COMBINATORICA
Akadémiai Kiadó – Springer-Verlag

COMBINATORICA 14 (1) (1994) 35–50

SPECTRAL PROPERTIES OF THRESHOLD FUNCTIONS

CRAIG GOTSMAN* and NATHAN LINIAL

Received May 8, 1990
Revised May 21, 1991

We examine the spectra of boolean functions obtained as the sign of a real polynomial of degree *d*. A tight lower bound on various norms of the lower *d* levels of the function's Fourier transform is established. The result is applied to derive best possible lower bounds on the influences of variables on linear threshold functions. Some conjectures are posed concerning upper and lower bounds on influences of variables in higher order threshold functions.

1. Introduction

The recent emergence of *neural networks* has revived interest in threshold functions, boolean functions obtained as the signs of real polynomials on the hypercube. Perhaps the most well-known examples of networks involving threshold gates are the Boltzmann machine [1] and the Hopfield associative memory model [12]. The back-propagation algorithm for training circuits with multiple layers of threshold gates ([25],[17]), a generalization of the celebrated Perceptron convergence procedure, has found many applications in real-world problem solving. This complements work done in the 60's, which focused mostly on linear threshold logic (e.g. [20,28]). In this paper we study the computational complexity of threshold functions.

The main mathematical tool we employ in our analysis is the Fourier transform. Use of harmonic analysis on the hypercube in the context of boolean functions can be traced back to the pioneering work of Ninomiya [21], where it was used for the classification of boolean functions. Only recently has its power been realized in the context of computational complexity, one of the most active fields in theoretical computer science today. These techniques have been applied successfully to derive significant results in circuit complexity (e.g. [7,19,4,14]).

Any boolean function can be represented uniquely as real 2^n-vector of its values at the points of the hypercube, or as a real 2^n-vector of its Fourier coefficients (the *spectrum* of the function). A recent focus of activity is characterizing the spectra of boolean functions in various complexity classes. This is mostly done by bounding

AMS subject classification codes (1991): 68 Q 15, 68 R 05

* Supported by an Eshkol fellowship, administered by the National Council for Research and Development — Israel Ministry of Science and Development.

l_p norms of the spectra of functions in these classes. For example, Bruck [8] has made extensive use of the l_1 and l_∞ spectral norms to characterize *polynomial threshold* functions, boolean functions which are obtained as the sign of a sparse real polynomial on the hypercube (one that consists of a polynomial number of terms). Use of these spectral norms enabled him to relate polynomial threshold functions to the class of AC^0 functions and the class of function computable by depth 2 circuits of *linear threshold* functions (signs of polynomials of degree 1). Linial et al. [19] have obtained results with a slightly different flavor. They show that the spectra of boolean functions in the low complexity class AC^0 have the following property:

Almost all the l_2 spectral norm is concentrated on the lower degree Fourier coefficients.

Exploiting this property, they were able to construct an efficient algorithm for *learning* AC^0 functions from examples. Kushilevitz et al. [16] have extended their algorithm to learn a wider class of boolean functions; those whose l_1 spectral norm is bounded from above by a polynomial (in n). This restriction eliminates functions whose spectrum is "spread out" over a large number of coefficients.

The result of Linial et al., reminiscent of typical signal processing applications, illuminates a possible connection between low complexity and concentration of l_p spectral norms on the lower Fourier coefficients. Since the Fourier coefficients of a boolean function f are just the coefficients of the representation of f as a real polynomial on the hypercube, this result implies that the boolean function defined as the sign of the low degree polynomial obtained by retaining only those terms with significant spectral weight is an excellent approximation for f. This interpretation led Gotsman [9] to conjecture that the opposite is also true, i.e. the l_2 spectral norm of a boolean function obtained as the sign of a low degree real polynomial has properties similar to the above. More specifically, calling the class of boolean functions obtained as the sign of a degree d real polynomial d-*threshold* functions, Gotsman conjectured that at least a constant fraction of the l_2 spectral norm is concentrated on the coefficients of degree $\leq d$. The main objective of this paper is to settle this conjecture. Towards this end, we obtain tight lower bounds on the l_p spectral norm ($1 \leq p \leq 2$) of the coefficients of degree $\leq d$ for d-threshold functions. This generalizes a lower bound of Bruck's (in a slightly different form in [7]) on the l_1 spectral norm. Bruck's bound seems to be weaker than ours, but it nontheless enabled him to obtain useful results on polynomial threshold functions. Since this paper was first submitted, Aspnes et al. [3] have obtained results complementary to ours, showing that the highest-degree Fourier coefficient of a d-threshold function cannot be too large.

The theme common to our results and those of [19] is that the following three types of boolean functions are related: 1. low-complexity functions, 2. functions obtained as the sign of a low degree polynomial, 3. functions whose spectral norms are concentrated on the lower degree Fourier coefficients.

