Vector Analysis of Threshold Functions*

Vwani Roychowdhury† Kai-Yeung Siu‡ Alon Orlitsky§ Thomas Kailath¶

Abstract

Viewing n-variable Boolean functions as vectors in $\mathcal{R}^{2^n}$, we invoke basic tools from linear algebra and linear programming to derive new results on the realizability of Boolean functions using threshold gates. Using this approach, we obtain: (1) a lower bound on the number of input functions required by a threshold gate implementing a given function; (2) a lower bound on the error incurred when a Boolean function is approximated by a linear combination of a set of functions; (3) a limit on the effectiveness of a well known lower-bound technique (based on computing correlations among Boolean functions) for the depth of threshold circuits implementing Boolean functions; (4) a construction showing that every Boolean function f of n input variables is a threshold function of polynomially many input functions, none of which is significantly correlated with f; (5) generalizations of some known results on threshold-circuit complexity, particularly those that are based on spectral analysis.

*Appeared in Information and Computation 120(1):22-31, July 1995. This copy printed on November 20, 1995.

†with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

‡with the Department of Electrical and Computer Engineering, University of California, Irvine, CA 92717.

§with AT&T Bell Laboratories, Murray Hill, NJ 07974.

¶with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, CA 94305.

1 Introduction

1.1 Background

An S-input threshold gate is characterized by S real weights $w_1, \ldots, w_S$. It takes S inputs: $x_1, \ldots, x_S$, each either $+1$ or -1, and outputs $+1$ if the linear combination $\sum_{i=1}^{S} w_i x_i$ is positive and -1 if the linear combination is negative.

Circuits consisting of threshold gates are called threshold circuits and have received considerable attention [20, 15, 17, 19, 4, 31, 29, 1, 8]. Threshold circuits have been recently shown to compute several functions of practical interest (including: Parity, Addition, Multiplication, Division, and Comparison) with fewer gates and smaller depth than conventional circuits using AND, OR, and NOT gates [5, 24, 2, 19, 27, 28]. However, some simple questions remain unanswered. It is not known, for example, whether there is a function that can be computed by a depth-3 threshold circuit with polynomially many gates but cannot be computed by any depth-2 circuit with polynomially many threshold gates. At an even more basic level, there is no efficient (unless P=NP) way to determine whether any Boolean function given in a disjunctive normal form can be computed by a single threshold gate [12, 23].

An S-input threshold gate corresponds to a hyperplane in $\mathcal{R}^S$. This geometric interpretation has been used for example to count the number of Boolean functions computable by a single threshold gate [7], and also to determine functions that cannot be implemented by a single threshold gate. However, threshold circuits of depth two or more do not carry a simple geometric interpretation in $\mathcal{R}^S$. The inputs to gates in the second level are themselves threshold functions, hence the linear combination computed at the second level can be a non-linear function of the inputs. Lacking a geometric view, researchers [6, 4] have used indirect approaches, applying spectral-analysis techniques to analyze threshold gates. These techniques, apart from their complexity, restricted the input functions of the gates to be of very special types: input variables, or parities of the input variables, thus not applying even to depth-two circuits.

In this paper, we outline a simple geometric approach to characterize the input/output relationship of a threshold gate. Using this approach, we obtain: (1) a lower bound on the number of input functions required by a threshold gate implementing a given function; (2) a lower bound on the error incurred when a Boolean function is approximated by a linear combination of a set of functions; (3) a limit on the effectiveness of a well known lower-bound technique (based on computing correlations among Boolean functions) for the depth of threshold circuits implementing Boolean functions; (4) a construction showing that every Boolean function f of n input variables is a threshold function of polynomially many input functions, none of which is significantly correlated with f; (5) generalizations of some known results on threshold-circuit complexity, particularly those that are based on spectral analysis.

1.2 Framework

An *n-variable Boolean function* is a mapping $f : \{-1,1\}^n \rightarrow \{-1,1\}$. We view f as a (column) vector in $\mathcal{R}^{2^n}$. Each of f's 2^n components is either -1 or $+1$ and represents $f(x)$ for a distinct value assignment x of the n Boolean variables. We view the weights of an S-input threshold gate as a *weight vector* $\mathbf{w} = (w_1, \cdots, w_S)^T$ in $\mathcal{R}^S$.

