Improved Asymptotic Formulas for Counting Correlation-Immune Boolean Functions

Eric Bach

Technical Report #1616

October 2007

Improved Asymptotic Formulas for Counting Correlation-Immune Boolean Functions

Eric Bach
Computer Sciences Dept.
University of Wisconsin
Madison, WI 53706

October 11, 2007

Abstract

A Boolean function is called correlation immune if every input is independent of the output, when the inputs are chosen from a uniform distribution. Such functions are of interest in machine learning and stream cipher design. We show how an asymptotic formula of Denisov, which approximately counts the n-variable correlation immune functions, can be improved so as to be accurate even for fairly small n. Such information is useful to designers of machine learning algorithms, as it indicates how often a greedy algorithm for learning decision trees will fail.

Technical Report 1616, Computer Sciences Dept., University of Wisconsin.

1 Introduction.

Let f be a Boolean function with n inputs and one output. Suppose we choose the n inputs at random, uniformly from the 2^n available choices. We call f *correlation immune* if, for all i, the random variables x_i and $y = f(x_1, \ldots, x_n)$ are independent.

There is an equivalent "physical" definition. Any Boolean function f can be viewed as a placement of of 0's and 1's on the vertices of an n-dimensional hypercube. Then, f is correlation immune if the center of mass for this placement is at the center of the hypercube. For example, when $n = 2$, consideration of the diagrams

$$\begin{array}{ccc} 0 & -- & 1 \\ | & & | \\ | & & | \\ 1 & -- & 0 \end{array} \qquad\qquad \begin{array}{ccc} 0 & -- & 1 \\ | & & | \\ | & & | \\ 1 & -- & 1 \end{array}$$

reveals that exclusive-or is correlation immune, whereas the ordinary or operation is not.

The term "correlation immunity" can be justified as follows. Since an input x_i and the output y each take at most two values, their independence is equivalent to the vanishing of their correlation coefficient. Also, this vanishing is unaffected by linear transformations, so there is no harm in assuming that the outputs of f are ± 1. In the remainder of this paper, we will, therefore, assume that $f : \{0, 1\}^n \to \{\pm 1\}$.

In this paper we focus on the following counting problem. Let B_n be the number of Boolean functions of n variables that are correlation immune. How large is B_n? Using analytic methods, Denisov [4] found an asymptotic formula for B_n, but said nothing about how well it approximates the true value. We will refine his formula so that it becomes accurate for small values of n, and provide rigorous bounds on the error in our approximations.

Correlation immune functions are of interest in machine learning [9], for the following reason. A simple "greedy" method for inferring a decision tree representation of a Boolean function from a sample of input-output pairs works by choosing the root of the tree to maximize some numerical score (information gain), and then proceeding recursively. If the function is correlation immune, the expectation of this numerical score is zero, and the method cannot distinguish between relevant and irrelevant variables. Having good answers to the counting problem, then, allows the designer of a machine learning procedure to know how many functions require more sophisticated methods.

Correlation immune functions are also useful in stream cipher cryptography [11, 12], as functions for combining two or more pseudo-random streams into a more complex keystream. They have also been tied coding theory [2]. For this reason, a number of authors have emphasized techniques for constructing them. For a survey of their properties from this point of view, and some results on combinatorial approaches to the counting problem, see [10].

We will also discuss a related question that is of interest in cryptography. What is the probability that a function is both correlation immune and balanced, in the sense that a random input produces a random bit as output? Such functions have been called *1-resilient* [3].

2 Previous Work

Let us first review Denisov's results. Recall that the Fourier coefficients of a function $f : \{0,1\}^n \to \{\pm 1\}$ are, for $z \in \{0,1\}^n$,

$$\hat{f}(z) = \sum_x f(x)(-1)^{z \cdot x}.$$

From this formula it follows that $\hat{f}(z)$ is always an integer. Then, since exactly n of the z's have Hamming weight 1,

$$Z := (\hat{f}(z))_{||z||=1}$$

is an element of $\mathbf{Z}^n$. Xiao and Massey [13] observed that f is correlation immune precisely when $Z = 0$.

Suppose that f is uniformly distributed over the 2^{2^n} possible functions. Then Z is a random variable, whose characteristic function $E[e^{i(t,Z)}]$ is

$$F_Z(t) = \prod_\sigma \cos(t^\sigma).$$

In this formula,

$$t^\sigma = \sigma_1 t_1 + \cdots + \sigma_n t_n,$$

where $\sigma \in \{\pm 1\}^n$ is a sign pattern, and we take the product over the 2^n possible such patterns. Applying Fourier inversion to the characteristic function of Z, one obtains

$$\Pr[Z = 0] = \frac{1}{(2\pi)^n} \int_{[-\pi,\pi]^n} F_Z(t)dt.$$

The function F_Z has a local maximum at $t = 0$, and by using a Laplace approximation around this point, Denisov found the limiting formula

$$D_n := \frac{1}{2} \left(\frac{8}{\pi}\right)^{n/2} 2^{-n^2/2} \cdot 2^{2^n}. \tag{1}$$

for the number of n-variable correlation immune functions. In particular, he showed that $B_n/D_n \to 1$ as $n \to \infty$.

