boolean function f is the minimal size of an and/or formula defining f.

Fix n, the number of variables. One natural way to define a probability distribution on Boolean functions f is to let T_m denote the total number of and/or trees of size m, let $T_m(f)$ be the number of these which compute f, and put

$$P(f) = \lim_{m\to\infty} \frac{T_m(f)}{T_m} .$$

Lefmann and Savický [4] seem to have been the first to show explicitly that for each choice of n this limit distribution P (which depends implicitly on n) is well defined, i.e., that there is convergence for all f, and that in fact the limit $P(f)$ is always strictly positive. A rather different proof can be given using the methods of Woods [9], who established the analogous results for non-binary and/or trees which take account of the associativity and commutativity of $\wedge$ and $\vee$, and whose size is taken to be the total number of nodes.

A second natural probability distribution $\pi(f)$ on Boolean functions f is obtained by generating an and/or tree by means of a random process. Start with the root and throw a fair coin. With probability $1/2$ decide to make the root a leaf, throw a fair $2n$-sided die to decide which literal will be its label, and then stop. With probability $1/2$ make the root an internal node and then throw the coin again to decide which connective $\wedge$ or $\vee$ will be its label. Then repeat the process with each of the two "daughter" nodes in place of the root.

Technically this is a *critical Galton–Watson branching process*. With probability 1 the tree is finite. The probability $\pi(f)$ is simply defined to be the sum of the probabilities associated with those finite and/or trees that compute f. Notice that as with the limit distribution P, the π distribution depends on n. We will be interested in their asymptotic behaviour as $n \to \infty$, as well as actually calculating or estimating probabilities when n is small. In this direction, in an early, but very interesting paper (predating the work of Lefmann and Savický) Paris, Venkovská and Wilmers [7] proved among many other things that $\lim_{n\to\infty} P(f) = 0$ for the constant functions $f \in \{True,\ False\}$.

The π distribution was first studied explicitly by Chauvin, Flajolet, Gardy and Gittenberger [1]. π is definitely different from P (even asymptotically for $n \to \infty$, as we will see below). However as they found, there are some important relationships between these distributions. The extensive calculations reported in [1] led them to also make conjectures regarding the relationship between the numerical values of $\pi(f)$ and $P(f)$ for particular functions f. Some of these conjectures are settled here. We will prove that $P(f) > \pi(f)$ for $f \in \{True,\ False\}$, while on the other hand, if f is a read-once function of some fixed set of r variables then for n sufficiently large, $P(f) < \pi(f)$.

For other variants of and/or formulas and the corresponding probability distributions see [2],[1] and [8]. Analogues of $P(True)$ have also been studied for tree-like formulas involving other connectives [6, 10, 11, 3, 5]. Mostly the results are restricted to explicit small values of n. Exceptionally Moczurad, Tyszkiewicz and Zaionc [6] have shown that for formulas in n variables (without negation) having implication as the only connective, the probability of a tautology $P(True)$ lies in the interval $[(4n+1)/(2n+1)^2, (3n+1)/(n+1)^2]$. However they do not seem to address the convergence issue for general values of n. In a similar vein, Matecki [5] has studied the probability of *True* when equivalence is the only connective, obtaining results valid for all n.

Which of these various models is of most significance? Well it depends on the situation. If short formulas are of importance, the π distribution may be suitable. If the formulas are large, then P (which is, roughly speaking, π conditioned on the size m being large) is more appropriate. As noted in [6], there is a correspondence between intuitionistic implicational tautologies (without negation) and inhabited types in λ-calculus. (However not all Boolean functions can be defined using only implication and variables.) In another arena, if a close relationship with the underlying Boolean function is needed, e.g., if the real aim as in [8] is to estimate the number of Boolean *functions* defined by and/or formulas of some type, then it may be desirable to regard formulas as being "the same" if they can be converted into each other by means of the commutative and associative laws for $\wedge$ and $\vee$. And so on. A decided advantage of the particular distributions π and P considered here is that they are less complicated to analyse than some of the others, while presumably often having qualitatively similar properties.

One reason for interest in probability distributions for Boolean functions f is the suggestion (appearing in [9] for an analogue of $P(f)$) that the probability of f might be related to its complexity $L(f)$. Lefmann and Savický [4] proved that for $P(f)$ this is indeed the case. In fact for some constant $c > 0$,

$$\frac{1}{4}\left(\frac{1}{8n}\right)^{L(f)+1} \le P(f) \le (1+O(1/n))\, exp\left(-c\frac{L(f)}{n^2}\right). \tag{1}$$

where the upper bound incorporates an improvement from Chauvin, Flajolet, Gardy and Gittenberger [1]. Lefmann and Savický prove their bounds by associating the limit distribution P with a distribution on certain sets of and/or trees having an *infinite* branch. Here we will sketch an alternative proof of a sharper lower bound by using generating series (and avoiding infinite trees). As a bonus, the proof also provides an analogous lower bound for $\pi(f)$.