Let the functions $f_1, \ldots, f_S$ be the inputs of a threshold gate with weight vector $\mathbf{w}$. The gate *computes* a function f (or f is the *output* of the gate) if the following vector equation holds:

$$f = \operatorname{sgn}\left(\sum_{i=1}^{S} f_i w_i\right) \tag{1}$$

where

$$\operatorname{sgn}(x) \stackrel{\text{def}}{=} \begin{cases} +1 & \text{if } x > 0, \\ -1 & \text{if } x < 0, \\ \text{undefined} & \text{if } x = 0. \end{cases}$$

Note that this definition requires that all components of $\sum_{i=1}^{S} f_i w_i$ be nonzero. It is convenient to write Equation (1) in matrix form:

$$f = \operatorname{sgn}(Y\mathbf{w}),$$

where the *input matrix*

$$Y = [f_1 \cdots f_S]$$

is a 2^n by S matrix whose columns are the input functions. The function f is a *threshold function* of $f_1, \ldots, f_S$ if there exists a threshold gate (namely, a weight vector $\mathbf{w}$) with inputs $f_1, \ldots, f_S$ that computes f.

Geometrically, each function f, being a ± 1 vector in $\mathcal{R}^{2^n}$, determines an *orthant* in $\mathcal{R}^{2^n}$ — the set of vectors whose non-zero coordinates agree in sign with the corresponding coordinates of f. The orthant's *interior* consists of all vectors in the orthant with no zero coordinates.

In this interpretation, f is the output of a threshold gate whose input functions are $f_1, \ldots, f_S$ if and only if the linear combination $Y\mathbf{w} = \sum_{i=1}^{S} f_i w_i$ defined by the gate lies in the interior of f's orthant. This view forms the basis of our results. Combined with some other basic linear-algebraic observations, it enables us to prove various threshold-circuit results as described in the next subsection.

The *correlation* of two n-variable Boolean functions f_1 and f_2 is

$$C_{f_1 f_2} \stackrel{\text{def}}{=} (f_1^T f_2)/2^n \ .$$

The functions f_1 and f_2 are *uncorrelated*, or *orthogonal*, if $C_{f_1 f_2} = 0$. Note that $C_{f_1 f_2} = 1 - d_H(f_1, f_2)/2^{n-1}$, where $d_H(f_1, f_2)$ is the *Hamming distance* between f_1 and f_2. Thus, correlation can be interpreted as a measure of how 'close' the two functions are.

1.3 Results

If two vectors lie in the same orthant, their inner product is non-negative. If one of the vectors is in the orthant's interior, and the other is nonzero, then the inner product is positive. We therefore have the following "folk theorem:"

Lemma 1 If f is orthogonal to the input functions $f_1, \ldots, f_S$, then f is not a threshold function of $f_1, \ldots, f_S$.

Proof: If f is orthogonal to all of $f_1, \ldots, f_S$, then f is orthogonal to any linear combination $Y\mathbf{w}$ of these functions. By the observations preceding the lemma, $Y\mathbf{w}$ cannot be in f's orthant, hence f cannot be a threshold function of $f_1, \ldots, f_S$. □

Fix the input functions $f_1, \ldots, f_S$ to a threshold gate. The *correlation vector* of a function f, with the input functions is

$$C_{fY} \stackrel{\text{def}}{=} (Y^T f)/2^n \ = \ [C_{ff_1} \, C_{ff_2} \, \cdots \, C_{ff_S}]^T.$$

The correlation of two n-variable Boolean functions can assume $2^n + 1$ different values, hence the correlation vector can assume at most $(2^n + 1)^S$ different values. Inevitably, many of the 2^{2^n} different functions share the same correlation vector.

In Section 2 we prove a uniqueness property: if f is a threshold function of $f_1, \ldots, f_S$, then the correlation vector C_{fY} is achieved only by f — for any other function g, $C_{gY} \neq C_{fY}$. This implies, for example, that any set of S input functions can give rise to at most $(2^n+1)^S$ different threshold functions.

The special case of the uniqueness property where the functions $f_1,\ldots,f_S$ are the input variables (or a constant function) had been proven in [6] (see also [15]). The proof used spectral-analysis tools such as Parseval's theorem and relied on the mutual orthogonality of the input functions (namely, $C_{x_i,x_j}=0$ for all $i\neq j$). Another special case where the input functions are parities of the input variables was proven in [4]. Essentially the same proof was used, again relying on the input functions being mutually orthogonal. However, in threshold circuits of depth larger than one, the input functions to the gates in the second layer are not necessarily orthogonal; hence these uniqueness results do not apply. Our proof shows that the spectral-analysis tools and assumptions are not needed for the uniqueness property to hold, thereby simplifying the proof and showing that the input functions need not be mutually orthogonal: the uniqueness property holds for all collections of functions.