The idea behind the Laplace approximation is the following. If we expand each cosine in a Taylor series, terms of odd degree in the t_i's will cancel, and we are left with

$$\prod_\sigma \cos(t^\sigma) = 1 - \frac{2^n \sum_{i=1}^n t_i^2}{2} + \cdots$$

This is also the Taylor expansion of the Gaussian

$$G(t) = e^{-2^n ||t||^2/2},$$

so the integral around 0 should be to be close to the integral of G over all of $\mathbf{R}^n$. If we include the idea that F_Z has multiple peaks, coming from the periodicity of the cosine, we get (1). (Actually, Denisov worked with a cosine product of $n+1$ variables; we have simplified his argument a bit.)

Denisov's formula has an interesting probabilistic interpretation. For a random Boolean function, each Fourier coefficient is the result of a 2^n-step random walk. The probability that such a walk returns to its origin is

$$\frac{\binom{2^n}{2^n/2}}{2^{2^n}} \sim \left(\frac{2}{\pi}\right)^{1/2} 2^{-n/2},$$

by Stirling's formula. If all the walks were independent, we would have a smaller result than (1). Thus, the vanishing of one Fourier coefficient "encourages" the others to vanish as well.

The main idea behind our improvement of Denisov's result is to use more terms in the Taylor expansion of F_Z. Of course, such enhancements will be meaningful only if the integrand falls off sharply enough away from its peaks, and most of our effort will go into verifying this.

3 Notation.

We now give some notation and definitions we will need later.

Unless noted otherwise, $||t||$ means the 2-norm of the vector $t \in \mathbf{R}^n$. Quantities estimated by big-O notation are allowed to be negative.

Let

$$L = \{(a_1\pi/2, \ldots, a_n\pi/2) : a_i \in \mathbf{Z}, \sum a_i \text{ even } \}.$$

This is a period lattice for F_Z, in the sense that $F_Z(t+v) = F_Z(t)$ for all $v \in L$. As a fundamental domain for L, we take the box

$$F = [-\pi/4, \pi/4]^{n-1} \times [-\pi/2, \pi/2].$$

Let $C = \{t \in \mathbf{R}^n : ||t||_1 \le \frac{\pi}{2}\}$. Note that $t \in C$ iff t^σ is. Also, the boundary of C is formed from the 2^n hyperplanes $t^\sigma = \frac{\pi}{2}$.

If v, w are distinct elements of L, we have $||v - w||_1 \ge \pi$. From this it follows that the sets $v + C$, for $v \in L$, have no interior points in common.

If we identify points whose coordinates agree mod 2π, the box $[-\pi, \pi]^n$ becomes a union of $4^n/2$ copies of F. Since F_Z is periodic, we then have

$$B_n = 2^{2^n} \frac{1}{(2\pi)^n} \int_{[-\pi,\pi]^n} F_Z(t) dt = 2^{2^n+2n-1} \frac{1}{(2\pi)^n} \int_F F_Z(t) dt. \qquad (2)$$

(Note that $2^{2^n} \frac{4^n}{2} = 2^{2^n+2n-1}$.)

4 Improved Analytic Bounds, With Error Estimates.

4.1 Bounds for Cosine Products.

For any t, we have $0 \le F_Z(t) \le 1$. The upper bound follows from the definition; to prove the lower bound, split the product in two, according to the sign of t_1. Since cosine is even, these parts have the same value.

Lemma 4.1 *If $t \in F \backslash C$,*

$$\prod_\sigma \cos(t^\sigma) \le \cos\left(\frac{\pi}{2n+1}\right)^{2^{n-1}}.$$

Proof. By the definition of F,

$$|t_1|, \ldots, |t_{n-1}| \leq \pi/4, \qquad\qquad |t_n| \leq \pi/2.$$

Let i be the least index such that $|t_i| > \pi/(2n)$; we know this exists because $t \notin C$. We claim there is an index j with the following property: for any σ, if we flip the sign of t_j and get t^τ, then at least one of $t^\sigma \bmod \pi$, $t^\tau \bmod \pi$, will be $\geq \pi/(2n+1)$ in absolute value. (We use symmetric residues.) The lemma follows easily from this claim.

We now prove the claim. If $i \leq n-1$, we may take $j = i$; in this case the claim follows from the inequalities $\pi/(2n+1) < |t_j| \leq \pi/4$.