2. Boolean functions and Fourier transforms

2.1. The basics

Denote by

$$\mathcal{F}_n = \{f : \{+1,-1\}^n \longrightarrow \mathcal{R}\}$$

the class of real functions on the hypercube, and by

$$\mathcal{B}_n = \{f : \{+1,-1\}^n \longrightarrow \{+1,-1\}\}$$

the class of boolean functions in n variables. Any function $f \in \mathcal{F}_n$ may be considered as a 2^n-vector in $\mathcal{R}^{2^n}$, each coordinate representing the function value on a point of the hypercube. We assume an arbitrary, but fixed, ordering of the points of the hypercube. What follows is standard harmonic analysis on the hypercube, and its use in the context of boolean functions can be traced back as far as [21], [28] and [18]. Consider the following *inner product* on $\mathcal{R}^{2^n}$:

$$(1) \qquad \langle f_1, f_2 \rangle = 2^{-n} \sum_{x \in \{+1,-1\}^n} f_1(x) f_2(x).$$

This is the *correlation* of the random variables f_1 and f_2 on the probability space $\{+1,-1\}^n$ equipped with the uniform probability distribution.

Denote $[n] = \{1,\dots,n\}$. Now consider the set of 2^n boolean functions

$$X = \{X^I = \prod_{i \in I} x_i : I \subseteq [n]\}$$

consisting of products of variables in all possible subsets of $\{x_1,\dots,x_n\}$ (including the empty set - the constant **1** function). The superscript I indicates the subset of variables participating in the product, and if $|I| = d$, X^I will be called *of order d*. These functions are also known as the *Walsh* functions. They are just the parity functions of the variables. It is easy to verify that X is an orthonormal basis for $\mathcal{R}^{2^n}$:

$$\langle X^I, X^J \rangle = \delta_{IJ}$$

so that any function $f \in \mathcal{F}_n$ can be expressed uniquely as

$$(2) \qquad f = \sum_{I \subseteq [n]} \hat{f}(I) X^I \quad ; \quad \hat{f}(I) = \langle f, X^I \rangle.$$

The vector $\hat{f}$ is the *Fourier transform* or *spectrum* of f, and the coefficients $\hat{f}(I)$ the *Fourier coefficients*, those corresponding to dth order Walsh functions called *of degree d*. There are $\binom{n}{d}$ Fourier coefficients of degree d. Note that (2) is a representation of f as a multilinear polynomial in $x_1,\dots,x_n$.

The orthonomality of the basis is precisely the reason the normalization constant 2^{-n} was introduced in the definition (1) of the inner product. This is convenient, but raises some other difficulties. The standard property of Fourier transforms $\hat{\hat{f}}=f$ does not hold exactly, but only up to a constant factor.

A fundamental tool in the investigation of spectra of functions are the relations between norms of f and $\hat{f}$. For $1\leq p\leq\infty$, denote

$$||f||_p = \left[2^{-n} \sum_{x\in\{+1,-1\}^n} |f(x)|^p\right]^{1/p}$$

and

$$|||\hat{f}|||_p = \left[\sum_{I\subseteq[n]} |\hat{f}_I|^p\right]^{1/p}.$$

Here too the normalization factor 2^{-n} must be taken into account, hence it appears in the definition of $||f||_p$, but not in $|||\hat{f}|||_p$. Let p be the *dual* of q (i.e. $1/p+1/q=1$). The *Hausdorff-Young inequalities* are [15]:

$$(3) \qquad ||f||_p \geq |||\hat{f}|||_q \qquad 1\leq p\leq 2$$

$$(4) \qquad ||f||_p \leq |||\hat{f}|||_q \qquad 2\leq p\leq \infty$$

Equality exists when $p=q=2$.

$$(5) \qquad ||f||_2 = \sqrt{\langle f,f\rangle} = |||\hat{f}|||_2.$$

More generally, for any two vectors $f_1, f_2 \in \mathcal{R}^{2^n}$, the orthonormality of the basis X implies *Parseval's equality*:

$$(6) \qquad \langle f_1, f_2\rangle = 2^n\langle \hat{f}_1, \hat{f}_2\rangle.$$

For boolean functions, $||f||_p=1$ for all p, so (3), (4) and (5) reduce to

$$(7) \qquad \begin{array}{ll} |||\hat{f}|||_p \geq 1 & 1\leq p\leq 2 \\ |||\hat{f}|||_p \leq 1 & 2\leq p\leq \infty \\ |||\hat{f}|||_2 = 1 = \sum_{I\subseteq[n]} \hat{f}^2(I). & \end{array}$$

2.2 Influences and average sensitivity

The following measures of complexity of a boolean function were introduced in [5]:

Definition 2.1. Let $f\in\mathcal{B}_n$ be a boolean function.