Lemma 1 showed that if $C_{fY}=0$ then f is not a threshold function of $f_1,\ldots,f_S$. In Section 3 we consider the case where $C_{fY}\neq 0$. Define the *generalized spectrum* of f with respect to $f_1,\ldots,f_S$ (name partially justified below) to be the S-dimensional vector:

$$\beta \stackrel{\text{def}}{=} (\beta_1,\ldots,\beta_S)^T \stackrel{\text{def}}{=} (Y^TY)^{-1}Y^Tf\ .$$

We show that if f is a threshold function of $f_1,\ldots,f_S$, then

$$S \geq \frac{1}{\max\{|\beta_i| : 1\leq i\leq S\}}\ . \tag{2}$$

This provides a way to lower bound the number of input functions. Specifically, if each β_i is exponentially small in n (the number of variables) then S must be exponentially large.

If the input functions are mutually orthogonal then $Y^TY=2^nI_S$ (the $S\times S$ identity matrix), hence

$$\beta = \frac{1}{2^n}Y^Tf = C_{fY}\ ,$$

and (2) reduces to

$$S \geq \frac{1}{\max\{|C_{ff_i}| : 1\leq i\leq S\}}\ . \tag{3}$$

We also prove an analog of this result for input functions that are asymptotically mutually orthogonal.

The special case of mutually-orthogonal input functions where the functions are parities of input variables was proved in [4]. As with the uniqueness property, the proof used spectral-analysis tools and assumptions. Using only basic linear-algebraic techniques, the geometric approach shows that these assumptions are not necessary, thereby simplifying the proofs and generalizing the results to arbitrary sets of input functions.

In Section 4 we use the generalized spectrum to lower bound the *error* incurred when a Boolean function is approximated by a linear combination of a set of functions. We apply the bound to show that the Majority function cannot be closely approximated by a sparse polynomial. Specifically, we show that if a polynomial of the input variables with only polynomially many (in n) monomials is used to approximate the n variable Majority function then the approximation error is $\Omega(1/(\log\log n)^{3/2})$.

Our lower bounds use the correlation coefficients C_{ff_i} to lower bound the number of input functions $f_1,\ldots,f_S$ required by a threshold gate computing f. An equivalent approach was used by Hajnal et. al. [11] to lower bound the depth of a threshold circuit computing the Inner-product-mod-2 function. In Section 5, we investigate this *correlation technique* in more detail and prove some limits to its effectiveness. We show that if the input functions are not mutually orthogonal then the number of input functions need not be inversely proportional to the largest correlation coefficient. We provide a construction showing that any n-variable function f is a threshold function of $2n$ input functions each having an exponentially small correlation with f.

2 Uniqueness

The correlation between two n-variable functions is a multiple of $2^{-(n-1)}$ and is bounded between -1 and 1, hence can assume 2^n+1 values. Given $Y=[f_1,\ldots,f_S]$, the correlation vector $C_{fY} = [C_{ff_1},\ldots,C_{ff_S}]^T$ can therefore assume at most $(2^n+1)^S$ different values. There are 2^{2^n} Boolean functions of n Boolean variables, hence many share the same correlation vector. However, the next theorem shows that a threshold function of $f_1,\ldots,f_S$ does not share its correlation vector with any other function.

Theorem 1 [**Uniqueness**] Let f be a threshold function of $f_1,\ldots,f_S$. For every Boolean function $g\neq f$,

$$C_{gY}\neq C_{fY}.$$

Proof: Let $(v)_i$ denote the i^{th} entry of a vector v. For any Boolean function g and for all $i \in \{1, \ldots, 2^n\}$,

$$(f-g)_i = \begin{cases} 0 & \text{if } (f)_i = (g)_i, \\ 2(f)_i & \text{if } (f)_i \neq (g)_i. \end{cases}$$

By assumption, there is a weight vector $\mathbf{w}$ such that $f = \text{sgn}(Y\mathbf{w})$. Hence, whenever $(f)_i \neq (g)_i$,

$$\text{sgn}(f-g)_i = \text{sgn}(f)_i = \text{sgn}(Y\mathbf{w})_i.$$

Moreover, if $f \neq g$, there must be an index i such that $(f)_i \neq (g)_i$, and by definition $(Y\mathbf{w})_i \neq 0$ for all i. Hence, $(f-g)^T Y\mathbf{w} > 0$, which implies $f^T Y \neq g^T Y$. □

The proof has a simple geometric interpretation. If f is a threshold function of $f_1, \ldots, f_S$ then some linear combination $Y\mathbf{w}$ of these functions lies in the interior of the orthant in $\mathcal{R}^{2^n}$ determined by f. But for any $g \neq f$, the nonzero vector $(f-g)$ lies in the same orthant. Hence, $(f-g)^T Y\mathbf{w} > 0$, which implies that $f^T Y \neq g^T Y$.

As an immediate consequence we derive a bound on the number of threshold functions of any set of input functions.