The plan of the paper is as follows: In Section 2, the connections between the two probability distributions and the generating functions for classes of and/or trees are recalled. These connections, which underlie the whole paper, are used in Section 3 to give improved lower bounds on $\pi(f)$ and $P(f)$ in terms of the complexity $L(f)$ and number $k(f)$ of minimal size representations of f. The main idea is to deal with a subset of the trees which compute $True$ which is both simple to describe and sufficiently large. Then in a move of particular significance for $P(f)$, the lower bounds for this set of tautologies are "transferred" to obtain lower bounds for any Boolean function f. In Section 4 we consider a variety of simple Boolean functions, comparing our lower bounds numerically with the exact values for small n, and with the Lefmann/Savický lower bound (1) when n is large. This is followed in Section 5 by comparisons between the *exact* values of the probabilities $P(f)$ and $\pi(f)$ for constant and read-once functions f. We conclude with some discussion and a conjecture in Section 6.

2 Generating functions for and/or trees

The generating function $T(z) = \sum_{m=1}^{\infty} T_m z^m$ enumerating the class $\mathcal{T}$ of all and/or trees by size m satisfies

$$T(z) = 2n\,z + 2\,T(z)^2.$$

Solving this gives

$$T(z) = \frac{1}{4}\left(1 - \sqrt{1-16nz}\right). \tag{2}$$

Expanding as a power series in z using the binomial theorem shows that the number of and/or trees of size m is

$$T_m = 2^{m-1}(2n)^m C_{m-1} \sim \frac{(16n)^m}{8m\sqrt{\pi m}},$$

where C_{m-1} is the $(m-1)$th Catalan number. Clearly $T(z)$ has radius of convergence $\rho = 1/(16n)$, and $T(\rho) = 1/4$. (More details of items in this section can be found in [1].)

Similarly for any class $\mathcal{E}$ of and/or trees, let $E(z) = \sum_{m=1}^{\infty} E_m z^m$ denote the corresponding generating series. It is easy to check that for any and/or tree τ of size m, the probability that τ is the tree generated by the Galton–Watson process described above is $2^{-2m+1}2^{-m+1}(2n)^{-m} = 4\,\rho^m$, so the definition of π can be extended to $\mathcal{E}$ by putting

$$\pi(\mathcal{E}) = 4\sum_{m=1}^{\infty} E_m \rho^m = 4\,E(\rho),$$

which always converges. In general E_m/T_m need not converge to a limit $P(\mathcal{E})$. However *if this limit does exist* it must satisfy

$$P(\mathcal{E}) = \lim_{m\to\infty} \frac{E_m}{T_m} = \lim_{z\to\rho-} \frac{E'(z)}{T'(z)}\ .$$

This follows from an easily proved *Abelian theorem* which uses only that the derivative $T'(z)$ has positive coefficients and diverges at $z = \rho$. To establish convergence, we will appeal to the following standard lemma, the idea being that (under certain conditions) if $E(z)$ has the same form of singularity at ρ as (2) then its coefficients will be asymptotic to those of $\beta_{\mathcal{E}}\, T(z)$, for some constant $\beta_{\mathcal{E}}$.

LEMMA 1 *Let $\mathcal{E}$ be a class of* and/or *trees. If the corresponding generating function $E(z)$ has on the circle $|z| = \rho$, a single dominant algebraic singularity at $\rho = 1/(16n)$, and around ρ has an expansion $E(z) = (\alpha_{\mathcal{E}} - \beta_{\mathcal{E}}\sqrt{1-16nz})/4 + o(\sqrt{1-16nz})$, then*

$$\pi(\mathcal{E}) = \alpha_{\mathcal{E}} = 4\, E(\rho)\,; \qquad P(\mathcal{E}) = \beta_{\mathcal{E}} = \lim_{z\to\rho-} \frac{E'(z)}{T'(z)}\ . \tag{3}$$