The i'th difference of f at x is

$$\Delta_i(f,x) = \frac{1}{2}[f(x_1,\dots,x_i,\dots,x_n) - f(x_1,\dots,-x_i,\dots,x_n)].$$

The *influence* of the i'th variable is

$$\mathbf{Inf}_i(f) = ||\Delta_i||_1 \quad i = 1,\dots,n. \tag{8}$$

$\mathbf{Inf}_i(f)$ is the probability that flipping the i'th variable of a random boolean input will flip the output, i.e. how influential the variable x_i is in determining the outcome of f. Since $\Delta_i \in \{0,+1,-1\}$, the influences may be expressed alternatively in terms of other l_p-norms for $p \geq 1$:

$$\mathbf{Inf}_i(f) = ||\Delta_i||_p^p \quad i = 1,\dots,n. \tag{9}$$

Definition 2.2. The *sensitivity* of f at x is

$$\mathbf{Sens}(f,x) = \sum_{i=1}^{n} |\Delta_i(f,x)|.$$

This is n times the probability that flipping any one bit of the input x will flip the output.

Definition 2.3. The *average sensitivity* of f is

$$\mathbf{AS}(f) = \mathbf{E}_x[\mathbf{Sens}(f,x)] = \sum_{i=1}^{n} \mathbf{Inf}_i(f).$$

Obviously $0 \leq \mathbf{Inf}_i(f) \leq 1$ for $i = 1,\dots,n$ and $0 \leq \mathbf{AS}(f) \leq n$. Regarding Δ_i as a real function on the n-cube, the following holds:

$$\Delta_i(x) = \sum_{\{I:i\in I\}} \hat{f}(I)X^I. \tag{10}$$

Applying (9) with $p = 2$, (5) and (10) yields:

$$\mathbf{Inf}_i(f) = \sum_{\{I:i\in I\}} \hat{f}^2(I)$$

$$\mathbf{AS}(f) = \sum_{I\subseteq[n]} |I|\hat{f}^2(I). \tag{11}$$

Combining (8) with the Hausdorff-Young inequality (3) for $p = 1$ immediately yields the following lower bound on variable influences:

Lemma 2.1. *For any boolean function f*

$$\mathbf{Inf}_i(f) \geq \max_{\{I:i\in I\}} \{|\hat{f}(I)|\} \qquad i = 1,\dots,n. \qquad \blacksquare$$

For monotone boolean functions, it was observed in [13] that the definitions imply:

Lemma 2.2. *For any monotone boolean function f*

$$\mathbf{Inf}_i(f) = |\hat{f}(\{i\})| \qquad i = 1, \dots, n. \qquad ▮$$

This also holds for *locally monotone* boolean functions, those which are monotonic increasing or decreasing in each variable. Lemmas 2.1 and 2.2 imply:

Corollary 2.1. *For any locally monotone boolean function f*

$$|\hat{f}(\{i\})| \geq \max_{\{I: i \in I\}} \{|\hat{f}(I)|\} \qquad i = 1, \dots, n. \qquad ▮$$

This means that for locally monotone boolean functions, the linear Fourier coefficients are the largest in absolute value.

The average sensitivity of a boolean function f also has a combinatorial interpretation. Consider the two-coloring of the points of the n-dimensional hypercube induced by f. Connect an edge between any two adjacent vertices whose colors differ. The average sensitivity is (twice) the number of these edges, i.e. the size of the *cut* of the two monochromatic sets.

3. The threshold function hierarchy

3.1. Definitions

In the previous section, we saw that any n-variable boolean function can be expressed *exactly* and uniquely as a real multilinear polynomial of degree n. Many boolean functions may be expressed as the *sign* of real polynomials of degree $<n$. For example, the majority function of 3 variables is expressed exactly as a cubic polynomial $\mathrm{maj}(x_1,x_2,x_3) = \frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2}x_3 - \frac{1}{2}x_1x_2x_3$, but also as the sign of the linear expression $(x_1+x_2+x_3)$. The following definition makes these notions precise:

Definition 3.1. The class of threshold functions of degree d in n variables is:

$$\mathcal{T}_n^d = \{f \in \mathcal{B}_n : f = \mathrm{sgn}(\sum_{|I| \leq d} w_I X^I) \text{ for some real vector } (w_I)\}.$$

If $f \in \mathcal{T}_n^d$, we say that f is **d-threshold**. The w_I are called the **realizing weights** of f.

Without loss of generality, we assume that the polynomial $P = \sum_{|I| \leq d} w_I X^I$ does not vanish on any point of the n-cube. The case $d = 1$ is the class of *linear* threshold functions. These functions are of major interest in the theory of neural networks. It is well known [20] that $\mathcal{T}_n^d \neq \mathcal{T}_n^{d+1}$ for $0 \leq d < n$. The separating functions are the parity functions of any $d+1$ variables, i.e. these functions cannot be expressed as the sign of a polynomial of degree d. The following lemma shows that the n-variable parity function is actually the *only* function which cannot be expressed as a polynomial of degree $n-1$.