Corollary 1 There are at most $(2^n+1)^S$ threshold functions of any set of S input functions. □

The innovation of Corollary 1 lies mostly in the simplicity of its proof. With considerably more work, one can obtain bounds that have the same asymptotic behavior when $S \ll 2^n$, but are stronger when S is comparable to 2^n [7, 20].

The following are two further observations regarding the uniqueness property:

1. Theorem 1 holds even if the functions $f_1, \ldots, f_S$ are real valued (not restricted to ± 1) and if their domain is of arbitrary dimension (not necessarily a power of 2).

2. The converse of Theorem 1 is *not* true. Fix a set $f_1, \ldots, f_S$ of input functions, and let f be a function satisfying $C_{fY} \neq C_{gY}$ for all $g \neq f$. Then f is *not* necessarily a threshold function of $f_1, \ldots, f_S$.

   To see that, let f be the parity of *all* the input variables, and let the input functions $f_1, \ldots, f_S$ be all the parity functions of subsets of the variables, except for f. Thus $S = 2^n - 1$. It is easy to verify that $f^T Y = 0$ while $g^T Y \neq 0$ for any $g \neq f$. Yet, by Lemma 1, the orthogonality of f to $f_1, \ldots, f_S$ implies that f is not a threshold function of them.

3 Generalized Spectrum

We define the generalized spectrum of a Boolean function f with respect to a set of input functions and use it to lower bound the number of input functions required by a threshold gate computing f. When the input functions are mutually orthogonal, the generalized spectrum of f reduces to its standard spectrum.

As in the introduction, $Y = [f_1 \ f_2 \ \cdots \ f_S]$ is the $2^n \times S$ input matrix whose columns are the input functions to a threshold gate. A basic linear-algebraic result guarantees that any function f can be expressed as:

$$f = Y\beta + Z, \tag{4}$$

where $\beta = [\beta_1 \ \beta_2 \ \cdots \ \beta_S]^T$ is an S-dimensional column vector, and $Z^T Y = 0$. Without loss of generality assume that Y is of full column rank. Then β can be computed as follows:

$$\beta = (Y^T Y)^{-1} Y^T f = 2^n (Y^T Y)^{-1} C_{fY} \tag{5}$$

where $C_{fY} = [C_{ff_1} \ C_{ff_2} \ \cdots \ C_{ff_S}]^T$ is the correlation vector defined in the introduction.

Geometrically, $Y\beta$ represents the orthogonal projection of the vector f onto the subspace spanned by the input vectors f_i, and Z is orthogonal to that subspace. For an interpretation of β, consider the case where the input functions form an orthogonal basis of $\mathcal{R}^{2^n}$. Such a basis is formed, for example, by the 2^n parity functions of subsets of the n input variables [30]. Then $f_i^T f_j = 0$ for $i \neq j$ and $f_i^T f_i = 2^n$. Hence, $2^n (Y^T Y)^{-1}$ is the $S \times S$ identity matrix and $\beta = C_{fY}$, namely, the β_i's are the corresponding *spectral coefficients*. In general, if the input functions f_i are mutually orthogonal then there exists an orthogonal basis $\mathcal{B} = \{f_1, f_2, \cdots, f_S, \tilde{f}_{S+1}, \cdots, \tilde{f}_{2^n}\}$, where $\tilde{f}_i \in \mathcal{R}^{2^n}$ (not necessarily restricted to have ± 1 entries). Since the functions f_i are mutually orthogonal, we again get $\beta = C_{fY}$, or β_i are the spectral coefficients corresponding to the orthogonal basis $\mathcal{B}$. For that reason, we call β the *generalized spectrum* of the output function f with respect to $f_1, \ldots, f_S$.

The next theorem uses the generalized spectral coefficients to lower bound the number of input functions required by a threshold gate computing a function.

Theorem 2 **[Spectral Bound]** If f is a threshold function of $f_1, \ldots, f_S$, then

$$\sum_{i=1}^{S} |\beta_i| \geq 1,$$

implying

$$S \geq \frac{1}{\max\{|\beta_i| : 1 \leq i \leq S\}}.$$

Proof: Consider the decomposition $f = Y\beta + Z$ described in (4). Suppose that $\sum_{i=1}^{S} |\beta_i| < 1$. Then $|(Y\beta)_i| < 1$ for all i. Thus, $\mathrm{sgn}(Z) = \mathrm{sgn}(f - Y\beta) = f$. However, by assumption, $f = \mathrm{sgn}(Y\mathbf{w})$ for some $\mathbf{w}$, which implies that $Z^T Y\mathbf{w} > 0$. This is a contradiction since $Z^T Y = 0$. □

Geometrically, if f is a threshold function of $f_1, \ldots, f_S$ then some linear combination $Y\mathbf{w}$ of these functions lies in the interior of the orthant in $\mathcal{R}^{2^n}$ determined by f. If $|(Y\beta)_i| < 1$ for all i, then the vector Z lies in the interior of f's orthant. However, this leads to a contradiction as, on the one hand Z is orthogonal to $f_1, \ldots, f_S$, while, on the other hand, it lies in the same orthant as the linear combination $Y\mathbf{w}$ of $f_1, \ldots, f_S$.