For any Boolean function f we will denote by $\mathcal{T}_f$ the class of all and/or trees which compute f. $T_f(z)$ will be the corresponding generating function. As noted in [1] (cf. [9]), on the circle $|z| = \rho$, $T_f(z)$ always has only an algebraic singularity at $\rho = 1/(16n)$, with

$$T_f(z) = \frac{1}{4}\left(\alpha_f - \beta_f\sqrt{1 - z/\rho}\right) + O(z-\rho)$$

near ρ for some constants $\alpha_f, \beta_f > 0$. So by the Lemma, $P(f)$ exists, $P(f)$ is positive, and

$$\pi(f) = \alpha_f = 4\, T_f(\rho)\,; \qquad P(f) = \beta_f = \lim_{z\to\rho-} \frac{T_f'(z)}{T'(z)}\ .$$

3 Improved lower bounds

For n variables, there is a system of 2^{2^n} quadratic equations in the generating functions $T_f(z)$ (with f ranging over all possible Boolean functions of n variables) which in principle can be solved for these 2^{2^n} generating functions. (See [1] for the details.) The underlying idea of our lower bound method is that simpler equations which are easier to solve can still give interesting bounds (instead of exact values) for the probabilities. Rather than work with the whole set $\mathcal{T}_f$ of all trees that compute f, we will work with a more easily described subset $\mathcal{E}_f \subseteq \mathcal{T}_f$ obtaining lower bounds on $\pi(f)$ and (provided $P(\mathcal{E}_f)$ exists) on $P(f)$.

Let us begin by considering the set of all and/or trees, and a *proper* subset $\mathcal{E}_{True} \subset \mathcal{T}_{True}$. So $\mathcal{E}_{True}$ is a set of *some* of the and/or trees that compute $True$. $\mathcal{E}_{True}$ is defined (in obvious notation) by

$$\begin{aligned}\mathcal{E}_{True} \quad = \quad & \oplus_{1\le i\le n}(\vee, x_i, \bar{x}_i) \,\oplus\, ... \oplus_{1\le i\le n}(\vee, \bar{x}_i, x_i) \,\oplus\, (\wedge, \mathcal{E}_{True}, \mathcal{E}_{True}) \oplus (\vee, \mathcal{E}_{True}, \mathcal{E}_{True}) \\ & \oplus (\vee, \mathcal{E}_{True}, \mathcal{T}\setminus\mathcal{E}_{True}) \oplus (\vee, \mathcal{T}\setminus\mathcal{E}_{True}, \mathcal{E}_{True}).\end{aligned}$$

A symmetrical equation defines a set $\mathcal{E}_{False}$ consisting of some of the trees that compute $False$. Now let $E_{True}(z)$ be the generating function that enumerates the set $\mathcal{E}_{True}$. It satisfies the following equation, in which $T(z)$ is the function enumerating all and/or trees on n literals:

$$E_{True}(z) = 2nz^2 + 2E_{True}(z)T(z).$$

We obtain $E_{True}(z) = (2nz^2)/(1-2T(z)) = zT(z)$; hence

$$E_{True}(z) = \frac{\rho(1-\sqrt{1-16nz})}{4} + O(z-\rho)$$

for z near ρ. Using Lemma 1 (the conditions for which are clearly satisfied) we can read off $\pi(\mathcal{E}_{True}) = \rho$, $P(\mathcal{E}_{True}) = \rho$. Recalling that $\rho = 1/(16n)$, these give the common lower bound:

THEOREM 2 $\pi(True) \ge \dfrac{1}{16n}\,; \qquad P(True) \ge \dfrac{1}{16n}\ .$

Of course the same bounds apply to $False$. Notice also that as $E_{True}(z) = z\,T(z)$, we even get a lower bound on the number $T_m(True)$ of trees of size m which compute $True$, namely

$$T_m(True) \ \ge\ 2^{m-2}(2n)^{m-1}C_{m-2} \quad \text{where } C_{m-2} \text{ is a Catalan number.}$$

Now define a subset $\mathcal{E}_x$ of the trees that compute the literal x by

$$\begin{aligned}\mathcal{E}_x \quad = \quad & \{x\} \oplus (\vee, \mathcal{E}_x, \mathcal{E}_x) \oplus (\wedge, \mathcal{E}_x, \mathcal{E}_x) \oplus (\wedge, \mathcal{E}_{True}, \mathcal{E}_x) \oplus (\wedge, \mathcal{E}_x, \mathcal{E}_{True}) \\ & \oplus (\vee, \mathcal{E}_{False}, \mathcal{E}_x) \oplus (\vee, \mathcal{E}_x, \mathcal{E}_{False}).\end{aligned}$$

The generating function $E_x(z)$ for this set satisfies the equation

$$E_x(z) = z + 2E_x(z)^2 + 4E_x(z)E_{True}(z),$$

which gives

$$E_x(z) = \frac{1}{4}\left(1 - z + z\sqrt{1-16nz} - \sqrt{1 - 10z + 2z^2 - 16nz^3 + 2z(1-z)\sqrt{1-16nz}}\right).$$