Lemma 3.1. *Let f be a n-variable boolean function which is not parity. Then $f \in \mathcal{T}_n^{n-1}$.*

Proof. Express f as a n-degree polynomial $f = \sum_I \hat{f}(I) X^I$ (its Fourier transform). Since f is not the parity function, the uniqueness of the transform implies that $|\hat{f}([n])| < 1$. This, in turn, implies that omitting $\hat{f}([n])$ will not effect the sign of the polynomial, therefore $f = \mathrm{sgn}(\sum_{|I|<n} \hat{f}(I) X^I)$. ∎

In the following sections, the majority function of (odd) n variables $f = \mathrm{sgn}(\sum_{i=1}^n x_i)$ will be constantly referred to as an extremal case. It is easily shown that for this function

$$\hat{f}(I) = \begin{cases} 0 & |I| \text{ is even} \\ O(n^{1/2-|I|}) & |I| \text{ is odd and constant} \end{cases} \tag{12}$$

and

$$|\hat{f}(I)| = |\hat{f}(J)| \quad \text{if} \qquad |I| + |J| = n + 1.$$

3.2 The main theorem

Bruck [7] has shown that a d-threshold function f is uniquely determined by its Fourier coefficients of degree $\leq d$. This is a generalization of the uniqueness property of the $n+1$ Chow parameters (the constant and linear Fourier coefficients) of *linear* threshold functions [28]. Consequently, any property of f may be expressed using only the Fourier coefficients of degree $\leq d$.

Recalling (7), Linial et al. [19] have shown that any boolean function $f \in AC^0$ (i.e. computable by a constant depth polynomial size circuit of $\vee, \wedge, \neg$ gates) satisfies

$$\sum_{|I| \leq O(\log^l n)} \hat{f}^2(I) = 1 - o(1) \tag{13}$$

where l is the depth of the circuit computing the function. Calling the sum of squares of the Fourier coefficients the l_2 *spectral norm*, (13) asserts that almost all the l_2 spectral norm of these functions is concentrated on a small $(2^{poly(\log n)})$ number of lower-degree coefficients. Gotsman [9], conjectured that a similar property holds for d-threshold functions, namely, most of their l_2 spectral norm is contained in the coefficients of degree $\leq d$. Independently, Bruck [7] obtained a constant lower bound on the l_1 spectral norm of the lower d levels of the transform. We generalize Bruck's result for any $1 \leq p \leq 2$ and confirm Gotsman's conjecture. Specifically, we bound from below the l_p spectral norm on the lower d levels with an expression independent of n for $1 \leq p \leq 2$. Our results, for the special case $p = 2$, imply that at least a constant fraction of the l_2 spectral norm is concentrated on the lower d levels. The mathematical tools we use are a set of inequalities, due to Bourgain, on the l_p norms of Walsh functions, which, for the case $d = 1$, reduce to the classical Khintchine inequalities:

Lemma 3.2. ([6]) *Let $P = \sum_{|I| \leq d} w_I X^I$ be a weighted sum of Walsh functions of degree $\leq d$. Then for any $1 \leq p \leq 2$*

$$||P||_1 \geq c_p^d ||P||_p$$

where $c_p \in (0,1]$ are constants depending on p alone. ▮

We are now ready to state the main theorem of this paper:

Theorem 3.3. *Let $f = \text{sgn}(P)$ be a d-threshold function. Denote $|||\hat{f}|||_p' = (\sum_{|I| \leq d} |\hat{f}(I)|^p)^{1/p}$. Then*

$$|||\hat{f}|||_p' \geq c_p^d \quad ; \quad 1 \leq p \leq 2$$

where c_p is a constant depending on p alone. Moreover, the inequalities are tight for the cases $p = 1, 2$, except for the numerical value of c_p.

Proof. Let q be the dual of p. Using the Hőlder inequality and Parseval's equality (6), noticing that w is the Fourier transform of P:

$$|||\hat{f}|||_p' \cdot |||w|||_q \geq \langle \hat{f}, w \rangle = \mathbf{E}[P \cdot \text{sgn}(P)] = ||P||_1.$$

Applying Lemma 3.2 and the Hausdorff-Young inequality (3) gives:

$$||P||_1 \geq c_p^d ||P||_p \geq c_p^d |||w|||_q$$

whence

$$|||\hat{f}|||_p' \geq c_p^d.$$

The strongest (and most useful) inequality is the case $p = 2$. To show that it is tight, it suffices to consider the case $d = n-1$. Take the n-variable parity function f and reverse the function value at any one point of the hypercube. By Lemma 3.1, the resulting Φ is a $(n-1)$-degree threshold function. The l_2 distance $||f - \Phi||_2^2$ is exponentially small, therefore (by Parseval's equality) so is the l_2 distance between their Fourier transforms $\hat{f}$ and $\hat{\Phi}$. The function $\hat{f}$ vanishes on the lower $n-1$ levels and has unit value on the highest coefficient. Therefore $\hat{\Phi}$ has exponentially small l_2 spectral norm on the lower $n-1$ levels. For smaller d, the same argument holds for the d-parity function on d arbitrary variables.