Suppose, then, that $i = n$. Without loss of generality, $t_n > 0$. If $t_n < \pi/2 - \pi/(4n+2)$, we can take $j = n$. Otherwise, suppose that for some ρ, we have $|t^\rho \bmod \pi| \leq \pi/(2n+1)$. (If not, we are done.) Since t_n is at least $\pi/2 - \pi/(4n+2)$, this cancellation is only possible if

$$\sum_{k=1}^{n-1} \rho_k t_k \geq \frac{\pi}{2} - \frac{3\pi}{2(2n+1)}.$$

Then, choose any j such that

$$\rho_j t_j \geq \frac{1}{n-1}\left[\frac{\pi}{2} - \frac{3\pi}{2(2n+1)}\right] = \frac{\pi}{2n+1}.$$

Then, $\pi/(2n+1) \leq |t_j| \leq \pi/4$, as needed. ■

Corollary 4.1 *For $t \in F \backslash C$, we have*

$$\prod_\sigma \cos(t^\sigma) \leq e^{-2^{n+1}/(2n+1)^2}, \tag{3}$$

Proof. Since $\cos(x) \leq 1 - 4x^2/\pi^2$ for $|x| \leq \pi/2$, we have

$$\cos\left(\frac{\pi}{2n+1}\right)^{2^{n-1}} \leq \left(1 - \frac{4}{(2n+1)^2}\right)^{2^{n-1}}.$$

The result then follows, since $(1 - y/m)^m \leq e^{-y}$ when $y > 0$. ■

Our next result deals with the behavior of F_Z on C.

Lemma 4.2 *For all $t \in C$,*

$$F_Z(t) \leq G(t). \tag{4}$$

Proof. On the boundary of C, F_Z vanishes, so we need only consider interior points. For $|z| < \pi/2$ we have the convergent Taylor series

$$\log \cos z = -\frac{z^2}{2} - \frac{z^4}{12} - \frac{z^6}{45} - \cdots \ .$$

All coefficients are negative. (See Abramowitz and Stegun [1], 4.3.72 for a formula for these, involving Bernoulli numbers.) Substituting $z = t^\sigma$ and summing over all σ, we get

$$\sum_\sigma \log \cos t^\sigma = -\frac{2^n||t||^2}{2} - \frac{T_4}{12} - \frac{T_6}{45} - \cdots \ , \tag{5}$$

where $T_{2k} = \sum_\sigma (t^\sigma)^{2k}$. Exponentiating both sides gives the result. ■

4.2 A Tail Bound for the Gaussian.

Let

$$\Gamma(\alpha, z) := \int_z^\infty u^{\alpha-1} e^{-u} du. \tag{6}$$

This is the incomplete Gamma function; see Abramowitz and Stegun [1].

Lemma 4.3 *Let $a \geq 0$. Then*

$$I_a := \frac{1}{(2\pi)^n} \int_{||t|| \geq a} e^{-2^n||t||^2/2} dt = \frac{2^{-n^2/2}}{(2\pi)^{n/2}} \frac{\Gamma(n/2, 2^{n-1}a^2)}{\Gamma(n/2)}.$$

Proof. By using spherical coordinates and integrating out the angles (as explained by Fleming [5], p. 181), we find

$$I_a := \frac{1}{(2\pi)^n} \int_{||t|| \geq a} e^{-||t||^2/(2s^2)} dt = \frac{\beta_n}{(2\pi)^n} \int_a^\infty r^{n-1} e^{-r^2/2s^2} dr, \tag{7}$$

where

$$\beta_n = \frac{2\ \pi^{n/2}}{\Gamma(n/2)}$$

is the surface area of the unit-radius sphere in $\mathbf{R}^n$.

Let $s^2 = 2^{-n}$. Substitute $u = r^2/(2s^2)$ in (7), to get

$$I_a = \frac{\beta_n s^n}{(2\pi)^n} 2^{n/2-1} \int_{\frac{a^2}{2s^2}}^{\infty} u^{n/2-1} e^{-u} du. \tag{8}$$

The constant is

$$\frac{\beta_n s^n}{(2\pi)^n} 2^{n/2-1} = \frac{2\ \pi^{n/2}}{\Gamma(n/2)} \frac{2^{-n^2/2}}{(2\pi)^n} 2^{n/2-1} = \frac{2^{-n^2/2}}{(2\pi)^{n/2}\Gamma(n/2)}, \tag{9}$$

and $a^2/(2s^2) = 2^{n-1}a^2$. ∎

Corollary 4.2 *Let $a = n2^{-n/2}$. Then for $n \geq 2$,*

$$I_a = \frac{2^{-n^2/2}}{(2\pi)^{n/2}} e^{-n^2/2 + O(n \log n)}.$$

Proof. The incomplete Gamma function has the asymptotic series

$$\Gamma(\alpha, z) \sim z^{\alpha-1} e^{-z} \left[1 + \frac{\alpha - 1}{z} + \frac{(\alpha-1)(\alpha-2)}{z^2} + \cdots \right].$$