Expanding $E_x(z)$ near its singularity $\rho = 1/(16n)$ gives

$$E_x(z) = \frac{1}{4}\left(\alpha_x - \beta_x\sqrt{1-16nz}\right) + O(1-16nz),$$

with $\alpha_x = (16n - 1 - \sqrt{\eta})/(16n)$ and $\beta_x = \alpha_x/\sqrt{\eta} = \rho\alpha_x/\sqrt{1-10\rho+\rho^2}$, where $\eta = 256n^2 - 160n + 1$. Hence

$$\pi(x) > \frac{16n-1-\sqrt{\eta}}{16n} = \frac{1}{4n} + \frac{3}{64n^2} + O(1/n^3); \qquad P(x) \geq \frac{16n-1-\sqrt{\eta}}{16n\sqrt{\eta}} = \frac{1}{64n^2} + \frac{1}{128n^3} + O(1/n^4).$$

What we have just done for literals can be mimicked for any Boolean function $f \notin \{True, False\}$. Let us consider a Boolean function $f \notin \{True, False\}$, let $L(f)$ be its complexity (i.e., the number of leaves in the trees of smallest size representing f), $\mathcal{M}(f)$ be the set of such trees of minimal complexity, and $k(f) = |\mathcal{M}(f)|$ the number of these trees. Next define a subset $\mathcal{E}_f$ of the trees that compute f by

$$\begin{aligned}\mathcal{E}_f \quad = \quad & \mathcal{M}(f) \oplus (\wedge, \mathcal{E}_f, \mathcal{E}_f) \oplus (\vee, \mathcal{E}_f, \mathcal{E}_f) \oplus (\wedge, \mathcal{E}_f, \mathcal{E}_{True}) \\ & \oplus (\wedge, \mathcal{E}_{True}, \mathcal{E}_f) \oplus (\vee, \mathcal{E}_f, \mathcal{E}_{False}) \oplus (\vee, \mathcal{E}_{False}, \mathcal{E}_f).\end{aligned}$$

The generating function $E_f(z)$ of $\mathcal{E}_f$ satisfies

$$E_f(z) = k(f)\, z^{L(f)} + 2\, E_f^2(z) + 4\, E_f(z)\, E_{True}(z)\ .$$

Using the form of $E_{True}(z)$ found above, it can be checked that $E_f(z)$ has only one dominant singularity on $|z| = \rho$, namely at ρ, and that this singularity is algebraic. (We omit the details.) Expanding $E_f(z)$ around ρ, we get

$$E_f(z) = \frac{1}{4}\left(\alpha_f - \beta_f\sqrt{1 - z/\rho}\right) + O(1 - z/\rho),$$

where, setting $\mu(f) = 8k(f)\rho^{L(f)}/(1-\rho)^2$, we have

$$\alpha_f = (1-\rho)\left(1 - \sqrt{1-\mu(f)}\right); \qquad \beta_f = \rho\left(\frac{1}{\sqrt{1-\mu(f)}} - 1\right).$$

Finally we apply Lemma 1 to get lower bounds for the probabilities $P(f)$ and $\pi(f)$:

THEOREM 3 *For any non-constant Boolean function f, if $L(f)$ is the complexity of f and $k(f)$ is the number of trees of minimal size $L(f)$ that compute f, then*

$$\pi(f) \geq (1-\rho)\left(1 - \sqrt{1-\mu(f)}\right); \qquad P(f) \geq \rho\left(\frac{1}{\sqrt{1-\mu(f)}} - 1\right),$$

where $\rho = 1/(16n)$ and

$$\mu(f) := \frac{8\ k(f)\ \rho^{L(f)}}{(1-\rho)^2}.$$

From this Theorem, we can obtain weaker bounds, easier to compute, but (in the form involving $k(f)$) asymptotically equivalent for large n.