The tightness of the inequality for the case $p = 1$ is treated separately in Corollary 3.4. ▮

Theorem 3.3 does not hold for spectral norms with $p > 2$. By (12), the spectral norms of the (linear threshold) majority function satisfy

$$|||\hat{f}|||_p' = \left[\sum_{|I| \leq 1} |\hat{f}(I)|^p \right]^{1/p} = O(n^{1/p - 1/2})$$

which decreases with n for $p > 2$. ▮

Bruck's lower bound on the l_1 norm of d-threshold functions is obtained as a special case of Theorem 3.3:

Corollary 3.4. *For any d-threshold function f*

$$\sum_{|I|\leq d} |\hat{f}(I)| \geq 1.$$

Proof. Apply Theorem 3.3 with $p=1$, noticing that $c_1=1$. This is tight for the parity function of any subset of the variables. ▮

Corollary 3.5. *Let f be a linear threshold function. Then*

$$\hat{f}^2(\emptyset) + \sum_{i=1}^{n} \hat{f}^2(\{i\}) \geq \frac{1}{2}.$$

Proof. Apply Theorem 3.3 with $p=2$ and $d=1$, observing that the best constant in the Khintchine inequality is $1/\sqrt{2}$ [10]. ▮

Note 3.1. The converse of Corollary 3.5 does not hold. Widner [28] gives an example of a linear threshold function f and a non-linear threshold function g such that

$$\sum_{|I|\leq 1} \hat{f}^2(I) = \sum_{|I|\leq 1} \hat{g}^2(I)$$

Note 3.2. The proof of Theorem 3.3 actually gives a slightly stronger result than that stated. If f is the sign of a polynomial supported on a subset of the Walsh functions $W \subset X$ such that $|I| \leq d$ for all $X^I \in W$, the l_p spectral norm of f on those **same** Walsh functions is at least c_p^d for $1 \leq p \leq 2$.

An immediate application of Theorem 3.3 is as a necessary condition for a boolean function to be d-threshold.

Corollary 3.6. *If f is a boolean function such that $\hat{f}(I)=0$ for all $|I| \leq d$, then $f \notin \mathcal{T}_n^d$.* ▮

Specifically, this eliminates the parity functions of $d+1$ variables, as mentioned in Section 3.1.

The lower bound for $p=2$ is the most intuitively appealing, because of (7), i.e. the spectrum is normalized to unity. This also seems to be the strongest and we use it in the sequel (Theorem 4.1). Nonetheless, Bruck has used the l_1 lower bound (Corollary 3.4) to obtain results concerning polynomial threshold functions. We hope that this family of bounds will prove useful in future analysis of the complexity of threshold functions.

4. Lower bounds on variable influences

Kahn et al. [13] give a tight lower bound on the sum of squares of the influences of variables of a boolean function. Denote by $\mathbf{E}(f)$ the expected value of f (this is also $\hat{f}(\emptyset)$). Then

$$\sum_{i=1}^{n} \mathbf{Inf}_i^2(f) \geq k(1-|\mathbf{E}(f)|)^2 \log^2 n/n$$

where k is an absolute constant. Consequently, there always exists a variable with influence $\geq \sqrt{k}(1-|\mathbf{E}(f)|)\log n/n$ (because of the relation between l_2 and l_∞ norms). Call f *balanced* if $\mathbf{E}(f)=0$. For a balanced boolean function, this implies the existence of a variable with influence $\Omega(\log n/n)$. The explicit function with precisely these influences is obtained as follows [5]: Denote

$$m(n) = \log n - \log\log n + O(1)$$

Divide the n variables into $n/m(n)$ blocks of $m(n)$ variables each. Define $f=1$ iff at least one block of variables are all 1.

Interestingly enough, f is a balanced $m(n)$-threshold function, implying that the smallest possible influence is already obtained by threshold functions of fairly low degree. In this section we address the question of lower bounds on influences for d-threshold functions, in the range $1 \leq d \leq m(n)$.

4.1. Linear threshold functions

The result of Section 3.2 enables us to prove a tight lower bound on the influences of variables of linear threshold functions:

Theorem 4.1. *Let f be a linear threshold function such that $|\mathbf{E}(f)| \leq 1/\sqrt{2}$. Then*

$$\sum_{i=1}^{n} \mathbf{Inf}_i^2(f) \geq 1/2 - \mathbf{E}^2(f).$$

This is tight up to a constant factor.