Furthermore, when $\alpha > 1$ and $z > 0$, its terms start out positive, and then alternate. If we use only the initial segment of positive terms, we get an upper bound. (This follows from the remainder estimate given by Abramowitz and Stegun [1], 6.5.32.) Our parameters are

$$\alpha = n/2, \qquad z = \frac{a^2}{2s^2} = \frac{n^2}{2},$$

so

$$\Gamma(n/2, n^2/2) \leq (n^2/2)^{n/2-1} e^{-n^2/2} \frac{n}{n-1},$$

since $\sum_{k \geq 0} \frac{1}{n^k} = (1 - 1/n)^{-1}$. The result follows from Lemma 4.3 and Stirling's formula. ∎

4.3 A Multi-Factor Cauchy-Schwartz Inequality.

Let $f_1, \ldots f_r$ be non-negative functions, in $L_1(\mu)$ for some positive measure μ. Then

$$\int f_1 \cdots f_r d\mu \leq \prod_{i=1}^{r} \left(\int f_i^r d\mu \right)^{1/r} . \tag{10}$$

This is a known result. See Polya and Szego [8], p. 68 for integrals over an interval. It can be proved by induction, as follows. The base case, $r = 2$, is Cauchy-Schwartz. To reduce r to $r-1$, apply Holder's inequality to $f_1 \cdots f_{r-1}$ and f_r, taking $p = r/(r-1)$, $q = r$.

4.4 Estimates for the Exponential Function.

Recall the Taylor series

$$e^{-t} = \sum_{k \geq 0} (-)^k \frac{t^k}{k!} = 1 - t + t^2/2 - \cdots .$$

Then, if $m \geq 2$ is even, we have

$$\sum_{k=0}^{m-1} (-)^k \frac{t^k}{k!} \leq e^{-t} \leq \sum_{k=0}^{m} (-)^k \frac{t^k}{k!}, \qquad \text{for all } t \geq 0 \ . \tag{11}$$

For a proof of this, see the discussion of "enveloping series" in [4], pp. 32. ff.

4.5 Bounds for Moments.

Recall that we defined $T_{2k} = \sum_\sigma (t^\sigma)^{2k}$, for $k \geq 2$. Our goal is now to estimate moments of these functions, against the Gaussian density

$$d\mu = \frac{1}{(2\pi)^{n/2}} e^{-||t||^2/(2s^2)} dt$$

on $\mathbf{R}^n$, where $s^2 = 2^{-n}$.

Lemma 4.4 *Let*

$$d_n = s^n/(2\pi)^{n/2} = \frac{2^{-n^2/2}}{(2\pi)^{n/2}}.$$

Then

$$\frac{1}{(2\pi)^n}\int_{\mathbf{R}^n} T_{2k_1}\cdots T_{2k_r}e^{-||t||^2/(2s^2)}dt = O\left(d_n \frac{n^{\sum k_i}}{2^{n\sum(k_i-1)}}\right). \qquad (12)$$

as $n \to \infty$*.*

Proof. Let

$$M = \int T_{k_1}\cdots T_{k_r}d\mu,$$

where $k_i \geq 2$, with repetitions allowed.

Replacing the T's by their definitions, we get

$$M = \sum_{\sigma_1,\ldots,\sigma_r}\int (t^{\sigma_1})^{2k_1}\cdots(t^{\sigma_r})^{2k_r}d\mu,$$

where the sum is over all the 2^{nr} possible r-tuples of sign combinations. By (10),

$$M \leq \sum_{\sigma_1,\ldots,\sigma_r}\prod_{i=1}^{r}\left(\int (t^{\sigma_i})^{2k_ir}d\mu\right)^{1/r}. \qquad (13)$$

The Gaussian measure is spherically symmetric, so we can replace the i-th integral in (13) by

$$J_i := \int (t_1+\cdots+t_n)^{2k_ir}d\mu.$$

Choose an orthogonal change of coordinates such that

$$u_1 = \frac{t_1+\cdots+t_n}{\sqrt{n}}.$$

(Values of $u_2,\ldots,u_n$ are not important here.) Then, since $\sum t_i = \sqrt{n}u_1$, we have

$$\begin{aligned} J_i &= \frac{n^{k_ir}}{(2\pi)^{n/2}}\int_{\mathbf{R}^n} u_1^{2k_ir}e^{-||u||^2/2s^2}du \\ &= \frac{n^{k_ir}s^{n-1}}{(2\pi)^{1/2}}\int_{\mathbf{R}} u_1^{2k_ir}e^{-u_1^2/2s^2}du_1 \\ &= \frac{n^{k_ir}s^{n+2rk_i}}{(2\pi)^{1/2}}\int_{\mathbf{R}} v^{2k_ir}e^{-v^2/2}dv \\ &= (2rk_i-1)(2rk_i-3)\cdots(3)(1)\frac{n^{k_ir}s^n}{2^{nrk_i}} \leq \frac{n^{k_ir}s^n}{2^{nrk_i}}(2rk_i)^{rk_i}. \end{aligned}$$