If the input functions are mutually orthogonal, *i.e.,* $C_{f_i f_j} = 0$ for all $i \neq j$, then $\beta_i = C_{ff_i}$. We therefore have:

Corollary 2 If f is a threshold function of $f_1, \ldots, f_S$, and if the f_i's are mutually orthogonal then

$$S \geq \frac{1}{\max\{|C_{ff_i}| : 1 \leq i \leq S\}}.$$

□

Next we prove a variation of Corollary 2 for the case where the input functions are asymptotically orthogonal. For every integer n, let f be a desired output function, let $\{f_1, \ldots, f_{S_n}\}$ (all Boolean functions of n Boolean variables — the index n is omitted from the function notation) be a set of input functions , and let

$$\hat{v}_n = \max\{|C_{f_i f_j}| : 1 \leq i < j \leq S_n\}.$$

We shall define the set of input functions $\{f_1, \ldots, f_{S_n}\}$ to be *strongly* asymptotically orthogonal if there exists an ϵ, $0 \leq \epsilon < 1$, such that for sufficiently large n,

$$\hat{v}_n < \epsilon / S_n.$$

That is the input functions are asymptotically mutually orthogonal, and their mutual correlation is bounded above by ϵ / S_n.

Define

$$\hat{C}_n \stackrel{\text{def}}{=} \max\{|C_{ff_i}| : 1 \leq i \leq S_n\}.$$

Theorem 3 If for every n, f is a linear threshold function of the strongly asymptotically orthogonal functions $f_1, \ldots, f_{S_n}$, then

$$S_n = \Omega(\frac{1}{\hat{C}_n}) \ .$$

Proof: One can represent $(Y^T Y)/2^n$ as follows:

$$\frac{1}{2^n} Y^T Y = I_{S_n} - E_{S_n},$$

where I_{S_n} is an $S_n \times S_n$ identity matrix and E_{S_n} is an $S_n \times S_n$ matrix, such that $(E_{S_n})_{ii} = 0$ and $\max|(E_{S_n})_{ij}| \leq \hat{v}_n$. Since $\{f_1, \ldots, f_{S_n}\}$ is strongly asymptotically orthogonal, we have for sufficiently large n, $\max|(E_{S_n})_{ij}| \leq \hat{v}_n \leq \epsilon/S_n$. Since $0 \leq \epsilon < 1$, one can now show that for sufficiently large n:
(1) $\max|(E^k_{S_n})_{ij}| \leq \epsilon^k/S_n$, and
(2) $(I_{S_n} - E_{S_n})^{-1} = \sum_{k=0}^{\infty} E^k_{S_n} = (I_{S_n} - F_{S_n})$, where

$$\max|(F_{S_n})_{ij}| \leq \frac{1}{S_n} \times \frac{\epsilon}{1-\epsilon}. \tag{6}$$

Now from Eqn. (5) we get

$$\beta = (\frac{1}{2^n} Y^T Y)^{-1} C_{fY} = (I_{S_n} - F_{S_n}) C_{fY}.$$

If $\hat{\beta}_n = \max\{|\beta_i| : 1 \leq i \leq S_n\}$ then using the upper bound on $\max|(F_{S_n})_{ij}|$ in Eqn. (6), it follows that

$$\hat{\beta}_n \leq \hat{C}_n (1 + \frac{\epsilon}{S_n(1-\epsilon)}).$$

The result of the theorem follows from the fact that $\hat{\beta}_n \geq 1/S_n$ (Theorem 2). □

3.1 Characterizations with Generalized L_1 Spectral Norms

Let $g_1, \ldots, g_{2^n}$ be a (not necessary orthogonal) basis for $\mathcal{F}$, the set of functions from $\{1,-1\}^n$ to $\{1,-1\}$. Every Boolean function $f \in \mathcal{F}$ can be expressed uniquely as

$$f = \sum_{i=1}^{2^n} \beta_i g_i$$

where β_i are the generalized spectral coefficients defined in (5).