Proof. Observe that linear threshold functions are locally monotone, therefore Lemma 2.2 applies:

$$\sum_{i=1}^{n} \mathbf{Inf}_i^2(f) = \sum_{|I|=1} \hat{f}^2(I).$$

By Corollary 3.5

$$\sum_{i=1}^{n} \mathbf{Inf}_i^2(f) \geq 1/2 - \mathbf{E}^2(f).$$

The bound is tight up to a constant factor, for the majority function with odd n is balanced and $\mathbf{Inf}_i(f)=2^{-n+1}\binom{n}{[n/2]}$ for $i=1,\dots,n$ ($[x]$ is the integer part of x). By Stirling's approximation:

$$\sum_{i=1}^{n}\mathbf{Inf}_i^2(f)=\frac{2}{\pi}(1+o(1)).$$ ∎

For balanced linear threshold functions, Theorem 4.1 implies:

Corollary 4.2. *If f is a balanced linear threshold function, then there exists a variable with influence $\Omega(n^{-\frac{1}{2}})$.* ∎

Here again, the majority function is extremal. Note that the lower bound on the l_1 spectral norm (Corollary 3.4) gives only $\Omega(1/n)$, which is even weaker than Kahn et al.'s lower bound ($\Omega(\log n/n)$) for general boolean functions.

4.2. Higher order threshold functions

We extend the construction of [5] to obtain a balanced d-threshold function with small influences in the range $1\le d\le m(n)$. For this we need the following lemma in probability theory, which is of independent interest and does not seem to be in the literature.

Lemma 4.3. *Let $p(x)$ be a non-increasing function of x such that $p(x)\in[0,1/2]$ for all x. Consider the family of binomial distributions $\{B(k,p(k)):k=1,2,\dots\}$ (k trials with success probability $p(k)$). There exists a K (possibly dependent on p) such that for all $k>K$, if $kp(k)$ is a positive integer, then $kp(k)$ is a median of $B(k,p(k))$.*

Proof. Consider the sequence $\{m_k=kp(k):k=1,2,\dots\}$. Without loss of generality, we deal only with the subsequence of m_k's which are integers. We distinguish between two cases:

1. There is a K such that for all $k>K$, $m_k>50$. This guarantees that for all $k>K$, $m_k(1-p(k))>25$. Now consider a $k>K$ and denote $\delta=(2\pi m_k(1-p(k))^{-1/2}$. Applying the normal approximation to the binomial distribution of $B(k,p(k))$ ([26] p. 130), we have

$$\mathrm{Prob}[B(k,p(k))\ge m_k]=1/2+\varepsilon\ge 1/2$$

where $0\le\varepsilon\le\delta/6$. By the same approximation, the m_k'th term of the probability density function of $B(k,p(k))$ is

$$\mathrm{Prob}[B(k,p(k))=m_k]=\binom{k}{m_k}p(k)^{m_k}(1-p(k))^{k-m_k}=\delta+\Delta$$

where $|\Delta|<1.2\delta^2$. Because $m_k(1-p(k))>50$, obviously $|\Delta|<5\delta/6$, implying:

$$\mathrm{Prob}[B(k,p(k))\le m_k]=1/2-\varepsilon+\delta+\Delta\ge 1/2+5\delta/6+\Delta\ge 1/2$$

as required.

2. There is a $m \leq 50$ and an infinite subsequence $\{m_{k_j} : j = 1,2,\ldots\}$ such that $m_{k_j} = m$ for all $j > 0$. The Poisson approximation to the distribution of $B(k_j,p(k_j))$ gives:

$$\text{Prob}[B(k_j,p(k_j)) \leq m] = e^{-m} \sum_{i=0}^{m} \frac{m^i}{i!} + \varepsilon(k_j)$$

where $|\varepsilon(k_j)| = O(1/k_j)$ ([26] p. 135). This indicates that the distribution function of $B(k_j,p(k_j))$ can be made arbitrarily close to Poisson(m) by increasing j. Choose K_m such that

$$|\varepsilon(K_m)| \leq 0.25e^{-m}\frac{m^m}{m!} \tag{14}$$

Watson [27], proving a conjecture of Ramanujan, has shown that

$$e^{-m}\left[\sum_{i=0}^{m-1} \frac{m^i}{i!} + y\frac{m^m}{m!}\right] = \frac{1}{2}$$

where $1/3 \leq y \leq 1/2$. Using this and (14) completes the proof for this m. Define $K = \max\{K_m : m = 1,\ldots,50\}$. The theorem now holds for all $k > K$. ▮

Denote $k = n/d$, $p = 2^{-d}$ and $t = kp$. Ignoring (for simplicity's sake) issues of divisibility and integrability, divide the n variables into k blocks of d variables each. Let f be the boolean function such that $f = 1$ iff at least t blocks of variables are all 1. By Lemma 4.3, f tends to be balanced for sufficiently large n. It is also d-threshold as it is equivalent to a linear threshold function composed with conjunctions of d variables, themselves expressable as polynomials of degree d. An easy calculation yields

$$\mathbf{Inf}_i(f) = 2^{1-n}\binom{k-1}{t-1}(2^d-1)^{(k-t)} \quad i = 1,\ldots,n. \tag{15}$$