Substituting this bound into (13), we see that

$$\frac{1}{(2\pi)^{n/2}}\int_{\mathbf{R}^n} T_{2k_1}\cdots T_{2k_r}e^{-||t||^2/(2s^2)}dt = O\left(s^n\frac{n^{\sum k_i}}{2^n\sum^{(k_i-1)}}\right),$$

with an implied constant depending on the k_i's. ■

Any moment can be computed exactly, should this be desired. We illustrate with three examples that we will use later.

Any T_{2k} can be handled by the same orthogonal transformation we used to prove Lemma 4.4. For example, we have

$$\frac{1}{(2\pi)^n}\int_{\mathbf{R}^n} G(t)T_4(t)dt = d_n\frac{3n^2}{2^n}, \qquad (14)$$

and

$$\frac{1}{(2\pi)^n}\int_{\mathbf{R}^n} G(t)T_6(t)dt = d_n\frac{15n^3}{4^n}, \qquad (15)$$

Moments involving cross products or powers are more involved. Consider, for example $(T_4)^2$. This is an even symmetric homogeneous polynomial of degree 8, so it must have the form

$$A\sum_i t_i^8 + B\sum_{i\neq j} t_i^6t_j^2 + C\sum_{i<j} t_i^4t_j^4 + D\sum_{\substack{i\neq j,k\\ j<k}} t_i^4t_j^2t_k^2 + E\sum_{i<j<k<\ell} t_i^2t_j^2t_k^2t_\ell^2.$$

Thus, our moment equals what we would have got had we integrated

$$Ant_1^8 + Bn(n-1)t_1^6t_2^2 + C\binom{n}{2}t_1^4t_2^4 + Dn\binom{n-1}{2}t_1^4t_2^2t_3^2 + E\binom{n}{4}t_1^2t_2^2t_3^2t_4^2.$$

To find the coefficients, we can make some judiciously chosen substitutions. First set $t_1 = 1$, and all other $t_i = 0$, and conclude that $A = 4^n$. To find B and C similarly set $t_1 = t_2 = 1$, and $t_1 = 2, t_2 = 1$ to get two linear equations, and then solve these to get $B = 12\cdot 4^n$ and $C = 38\cdot 4^n$. Finally, from $t_1 = t_2 = t_3 = 1$, get $D = 84\cdot 4^n$ and from $t_1 = t_2 = t_3 = t_4 = 1$, get $E = 216\cdot 4^n$. Since our multivariate Gaussian density factors, computing the integral for T_4^2 reduces to computing moments of a univariate normal distribution, which are known [6, p. 227]. Proceeding in this way, we get

$$\frac{1}{(2\pi)^n}\int_{\mathbf{R}^n} G(t)T_4^2dt = d_n\frac{3n(3n^3+24n^2+24n-16)}{2^{2n}}. \qquad (16)$$

Applied to a polynomial of degree m, this procedure finds $P(m/2)$ coefficients, where $P(n)$ denotes the number of partitions of n. It is thus more efficient than a brute force expansion of a product of T_{2k}'s.

4.6 Asymptotics for B_n.

Theorem 4.1 *As $n \to \infty$, we have*

$$B_n = D_n \left(1 - \frac{n^2}{4 \cdot 2^n} + O(\frac{n^4}{2^{2n}})\right).$$

Proof.

For sufficiently small $|z|$,

$$-\frac{z^2}{2} - \frac{z^4}{12} - \frac{2z^6}{45} \leq \log\cos z \leq -\frac{z^2}{2} - \frac{z^4}{12}.$$

(In fact, $|z| \leq z_0 := 1.3228...$ is small enough.) This implies

$$-\frac{2^n \sum t_i^2}{2} - \frac{T_4}{12} - \frac{2T_6}{45} \leq \sum_\sigma \log\cos(t^\sigma) \leq -\frac{2^n \sum t_i^2}{2} - \frac{T_4}{12}.$$

So

$$e^{-2^n \sum t_i^2/2} e^{-T_4/12 - 2T_6/45} \leq F_Z(t) \leq e^{-2^n \sum t_i^2/2} e^{-T_4/12}.$$

Applying (11) with $m = 2$ to the factors involving T_4 and T_6, we get

$$G(t)(1 - \frac{T_4}{12} - \frac{2T_6}{45}) \leq F_Z(t) \leq G(t)(1 - \frac{T_4}{12} + \frac{T_4^2}{288}).$$

The upper bound holds for any $t \in C$, and the lower bound additionally $||t|| \leq z_0$.