We define the *generalized L_1 spectral norm of f with respect to the basis* $\{g_i | i = 1, \ldots, 2^n\}$ to be:

$$||f||_g \stackrel{\text{def}}{=} \sum_{i=1}^{2^n} |\beta_i| \ .$$

The generalized spectral norm reduces to the spectral norm when the g_i's are the parity functions. In [27] it was shown that if the spectral norm of a function is polynomially bounded, then it can be expressed as a threshold function of polynomially many parity functions or monomials. We present here an immediate generalization of this result in terms of the generalized spectral coefficients (proof omitted).

Theorem 4 If the generalized L_1 spectral norm of f is bounded by a polynomial, *i.e.*, $||f||_g \leq n^c$ for some $c > 0$, then for any $k > 0$, there exists $S \subset \{1, ..., 2^n\}$ such that $|S| \leq n^p$ (i.e., S has only polynomially many elements in it) and

$$|f(X) - \sum_{i \in S} w_{g_i} g_i(X)| \leq n^{-k} \ .$$

Consequently, $f(X) = \text{sgn}(\sum_{i \in S} w_{g_i} g_i(X))$. □

4 Spectral Approximation of Boolean Functions

We present results on approximating Boolean functions by a linear combination of a set of functions. A lower bound on the *approximation error* is derived in terms of the generalized spectral coefficients. We then apply this lower bound to the special case of approximating Boolean functions by polynomials and show that relatively simple functions, such as Majority, cannot be closely approximated by sparse polynomials. In particular, we show that if a polynomial of the input variables with only polynomially many (in n) monomials is used to approximate the n variable Majority function then the approximation error is $\Omega(1/(\log \log n)^{3/2})$. This provides a direct approach (based on the spectrum of the Majority function) for proving lower bounds on the approximation error and improves approximation results reported in [27].

Given the Boolean functions f and $f_1, \ldots, f_S$, the *approximation error*, $0 \leq \epsilon_{fY}$, is defined as

$$\epsilon_{fY} = \min\{||f - Y\mathbf{w}||_\infty \ : \ \mathbf{w} \in \mathcal{R}^S\} \ .$$

The set of functions $Y = [f_1, \cdots, f_S]$ is said to *approximate* f, if $\epsilon_{fY} < 1$.

Note that a set of functions $Y = [f_1 \cdots f_S]$ approximates a function f (*i.e.* , $0 \leq \epsilon_{fY} < 1$), if and only if f is a threshold function of $f_1, \cdots, f_S$.

Theorem 5 If a Boolean function f is approximated by a set of S other functions, $f_1, f_2, \cdots, f_S$, then

$$4\epsilon_{fY} \geq 1 - \frac{||Y\beta||^2}{2^n}$$

where, $Y = [f_1 \cdots f_S]$ and β is the generalized spectrum as defined in (5) and $||\cdot||$ is the L_2 norm.

Proof: Let $X = Y\mathbf{w}$, such that $|(f - X)_i| \leq \epsilon_{fY} < 1$. Since $2^n C_{fY} = (Y^T Y)\beta$, we obtain

$$f^T X = f^T Y\mathbf{w} = 2^n C_{fY}^T \mathbf{w} = \beta^T Y^T Y \mathbf{w} = (Y\beta)^T X.$$

Using Cauchy-Schwarz inequality we get,

$$f^T X \leq ||Y\beta|| \, ||X||.$$

Hence,

$$||Y\beta|| \geq f^T X / ||X||.$$

However, we know that $f^T X = \sum_{i=1}^{2^n} |X_i| \geq (1 - \epsilon_{fY})2^n$, and $||X||^2 = \sum_{i=1}^{2^n} X_i^2 \leq 2^n(1 + \epsilon_{fY})^2$. Hence we obtain,

$$\frac{||Y\beta||^2}{2^n} \geq (1 - \epsilon_{fY})^2/(1 + \epsilon_{fY})^2.$$

Now the theorem follows by observing that

$$4\epsilon_{fY} \geq 1 - \frac{(1 - \epsilon_{fY})^2}{(1 + \epsilon_{fY})^2} .$$

□

If we restrict to the case where columns of Y are orthogonal, then we have $\beta = C_{fY}$, and $||Y\beta||^2 = 2^n ||C_{fY}||^2$; hence, we obtain

$$4\epsilon_{fY} \geq 1 - ||C_{fY}||^2. \tag{7}$$

We now apply Theorem 5 to the more specific case of polynomial approximation of Boolean functions. From the spectral representation theory [18, 14, 21, 10, 30, 25] we know that any Boolean function, $f(X)$, of n variables $X_1, \cdots X_n$, can be written as $f(X_1, \cdots, X_n) = \sum_{\alpha \in \{0,1\}^n} a_\alpha X^\alpha$, where a_α are the spectral coefficients, and X^α are the monomials. Each monomial X^α is a parity function of the appropriate number of input variables (*i.e.* , those variables for which $\alpha_i = 1$). Since the parity functions are mutually orthogonal, we have $a_\alpha = C_{f\phi_\alpha}$, where $C_{f\phi_\alpha}$ denotes the correlation of function f with the parity function (or monomial) X^α.