In the extreme case $d = 1$ we have $k = n$ and $t = n/2$, the majority function. At the other extreme $d = m(n)$, we have $k = n/m(n)$ and $t = 1$, the construction of [5]. In the intermediate range, (15) can be simplified to

$$\mathbf{Inf}_i(f) = \Theta(\sqrt{\frac{d}{2^d n}}) \quad i = 1,\ldots,n. \tag{16}$$

This shows that there exist balanced d-threshold functions with small influences, so that any lower bound on the influences of variables on balanced functions in $\mathcal{T}_n^d$ cannot exceed (16).

5. Upper bounds on the average sensitivity

We now present some upper bounds and conjectures on the average sensitivity of threshold functions.

It is well known [22,2] that for a linear threshold function f

$$\mathbf{AS}(f) \leq 2^{-n+1}\binom{n}{[n/2]}(n-[n/2]) = O(\sqrt{n}) \tag{17}$$

where $[x]$ is the integer part of x. Equality is obtained for the majority function. The correct order of magnitude may be obtained simply from the Fourier coefficients by applying Lemma 2.2:

$$\mathbf{AS}(f) = \sum_{i=1}^{n} \mathbf{Inf}_i(f) = \sum_{|I|=1} |\hat{f}(I)|.$$

Bounding this with the Cauchy-Schwartz inequality:

$$\mathbf{AS}(f) \leq \sqrt{n}\sqrt{\sum_{|I|=1} \hat{f}^2(I)} \leq \sqrt{n}.$$

The exact bound is related to the central binomial coefficient. Higher order threshold functions are more difficult to analyze. We conjecture :

Conjecture. Let f be a d-threshold function. Then

$$\mathbf{AS}(f) \leq 2^{-n+1}\sum_{k=0}^{d-1}\binom{n}{[(n-k)/2]}(n-[(n-k)/2]) \tag{18}$$

which is related to the sum of the d central binomial coefficients. For $d=o(\sqrt{n})$ the r.h.s. of (18) is $O(d\sqrt{n})$, but in order to obtain an average sensitivity of $n(1-o(1))$ it is necessary to increase d to $\Omega(n)$. The symmetric d-threshold function which cuts the middle d layers of the hypercube attains this bound. For the same reason, this is a tight upper bound on the average sensitivity of *symmetric* d-threshold functions (a symmetric binary function is $f(x_1,\ldots,x_n)=g(\sum_{i=1}^{n} x_i)$).

6. Related and open questions

The results presented here could be improved considerably once answers to the following questions are obtained:

1. For a d-threshold function, how do the Fourier coefficients of degree $\leq d$ determine those of higher degree?
2. For a d-threshold function f, where $d \geq 2$, how may $\mathbf{Inf}_i(f)$ be expressed as a function of the Fourier coefficients of degree $\leq d$?

After the submission of this paper, we were informed of a result complementary to ours obtained by Aspnes et al. [3]. They show that for any d-threshold function f:

$$|\hat{f}([n])| = O(d/\sqrt{n})$$

for constant d. This is obtained by the symmetric function cutting the middle d layers of the n-cube.

Recall the combinatorial interpretation of the average sensitivity of a d-threshold function $f = \mathrm{sgn}(P)$ as the number of edges of the n-cube cut by the polynomial P. A related question is the number of d-degree polynomials required to cut *all* the edges of the n-cube. For the linear ($d = 1$) case, M. Paterson [23] (see also remark in [22]) has established an upper bound of $5n/6$ by constructing 5 hyperplanes which cut all the edges of the 6-dimensional hypercube, contradicting a common belief that exactly n hyperplanes are required. J. Håstad has observed [11] that a lower bound on this quantity would also be a lower bound on the size of a depth two linear threshold circuit computing parity. A trivial lower bound of $\Omega(\sqrt{n})$ may be obtained immediately from (17), but it is believed to be much closer to $O(n)$. Paturi et al. [24] have shown that in the special case where the realizing weights are bounded by a polynomial in n, a lower bound of $\Omega(n/\log^2 n)$ holds.

However, if the realizing weights (the linear coefficients) are restricted to be positive, n hyperplanes are necessary and sufficient. This is shown easily by observing the path of n edges connecting the following vertices of the n-cube:

$$(-1,-1,-1,\dots,-1),(1,-1,-1,\dots,-1),(1,1,-1,\dots,-1),\dots,(1,1,1,\dots,1)$$

Any given hyperplane cuts exactly one of the edges, therefore at least n are required. This is also sufficient as the family of n hyperplanes with positive coefficients

$$\{\sum_{i=1}^{n} x_i + 2t = 0 : t = -[n/2],\dots,[n/2]\}$$

cuts all the edges.