Recall that B_n is expressed using the integral (2) over F. If we write

$$F = \{t \in F \cap C : ||t|| < n2^{-n/2}\} + \{t \in F \cap C : ||t|| \geq n2^{-n/2}\} + (F \backslash C)$$

and consider each piece separately, we get

$$L_n \leq \frac{1}{(2\pi)^n} \int_F F_Z(t) dt \leq U_n,$$

where

$$L_n = \frac{1}{(2\pi)^n}\int_{\mathbf{R}^n} G(t)(1 - \frac{T_4}{12} - \frac{2T_6}{45})dt - \frac{1}{(2\pi)^n}\int_{||t||>n2^{-n/2}} G(t)dt.$$

(Note that $n2^{-n/2} < z_0$ for all $n \geq 1$.) Similarly, using $F = (F \cap C) + (F \backslash C)$, we obtain

$$U_n = \frac{1}{(2\pi)^n}\int_{\mathbf{R}^n} G(t)(1 - \frac{T_4}{12} + \frac{T_4^2}{288})dt + \frac{1}{(2\pi)^n}\int_{F\backslash C} F_Z(t)dt.$$

Using (14), Lemma 4.4, and Corollary 4.2, we get

$$L_n = d_n\left(1 - \frac{n^2}{4\cdot 2^n} + O(\frac{n^3}{2^{2n}})\right) + d_n e^{-n^2/2+O(n\log n)}.$$

Similarly, but now using Corollary 4.1, we have

$$U_n = d_n\left(1 - \frac{n^2}{4\cdot 2^n} + O(\frac{n^4}{2^{2n}})\right) + d_n e^{-2^{n+1}/(2n+1)^2+O(n^2)}.$$

Asymptotically, the rightmost error terms can be absorbed into the big-O terms, so we can multiply by $2^{2^n}\cdot 4^n/2$ to get the theorem. ■

Using more terms in our Taylor expansions, we can get more accurate formulas. This involves significantly more computation, but no new ideas. Accordingly, we just state the result.

Theorem 4.2 *There are polynomials P_1, P_2, ... such that*

$$B_n = D_n\left(1 + \sum_{k=1}^{\nu}\frac{P_k(n)}{2^{kn}} + O(n^{2\nu+2}2^{-(\nu+1)n})\right).$$

The first three P_i are

$$P_1 = -\frac{n^2}{4};$$

$$P_2 = \frac{n(3n^3 - 8n^2 + 24n - 16)}{96};$$

$$P_3 = -\frac{n(n^5 - 8n^4 + 72n^3 - 336n^2 + 512n - 256)}{384}.$$

It can be shown that P_k has degree $2k$.

Taking $\nu = 1, 2, 3$ in the theorem, we get approximations which we denote by E_n, F_n, and G_n. The following data show that they are accurate even for small n.

n	D_n	E_n	F_n	G_n	B_n
1	$2.25676e+00$	$1.97466e+00$	$1.99229e+00$	$2.00331e+00$	2
2	$5.09296e+00$	$3.81972e+00$	$3.97887e+00$	$4.01866e+00$	4
3	$2.29872e+01$	$1.65220e+01$	$1.72516e+01$	$1.73235e+01$	18
4	$8.30023e+02$	$6.22517e+02$	$6.41971e+02$	$6.41971e+02$	648
5	$3.83624e+06$	$3.08697e+06$	$3.14141e+06$	$3.13984e+06$	3140062
6	$5.80992e+14$	$4.99290e+14$	$5.03616e+14$	$5.03489e+14$	503483766022188

We obtained values of B_n from the data of Palmer, Read, and Robinson [7], who enumerated correlation immune functions by Hamming weight. We note that $B_n \le D_n$ for all $n \le 6$. One suspects that this bound actually holds for all n.

4.7 Explicit Bounds.

The estimates of the last section can be converted into explicit bounds. Using (14)–(16) and Lemma 4.3, we have the lower bound

$$B_n \ge D_n \left(1 - \frac{n^2}{4 \cdot 2^n} - \frac{2n^3}{3 \cdot 4^n} - 2\frac{(n^2/2)^{n/2} e^{-n^2/2}}{n(n-1)\Gamma(n/2)}\right).$$

Similarly, but with Lemma 4.1, we get an upper bound. This bound, however, is only sharp for $n \ge 15$. This is because it uses a worst-case estimate for the cosine product. If we use estimates that vary with t, and average over possible t, we can do better.

For convenience in calculation, we consider only the set $F \cap \{t_i \ge 0\}$, and assign it the mass 1, using a product measure. Following the ideas in Lemma 4.1, there are four cases to consider.