COROLLARY 4 $\pi(f) \geq \dfrac{4k(f)}{(16n)^{L(f)}} \geq \dfrac{2}{(8n)^{L(f)}}; \qquad P(f) \geq \dfrac{4k(f)}{(16n)^{L(f)+1}} \geq \dfrac{1}{(8n)^{L(f)+1}}.$

Here we have used the inequality $k(f) \geq 2^{L(f)-1}$. This is related to the "folklore" fact that minimal and/or trees for f are rigid, and can be proved by induction on $L(f)$. The case $L(f) = 1$ is trivial. If $L(f) > 1$, observe that in a tree representation of f of minimal size, the root has two daughters computing f_1 and f_2, say. Either $f = f_1 \vee f_2$ or $f = f_1 \wedge f_2$. Notice that $f_1 \neq f_2$. For if $f_1 = f_2$ then $f = f_1 = f_2$ and the representation of f cannot be minimal. Clearly the two daughters must also be of the minimal sizes $L(f_1)$ and $L(f_2)$, and $L(f) = L(f_1) + L(f_2)$. By the induction hypothesis, $k(f_1) \geq 2^{L(f_1)-1}$ and $k(f_2) \geq 2^{L(f_2)-1}$ giving $2^{L(f_1)-1}\,2^{L(f_2)-1}$ distinct minimal trees computing f. In each case we can exchange the daughter subtrees without modifying the function computed. As $f_1 \neq f_2$ the representations of f resulting from doing this are all *different*, so

$$k(f)\ \geq\ 2\, 2^{L(f_1)-1}\, 2^{L(f_2)-1} = 2^{L(f_1)+L(f_2)-1} = 2^{L(f)-1}\ .$$

If we know $k(f)$, or a better lower bound on $k(f)$, we may get a substantial improvment on the bound of Lefmann and Savický for $P(f)$. (See the numerical results below.) Even if we do not know $k(f)$, we still get at least four times their bound.

4 Numerical results

For several Boolean functions, we will compare our lower bound for $P(f)$ with that of Lefmann and Savický, and numerical values of our best lower bounds with the exact values for $n \leq 3$.

- For the constants $True$ and $False$, π and P are greater than $1/(16n)$, which is much better than Lefmann and Savický's bound of $1/(2\,048n^3)$.

	$\pi(True)$	Lower bound on $\pi(True)$	$P(True)$	Lower bound on $P(True)$
$n=1$	0.1339	0.0625	0.2886	0.0625
$n=2$	0.08642	0.03125	0.2094	0.03125
$n=3$	0.0642	0.015625	0.165	0.015625

- For a literal x, $k(x) = L(x) = 1$, and Lefmann and Savický's bound on $P(x)$ is $1/(256n^2)$. Our lower bound on $P(x)$ is

$$P(x) \geq \frac{1}{16n}\left(\frac{1}{\sqrt{1-\frac{1}{2n(1-1/(16n))^2}}}-1\right) \sim \frac{1}{64n^2}.$$

The lower bound on $\pi(x)$ is

$$\pi(x) \geq \left(1-\frac{1}{16n}\right)\left(1-\sqrt{1-\frac{1}{2n(1-1/(16n))^2}}\right) \sim \frac{1}{4n}.$$

Let us see how these bounds compare with the actual values for $n \leq 3$:

	$\pi(x)$	Lower bound on $\pi(x)$	$P(x)$	Lower bound on $P(x)$
$n=1$	0.3660	0.3219	0.2113	0.03268
$n=2$	0.1595	0.1390	0.06717	0.005235
$n=3$	0.0994	0.08916	0.0314	0.002087

- For the functions $l_1 \wedge l_2$ or $l_1 \vee l_2$ (for literals $l_1 \neq l_2, \bar{l}_2$), we have that $L(f) = 2 = k(f)$. Lefmann and Savický's bound on $P(l_1 \wedge l_2)$ is $1/(2\,048\,n^3)$. Our lower bound is

$$P(l_1 \wedge l_2) \geq \frac{1}{16n}\left(\frac{1}{\sqrt{1-\frac{1}{16n^2(1-1/(16n))^2}}}-1\right) \sim \frac{1}{512n^3},$$

and the lower bound on $\pi(l_1 \wedge l_2)$ is

$$\pi(l_1 \wedge l_2) \geq \left(1-\frac{1}{16n}\right)\left(1-\sqrt{1-\frac{1}{16n^2(1-1/(16n))^2}}\right) \sim \frac{1}{32n^2}.$$

Again we compare these lower bounds with the actual values for $n = 2, 3$:

	$\pi(l_1 \wedge l_2)$	L. B. on $\pi(l_1 \wedge l_2)$	$P(l_1 \wedge l_2)$	L. B. on $P(l_1 \wedge l_2)$
$n=2$	0.02345	0.008098	0.03848	0.0002634
$n=3$	0.00776	0.00355	0.00995	$0.7586\,10^{-4}$