Once again, the case $d > 1$ remains an open question.

Acknowlegdements. We would like to thank Mario Szegedy and Ilan Newman for interesting discussions during this work.

References

[1] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski: A learning algorithm for Boltzmann machines, *Cognitive Science* **9** (1985), 147–169.

[2] R. Ahlswede, and Z. Zhang: An identity in combinatorial extremal theory, *Advances in Mathematics* **80** (1990), 137–151.

[3] J. Aspnes, R. Beigel, M. Furst, and S. Rudich: The expressive power of voting polynomials, in: *Proceedings of the 23rd Symposium on the Theory of Computing*, 402–409, ACM, 1991.

[4] D. A. Mix Barrington, H. Straubing, and D. Thérien: Non-uniform automata over groups, *Information and Computation* **89** (1990), 109–132.

[5] M. BEN-OR, and N. LINIAL: Collective coin flipping, in: *Randomness and Computation*, (Micali S., editor). Academic Press, 1989.

[6] J. BOURGAIN: Walsh subspaces of l_p product spaces, in: *Seminaire D'Analyse Fonctionnelle*, 4.1–4.9. Ecole Polytechnique, Centre de Mathematiques, 1979-1980.

[7] J. BRUCK: Harmonic analysis of polynomial threshold functions, *SIAM Journal of Discrete Maths.* **3** (1990), 168–177.

[8] J. BRUCK: Polynomial threshold functions, AC^0 functions and spectral norms, *SIAM Journal of Computing* **21** (1) (1992), 33–42.

[9] C. GOTSMAN: On boolean functions, polynomials and algebraic threshold functions, Technical Report TR-89-18, Dept. of Computer Science, Hebrew University.

[10] U. HAAGERUP: The best constants in the Khintchine inequality, *Studia Mathematica* **70** (1982), 231–283.

[11] J. HÅSTAD: Personal communication, 1990.

[12] J. J. HOPFIELD: Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sc. USA* **79** (1982), 2554–2558.

[13] J. KAHN, G. KALAI, and N. LINIAL: The influence of variables on boolean functions, in: *Proceedings of the 29th Symposium on the Foundations of Computer Science*, 68–80, IEEE, 1988.

[14] J. KAHN, and R. MESHULAM: On mod p transversals, *Combinatorica* **11** (1) (1991), 17–22.

[15] Y. KATZNELSON: *An Introduction to Harmonic Analysis*, Wiley, 1968.

[16] A. KUSHILEVITZ, and Y. MANSOUR: Learning decision trees using the Fourier spectrum, in: *Proceedings of the 23rd Symposium on the Theory of Computing*, 455–464, ACM, 1991.

[17] Y. LE CUN: Une procedure d'apretissage pour reseau a seuil asymmetrique (A learning procedure for asymmetric threshold networks), in: *Proceedings of Cognitiva 85*, 599–604. Paris, 1985.

[18] R. J. LECHNER: Harmonic analysis of switching functions, in: *Recent Developments in Switching Theory*, (A. Mukhopadhyay, editor) Academic Press, 1971.

[19] N. LINIAL, Y. MANSOUR, and N. NISAN: Constant depth circuits, Fourier transforms and learnability, in: *Proceedings of the 30th Symposium on the Foundations of Computer Science*, 574–579, IEEE, 1989.

[20] M. MINSKY, and S. PAPERT: *Perceptrons: An Introduction to Computational Geometry*, MIT Press, 1968.

[21] I. NINOMIYA: A theory of the coordinate representation of switching functions (review), *IEEE Transactions on Electronic Computers* **12** (1963), 152.

[22] P. O'NEIL: Hyperplane cuts of an n-cube, *Discrete Maths.* **1** (1971), 193–195.

[23] M. PATERSON: Personal communication, 1990.

[24] R. PATURI, and M. SAKS: On threshold circuits for parity, in: *Proceedings of the 31st Symposium on the Foundations of Computer Science*, 397–404, IEEE, 1990.

[25] D. RUMELHART, G. HINTON, and J. WILLIAMS: Learning internal representations by error propogation, in: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol 1.*, (Rumelhart J., McLelland J., and the PDP research group, editors) M.I.T. Press, 1986.

[26] J. V. USPENSKY: *Introduction to Mathematical Probability*, McGraw-Hill, 1937.

[27] G. N. WATSON: Theorems stated by Ramanujan (v): Approximations connected with e^x, *Proc. London Math. Society (2nd Series)* **29** (1929), 293–308.

[28] R. WIDNER: Chow parameters in threshold logic, *Journal of the ACM* **18** (1971), 265–289.

Craig Gotsman

Department of Computer Science
Technion — Israel Institute of Technology
Haifa 32000
Israel
gotsman@cs.technion.ac.il

Nathan Linial

Department of Computer Science
The Hebrew University
Jerusalem 91904
Israel
nati@cs.huji.ac.il