If $f(X)$ is approximated by a polynomial, then the following corollary is directly implied by Theorem 5.

Corollary 3 If a function f is approximated by a polynomial where the monomials are chosen from the set $\{X^\alpha \ : \ \alpha \in Q, \ Q \subseteq \{0,1\}^n\}$, then the approximation error $0 \leq \epsilon_p < 1$ satisfies the following:

$$4\epsilon_p \ \geq \ 1 - \sum_{\alpha \in Q} a_\alpha^2 \ = \ \sum_{\alpha \notin Q} a_\alpha^2 \ .$$

□

Thus, the approximation error is bounded below by the total *spectral power* concentrated in the monomials that are not included in the approximation. This complements a result due to Linial *et. al.* [16].

A case of particular interest is when only polynomially many monomials are used to approximate a given function. A polynomial of n variables is a *sparse polynomial* if the number of monomials in it is polynomially bounded in n.

The class $\widetilde{SP}$ consists of all Boolean functions that can be closely approximated by sparse polynomials, more precisely, given any $f(X) \in \widetilde{SP}$, $k > 0$, there exist polynomially many (in n) c_α's such that

$$|f(X) - \sum c_\alpha X^\alpha| < n^{-k} \text{ for all } X \in \{1,-1\}^n \ .$$

$\widetilde{SP}$ constitutes a rich class of functions and contains all functions with polynomially bounded spectral norm [27]. Examples of functions in $\widetilde{SP}$ include functions such as AND, OR, Comparison and Addition.

Next we use our lower bound to show that the Majority function does not belong in $\widetilde{SP}$. The Majority function, for odd n, is defined as follows:

$$\text{Majority}(X_1, \cdots X_n) \ = \ \text{sgn}(\sum_{i=1}^{n} X_i).$$

Corollary 4 If the Majority function is approximated by a sparse polynomial then the approximation error ϵ_M is $\Omega(1/(\log\log n)^{3/2})$.

Proof: For any given $\alpha \in \{0,1\}^n$, let $|\alpha| = \sum_{i=1}^n \alpha_i$, *i.e.* , $|\alpha|$ denotes the number of 1's in α. For any monomial X^α, its degree is defined to be $|\alpha|$. Now, the spectral coefficients of the Majority function can be listed as follows (note that from symmetry, the monomials with the same degree have the same spectral coefficients) [3]:

$$a_\alpha = \begin{cases} 0 & \text{if } |\alpha| \text{ is even,} \\ \frac{2(-1)^{(|\alpha|-1)/2}}{2^n} \frac{(|\alpha|-1)!}{(\frac{|\alpha|-1}{2})!} \frac{(n-|\alpha|)!}{(\frac{n-|\alpha|}{2})!\,(\frac{n-1}{2})!} & \text{if } |\alpha| \text{ is odd.} \end{cases}$$

Now consider the spectral coefficients for which $|\alpha| = \log\log n$, and assume without loss of generality that $\log\log n$ is odd. There are $\binom{n}{\log\log n}$ of such spectral coefficients and they all have the same value. Hence, the spectral power concentrated in these coefficients is

$$\sum_{\alpha \ni |\alpha| = \log\log n} a_\alpha^2 = \binom{n}{\log\log n} a^2_{|\alpha|=\log\log n} = \Omega(1/(\log\log n)^{3/2}).$$

Suppose that a sparse polynomial, $R(X)$, approximating the Majority function, has less than n^c monomials for some constant c. Then applying Corollary 3, we get

$$4\epsilon_M \geq \sum_{X^\alpha \notin R} a_\alpha^2 \geq \sum_{\alpha \ni |\alpha| = \log\log n;\ X^\alpha \notin R} a_\alpha^2 \ .$$

There are $\binom{n}{\log\log n}$ monomials of degree $\log\log n$, and at least a fraction of $(1 - \frac{n^c}{\binom{n}{\log\log n}})$ of them do not appear in R; hence,

$$4\epsilon_M \geq \Omega(1/(\log\log n)^{3/2}(1 - \frac{n^c}{\binom{n}{\log\log n}})) = \Omega(1/(\log\log n)^{3/2}) \ .$$

□

The above corollary uses the spectrum of the Majority function to show that it is not in $\widetilde{SP}$. An indirect proof was given in [27] that showed that the approximation error is $\Omega(1/n)$; our direct approach improves the lower bound considerably.