- For a function $l_1 \wedge l_2 \wedge l_3$ (with l_1, l_2, l_3 literals in distinct variables), $L(f) = 3$ and $k(f) = 12$. Lefmann and Savický's bound on $P(l_1 \wedge l_2 \wedge l_3)$ is $1/(16\,384\,n^4)$. Our lower bound is now

$$P(l_1 \wedge l_2 \wedge l_3) \geq \frac{1}{16n}\left(\frac{1}{\sqrt{1-\frac{3}{128n^3(1-1/(16n))^2}}}-1\right) \sim \frac{3}{4096n^4},$$

and the lower bound on $\pi(l_1 \wedge l_2 \wedge l_3)$ is

$$\pi(l_1 \wedge l_2 \wedge l_3) \geq \left(1-\frac{1}{16n}\right)\left(1-\sqrt{1-\frac{3}{128n^3(1-1/(16n))^2}}\right) \sim \frac{3}{256n^3},$$

The exact values for $n = 3$ are:

	$\pi(l_1 \wedge l_2 \wedge l_3)$	L. B. on π	$P(l_1 \wedge l_2 \wedge l_3)$	L. B. on P
$n=3$	0.00282	0.0004433	0.00768	$0.943\,10^{-5}$

- For a function $l_1 \wedge (l_2 \vee l_3)$ (with l_1, l_2, l_3 literals in distinct variables), of similar complexity $L(f) = 3$ but smaller $k(f) = 4$, $\pi(f) \geq 1/(256n^3)$. Lefmann and Savický's bound on $P(l_1 \wedge (l_2 \vee l_3))$ is $1/(16\,384\, n^4)$, i.e. the same as for the functions of the type $l_1 \wedge l_2 \wedge l_3$. Our lower bound is

$$P(l_1 \wedge (l_2 \vee l_3)) \geq \frac{1}{16n} \left(\frac{1}{\sqrt{1 - \frac{1}{128n^3(1-1/(16n))^2}}} - 1 \right) \sim \frac{1}{4096n^4},$$

and the lower bound on $\pi(l_1 \wedge (l_2 \vee l_3))$ is

$$\left(1 - \frac{1}{16n}\right) \left(1 - \sqrt{1 - \frac{1}{128n^3(1-1/(16n))^2}}\right) \sim \frac{1}{256n^3}.$$

We check the lower bounds against the exact values for $n = 3$:

	$\pi(l_1 \wedge (l_2 \vee l_3))$	L.B. on π	$P(l_1 \wedge (l_2 \vee l_3))$	L. B. on P
$n = 3$	0.000817	0.0001477	0.00211	$0.3144\, 10^{-5}$

- For the function $f = x_1 \, xor \, x_2$, $L(f) = 4$ and $k(f) = 16$; we basically have two minimal representations: $(x_1 \wedge \bar{x}_2) \vee (\bar{x}_1 \wedge x_2)$ and $(x_1 \vee x_2) \wedge (\bar{x}_1 \vee \bar{x}_2)$, and each representation gives eight different trees. This gives:

  1. For $n = 2$: the lower bound on π is $0.630\ 10^{-4}$ and the lower bound on P is $0.203\ 10^{-5}$ (the actual values are 0.000635 for π and $0.00229...$ for P).
  2. For $n = 3$: the lower bound on π is $0.123\, 10^{-4}$ and the lower bound on P is $0.261\, 10^{-6}$ (the actual values are $0.635\, 10^{-3}$ for π and $0.192\, 10^{-3}$ for P).
  3. For large n, $\pi(x_1 \, xor \, x_2) \geq 1/(1\,024n^4)$ and $P(x_1 \, xor \, x_2) \geq 1/(16\,384n^5)$.

All these numerical computations show that the lower bounds for $P(f)$ are quite far from the actual values of the probabilities, when we know them! For $\pi(f)$ the gap is not quite so large, perhaps hinting at the major contribution of trees of the minimal size $L(f)$ to both $\pi(f)$ and our lower bound.

5 Comparison of $P(f)$ and $\pi(f)$

We will now compare the probabilities $P(f)$ and $\pi(f)$ for some particular Boolean functions f. If S is a set of Boolean functions, write $P(S) = \sum_{f \in S} P(f)$.

LEMMA 5 (PARIS, VENCOVSKÁ AND WILMERS [7]) *Fix k in the interval $0 \leq k \leq 1$. Let $S(k)$ be the set of all Boolean functions $f : \{True, False\}^n \to \{True, False\}$ such that $2^{-n} |\{\mathbf{x} \in \{True, False\}^n : f(\mathbf{x}) = True\}| = k$. Then $P(S(k)) \to 0$ as $n \to \infty$.*

Let f be a function of $x_1, x_2, \ldots, x_r$. Considering f to be a function of $x_1, x_2, \ldots, x_n$ which does not depend on the variables $x_{r+1}, x_{r+2}, \ldots, x_n$, the probabilities $P(f)$ and $\pi(f)$ make sense for all $n \geq r$.