First, suppose there is some i, $i = 1, \ldots, n$ such that $t_i = \alpha$, with $\pi/(2n) < t_i < \pi/4$. Then, since flipping the sign of t_i will bounce any $t^\sigma \bmod \pi$ out of the interval $(-\alpha, \alpha)$, we have $F_Z \le \cos(\alpha)^{2^{n-1}}$. The mass associated with the interval $(\alpha, \alpha + d\alpha)$ is (at most)

$$dV = \frac{1}{2 \cdot n!} \times n \left(\frac{\alpha}{\pi/4}\right)^{n-1} d\left(\frac{\alpha}{\pi/4}\right).$$

(We divide by 2 because t_n only goes up to $\pi/4$, and by $n!$ because we only care about $t \notin C$.) Integrating over possible α, we see that the total contribution for this case is at most

$$E_1 := \frac{2}{\pi(n-1)!} \int_{\pi/2n}^{\pi/4} \left(\frac{\alpha}{\pi/4}\right)^{n-1} (\cos\alpha)^{2^{n-1}} d\alpha.$$

Second, suppose $t_i < \pi/(2n)$ for $i = 1, \ldots, n-1$, and $t_n = \pi/2 - \beta$, with $\pi/(2n) < \beta < \pi/4$. By similar reasoning, $t^\sigma \bmod \pi$ is outside $(-\beta, \beta)$ for at least half of the σ's, so this case contributes at most

$$E_2 := \frac{2}{\pi}\left(\frac{2}{n}\right)^{n-1} \int_{\pi/2n}^{\pi/4} (\cos\beta)^{2^{n-1}} d\beta.$$

Third, suppose $t_1, \ldots, t_{n-1}$ are as in the previous case, but now $t_n = \gamma$, with $\pi/2 - \pi/(2n) < \gamma < \pi/2$, and for all σ, we have $t^\sigma \bmod \pi \notin (-\pi/(2n+1), \pi/(2n+1))$. The contribution from this case is bounded by

$$E_3 := \left(\frac{2}{n}\right)^{n-1} \cdot \frac{1}{n}\left(\cos\frac{\pi}{2n+1}\right)^{2^n}.$$

Fourth, let the t_i be as in the previous case, but assume there is some σ with $t^\sigma \bmod \pi \in (-\pi/(2n+1), \pi/(2n+1))$. As noted in the proof of Lemma 4.1, there must be some $j \leq n-1$ with $t_j \geq \pi/(2n+1)$. Since $\max\{t_1, \ldots, t_n\}$ is at most $\pi/(2n)$ but at least $\pi/(2n+1)$, the mass associated with this case is bounded by

$$\frac{1}{n}\left[\left(\frac{\pi}{2n} - \frac{\pi}{2n+1}\right) \cdot \frac{4}{\pi}\right]^{n-1} = \frac{1}{n}\left[\frac{4}{2n(2n+1)}\right]^{n-1}.$$

Hence this case contributes at most

$$E_4 := \frac{1}{n}\left(\frac{4}{2n(2n+1)}\right)^{n-1}\left(\cos\frac{\pi}{2n+1}\right)^{2^{n-1}}.$$

With this case analysis, we get the upper bound

$$B_n \leq D_n\left(1 - \frac{n^2}{4 \cdot 2^n} + \frac{n(3n^3 + 24n^2 + 24n - 16)}{96 \cdot 4^n}\right) + (E_1 + E_2 + E_3 + E_4)2^{2^n}.$$

The integrals in E_1 and E_2 are amenable to numerical integration.

The table below displays these bounds. Since B_n is not known for $n \geq 7$, we presume that G_n is accurate enough to use as a reference value. We have not continued beyond $n = 16$ because all columns agree with G_n to the precision given.

n	lower bound	G_n	upper bound
7	$1.682368e+32$	$1.715225e+32$	$3.038417e+33$
8	$5.284468e+68$	$5.322504e+68$	$2.413768e+70$
9	$2.773184e+143$	$2.780329e+143$	$2.826073e+145$
10	$8.334293e+294$	$8.341773e+294$	$1.289014e+297$
11	$1.668020e+600$	$1.668529e+600$	$1.511604e+602$
12	$2.988472e+1213$	$2.988776e+1213$	$1.807850e+1214$
13	$8.630092e+2442$	$8.630378e+2442$	$8.653684e+2442$
14	$1.299441e+4905$	$1.299454e+4905$	$1.299459e+4905$
15	$1.066098e+9833$	$1.066102e+9833$	$1.066103e+9833$
16	$5.200232e+19692$	$5.200237e+19692$	$5.200239e+19692$

5 Counting 1-Resilient Functions.

Our methods extend to count correlation immune Boolean functions that are also balanced, in the sense that there are an equal number of inputs giving 0 and 1 outputs. Such functions have been called 1-resilient [3]. In this section we indicate how our results extend to this case, without giving any of the proofs.