THEOREM 6 *Suppose that r is fixed and $f(x_1, x_2, \ldots, x_r)$ is any Boolean function that depends essentially on all of the r variables $x_1, x_2, \ldots, x_r$. Then $P(f) = o(n^{-r})$ as $n \to \infty$.*

PROOF: As the function f depends essentially on all of $x_1, x_2, \ldots, x_r$, distinct choices of $1 \leq i_1 < i_2 < \cdots < i_r \leq n$ correspond to *distinct* functions $f(x_{i_1}, x_{i_2}, \ldots, x_{i_r})$. Let S be the set of all such functions. Clearly,

$$P(S) = \binom{n}{r} P(f)$$

and $S \subseteq S(k)$ for the fixed real number $k = 2^{-r} |\{\mathbf{x} \in \{True, False\}^r : f(\mathbf{x}) = True\}|$. Applying Lemma 5 shows that $P(S) \leq P(S(k)) = o(1)$. Consequently,

$$P(f) = P(S) \Big/ \binom{n}{r} = o(n^{-r}) . \qquad \square$$

An and/or formula is *read–once* if each variable appears at most once (possibly negated). It is well known (see e.g. [8]) that the function defined by a read–once formula depends essentially on all the variables appearing. A Boolean function is *read–once* if there is some read–once and/or formula which defines it.

THEOREM 7 *Fix r and suppose that $f(x_1, x_2, \ldots, x_r)$ is any read–once Boolean function of r variables. Then $\lim_{n\to\infty} \frac{P(f)}{\pi(f)} = 0$ so certainly $P(f) < \pi(f)$ once n is sufficiently large.*

PROOF: We can assume that f depends essentially on all of the variables $x_1, x_2, \ldots, x_r$. By Theorem 6, it is only necessary to show that $\pi(f) \geq c_r n^{-r}$ for some constant $c_r > 0$. However we saw in Section 3 that $\pi(f) \geq 2\,(8n)^{-L(f)}$, and as $L(f) = r$ for the read–once function f, this lower bound is indeed of the form $c_r n^{-r}$. □

For example, $P(x_1) < \pi(x_1)$, $P(x_1 \vee x_2) < \pi(x_1 \vee x_2)$ and $P(\bar{x}_1 \vee (\bar{x}_2 \wedge x_3)) < \pi(\bar{x}_1 \vee (\bar{x}_2 \wedge x_3))$ once n is large enough.

We now return to considering the probability that an and/or formula is a tautology.

THEOREM 8 *$P(True) > \pi(True)$ for all n.*

PROOF: As before, $\mathcal{T}$, $\mathcal{T}_x$ and $\mathcal{T}_{True}$ will denote respectively the class of all and/or trees, the class of all trees computing the literal x, and the class of all trees computing the constant function *True*. The corresponding generating functions are $T(z)$, $T_x(z)$ and $T_{True}(z)$. Consider the class

$$\begin{array}{rcll} \mathcal{G} & = & \oplus_{1\leq i\leq n}(\vee, \mathcal{T}_{x_i}, \mathcal{T}_{\bar{x}_i}) & \oplus \ldots \quad \oplus_{1\leq i\leq n}(\vee, \mathcal{T}_{\bar{x}_i}, \mathcal{T}_{x_i}) \\ & \oplus & (\wedge, \mathcal{T}_{True}, \mathcal{T}_{True}) & \oplus \quad (\vee, \mathcal{T}_{True}, \mathcal{T}_{True}) \\ & \oplus & (\vee, \mathcal{T} \setminus \mathcal{T}_{True}, \mathcal{T}_{True}) & \oplus \quad (\vee, \mathcal{T}_{True}, \mathcal{T} \setminus \mathcal{T}_{True}). \end{array}$$

Clearly $\mathcal{G} \subseteq \mathcal{T}_{True}$. The generating series $G(z)$ for $\mathcal{G}$ is given by

$$G(z) = 2n\, T_x(z)^2 + 2\, T_{True}(z)\, T(z)\ .$$

Each of the functions $T(z)$, $T_{True}(z)$ and $T_x(z)$ on the right has radius of convergence $\rho = 1/(16n)$ and on their circle of convergence only an algebraic singularity at $z = \rho$, so the same is clearly true of $G(z)$. Similarly, for z near ρ, $G(z) \approx \left(\alpha - \beta\sqrt{1 - z/\rho}\right)/4$ for some positive constants α, β. Using Lemma 1 we see that $P(True) \geq P(\mathcal{G}) = \lim_{z\to\rho-}(G'(z)/T'(z))$. Now

$$G'(z) = 4n\, T_x(z)\, T_x'(z) + 2\, T_{True}'(z)\, T(z) + 2\, T_{True}(z)\, T'(z)$$

and dividing by $T'(z)$ gives

$$\begin{aligned} P(True) \;\geq\; \lim_{z\to\rho-} \frac{G'(z)}{T'(z)} &= 4n\, T_x(\rho) \lim_{z\to\rho-} \frac{T_x'(z)}{T'(z)} \;+\; 2\, T(\rho) \lim_{z\to\rho-} \frac{T_{True}'(z)}{T'(z)} \;+\; 2\, T_{True}(\rho) \\ &= n\, \pi(x) P(x) \;+\; \frac{1}{2}\, P(True) \;+\; \frac{1}{2}\, \pi(True)\ . \end{aligned}$$

So $P(True) \geq 2n\, \pi(x) P(x) + \pi(True) > \pi(True)$, as the probabilities $\pi(x)$ and $P(x)$ of computing the literal function x are strictly positive for all n. □

6 Final remarks

Notice that in the case of read–once functions f of r variables, with r fixed, the lower bound $\pi(f) \geq c_r n^{-r}$ from Section 3 differs from the trivial upper bound

$$\pi(f) \;\leq\; 1 \Big/ \binom{n}{r}$$

proved similarly to Theorem 6, by only a constant factor. (The constant depends on r). For $P(f)$ the agreement is not quite so good, the lower bound from Section 3 differing from the upper bound in Theorem 6 by a factor of order $o(n)$.

CONJECTURE 1 *Suppose that $f(x_1, x_2, \ldots, x_r)$ is a read–once Boolean function of r variables, with r fixed. Then there exist constants b_f and B_f such that $\pi(f) \sim b_f\, n^{-r}$ and $P(f) \sim B_f\, n^{-r-1}$ as $n \to \infty$.*

The conjecture asserts that, aside from constant factors depending on f, the lower bounds in Corollary 4 give correct asymptotic formulas when f is a read-once function. All the examples given in Section 4, apart from the first and last, are read-once functions. So in particular, the asymptotic formulas for them should look like the computed asymptotic lower bounds except for the values of the constant factors.

Random generation of and/or trees, for $n = 2$, has been simulated by F. Quessette and D. Villa Moreira. This was done for two variables x_1 and x_2, and e.g. the number of internal nodes equal to 1000, with 10^6 trees generated at random. The random generation algorithm is as follows: first a random binary tree is generated, using Remy's algorithm, then a random labelling of internal nodes and of leaves takes place.

This simulation gave good agreement with the calculated values of the probabilities $P(f)$, for the 16 Boolean functions considered. We then computed, for each Boolean function, the following parameters: height, width, number of occurrences of $\wedge$, number of occurrences of a specified literal. Simulations appear to indicate that in each case, these parameters follow *the same* Gaussian limiting distribution whatever the Boolean function.

Acknowledgements

The authors are pleased to thank B. Chauvin for many fruitful discussions.

References

[1] Brigitte Chauvin, Philippe Flajolet, Daniele Gardy, and Bernhard Gittenberger. And/Or trees revisited. *Combinatorics, Probability and Computing*, 13(4-5):475–497, 2004.

[2] Brigitte Chauvin, Daniele Gardy, and Alan Woods. *Commutative and associative tree representations for Boolean functions*, 2005. Technical report, U.V.S.Q., to appear.

[3] Z. Kostrzycka and M. Zaionc. Statistics of intuitionnistic versus classical logic. *Studia Logica*, 76(3), 2004.

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

[5] G. Matecki. Asymptotic density for equivalence. *Electronic Notes in Theoretical Computer Science*. to appear.

[6] M. Moczurad, J. Tyszkiewicz, and M. Zaionc. Statistical properties of simple types. *Mathematical Structures in Computer Science*, 10(5):575–597, 2000.

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

[8] P. Savický and A. Woods. The number of Boolean functions computed by formulas of a given size. *Random Structures and Algorithms*, 13:349–382, 1998.

[9] Alan Woods. Coloring rules for finite trees, and probabilities of monadic second order sentences. *Random Structures and Algorithms*, 10, 1997.

[10] M. Zaionc. Statistics of implicational logic. *Electronic Notes in Theoretical Computer Science*, 84, 2003.

[11] M. Zaionc. On the asymptotic density of tautologies in logic of implication and negation. *Reports on Mathematical Logic*, 38, 2004.