First, the probability that a Boolean function of n variables is 1-resilient is

$$\frac{1}{(2\pi)^{n+1}} \int_{[\pi,\pi]^{n+1}} F'_Z(t)dt,$$

where

$$F'_Z(t) = \prod_{\text{all signs}} \cos(t_0 \pm t_1 \pm \cdots \pm t_n).$$

Note that there are now $n+1$ variables but 2^n sign combinations.

The analog of Denisov's formula is

$$D'_n := \frac{1}{2}\left(\frac{8}{\pi}\right)^{(n+1)/2} 2^{-n(n+1)/2} \cdot 2^{2^n}.$$

Asymptotic corrections can also be made to this formula, and we can re-use a lot of our work if we note that

$$F'_Z(t_0, \dots, t_n)^2 = F_Z(t_0, \dots, t_n).$$

Furthermore, $F'_Z \geq 0$ on $C' = \{t : ||t|| \geq \pi/2\}$, where the action is.

The analog of Theorem 4.2 holds, but now with polynomials $P'_i(n) = P_i(n+1)$. (That is, we use the same polynomials, but now plug in $n+1$.) For example, if B'_n denotes the number of n-variable 1-resilient Boolean functions, then we have, as $n \to \infty$,

$$B'_n = D'_n \left(1 - \frac{(n+1)^2}{4 \cdot 2^n} + O(\frac{n^4}{2^{2n}})\right).$$

Here is some numerical data.

n	D'_n	E'_n	F'_n	G'_n	B'_n
1	$2.55e+00$	$1.27e+00$	$1.59e+00$	$1.750704e+00$	0
2	$4.06e+00$	$1.78e+00$	$2.29e+00$	$2.395397e+00$	2
3	$1.30e+01$	$6.48e+00$	$7.70e+00$	$7.700410e+00$	8
4	$3.31e+02$	$2.02e+02$	$2.21e+02$	$2.194959e+02$	222
5	$1.08e+06$	$7.78e+05$	$8.10e+05$	$8.081522e+05$	807980
6	$1.16e+14$	$9.37e+13$	$9.53e+13$	$9.526281e+13$	95259103924394
7	$2.67e+31$	$2.33e+31$	$2.35e+31$	$2.347810e+31$	(see below)

The value $B'_7 = 23478015754788854439497622689296$ is not displayed in the table, for typographical reasons.

Here, E'_n, F'_n, G'_n are analogous to their unprimed counterparts. The exact values for $n \leq 6$ come from Palmer, Read, Robinson [7], and B'_7 was recently computed by Alfredo Viola.

6 Acknowledgements.

I would like to thank David Page and Lisa Hellerstein for getting me interested in this problem.

Support by the National Science Foundation, via grants CCF-0523680 and CCF-0635355, is gratefully acknowledged. We also acknowledge support from a Vilas Research Associate Award from the University of Wisconsin.

References

[1] M. Abramowitz and I. Stegun, eds., Handbook of Mathematical Functions, Dover, 1972.

[2] P. Camion, C. Carlet, P. Charpin, and N. Sendrier, On correlation-immune functions, Proc. CRYPTO 91, pp. 86-100.

[3] B. Chor et al., The bit extraction problem or t-resilient functions, Proc. 26th IEEE Symp. Found. Computer Science, 1985, pp. 396-407.

[4] O. V. Denisov, An asymptotic formula for the number of correlation-immune of order k Boolean functions, Discrete Math. Appls. 2, 407-426, 1992.

[5] W. Fleming, Functions of Several Variables, Addison-Wesley, 1965.

[6] B. V. Gnedenko, The Theory of Probability, Chelsea, 1967.

[7] E. M. Palmer, R. C. Read, and R. W. Robinson, Balancing the n-cube: a census of colorings, J. Algebraic Combin., 1, 1992, 257-273.

[8] G. Polya and G. Szego, Problems and Theorems in Analysis I, Springer-Verlag, New York, 1972.

[9] B. Rosell, L. Hellerstein, S. Ray, and D. Page, Why skewing works: learning difficult Boolean functions with greedy tree learners, Proc. 22nd Intl. Conf. Machine Learning, 2005, pp. 728-735.

[10] B. Roy, A Brief Outline of Research on Correlation Immune Functions, Proc. ACISP 2002, Springer-Verlag, 2002, pp. 379-394.

[11] T. Siegenthaler, Correlation-immunity of nonlinear combining functions for cryptographic applications, IEEE Trans. Inform. Theory, IT-30, 1984, pp. 776-780.

[12] T. Siegenthaler, Decrypting a class of stream ciphers using ciphertext only, IEEE Trans. Computers, C-34, 1985, pp. 80-84.

[13] G.-Z. Xiao and J. L. Massey, A spectral characterization of correlation-immune combining functions, IEEE Trans. Inform. Theory, v. 34, 569-571, 1988.