5 On the Correlation Technique

Section 3 showed that the correlation of a function f with a set of input functions can be used to lower bound the number of input functions required to compute f using a threshold gate. In particular, for mutually orthogonal (or strongly asymptotically orthogonal) input functions, if the correlation of f with each input function is exponentially small (in the number of variables) then the set of input functions must be exponentially large. In this section we investigate this correlation technique when the input functions are not orthogonal.

The next proposition shows that for *any* set of input functions, if the gate's weights are polynomially bounded integers (in the number of variables) then exponentially small correlations still imply that exponentially many input functions are required. The proposition was proven in [11]. We restate the proof using the terminology used here.

Proposition 1 ([11]) Let f and $f_1, \ldots, f_S$ be Boolean functions such that

$$f = \mathrm{sgn}(\sum_{i=1}^{S} f_i w_i)$$

where the weights are integers. Then,

$$S \geq \frac{1}{\hat{w}\hat{C}}$$

where $\hat{w} \stackrel{\text{def}}{=} \max\{|w_i| : 1 \leq i \leq S\}$ and $\hat{C} \stackrel{\text{def}}{=} \max\{|C_{ff_i}| : 1 \leq i \leq S\}$.

Proof: Since all weights are integers, $Y\mathbf{w}$ is an integer vector, agreeing in sign with f. Hence,

$$2^n \leq f^T(Y\mathbf{w}) = (f^T Y)\mathbf{w} = 2^n C_{fY}^T \mathbf{w} = 2^n \sum_{i=1}^{S} C_{ff_i} w_i \leq 2^n S \hat{C} \hat{w} \ ,$$

implying the result. □

Remark: [11] applied the above proposition to establish the first known separation result for threshold circuits. Let LT_d be the set of functions computable by depth-d threshold circuits with size polynomially bounded in the number of input variables, and let $\widehat{LT}_d$ be a subclass of LT_d, where the weights of the threshold gates are restricted to be integers, which are polynomially bounded in the number of input variables.

The inner-product-mod-2 function of two n-bit binary sequences x and y is defined by:

$$IP_2(x, y) \ = \ \bigoplus_{i=1}^{n} x_i \wedge y_i \ .$$

By showing that IP_2 has exponentially small correlation with any function in LT_1, and applying the proposition, [11] showed that IP_2 requires exponentially many gates when implemented by a depth-2 threshold circuit with polynomially-bounded integer weights, hence

$IP_2 \notin \widehat{LT}_2$. Since IP_2 can be computed by a linear-size depth-3 threshold circuit, it follows that $\widehat{LT}_2 \subsetneq \widehat{LT}_3$. In a more recent paper [13] an alternate geometric approach is used to show that $IP_2 \notin \widehat{LT}_2$; in fact related results in [13] imply that $IP_m \notin \widehat{LT}_2$.

We should also note that the computational power of threshold circuits where all weights are polynomially bounded integers has recently been explored [27, 8, 9]. For example, it is shown in [8, 9] that $LT_1 \subset \widehat{LT}_2$, and in general $LT_d \subseteq \widehat{LT}_{d+1}$. The method of correlation and other related techniques have been used in [8] to obtain these and other related separation results.

We have therefore seen that if f is to be computed by a threshold gate with input functions $f_1, \ldots, f_S$, then exponentially small correlation of f with each f_i implies that S is exponentially large in any of the following cases:

1. The input functions are orthogonal.

2. The input functions are strongly asymptotically orthogonal.

3. The input functions are arbitrary and the weights of the threshold gate are integers bounded by a polynomial in the number of inputs.

It is natural to ask whether the same holds for arbitrary input functions when the weights are not restricted. Such a result would show, for example, that any depth-2 threshold circuit for IP_2 requires exponentially many gates (regardless of the weights).

In this section we show that the above is *not* true. We prove that any n-variable (for n even) function f is a threshold function of $2n$ input functions such that the correlation of f with each input function is exponentially small (in fact, with one input function the correlation is $2^{-(n-1)}$ — the smallest nonzero value, and with all other input functions, the correlation is zero). Note that Lemma 1 implies that f must have a nonzero correlation with at least one input function.

The proof uses a well known linear-programming result, described in the next lemma. Let $Y = [f_1\ f_2\ \cdots f_S]$. Define Y^f to be the matrix obtained by negating the rows of Y that correspond to the -1 entries of f, that is,

$$Y^f_{ij} = f_i Y_{ij} \ .$$

Clearly, $f^T Y = \mathbf{1}^T Y^f$ where $\mathbf{1}$ is the all-1 vector. Also, $(Y^f)^f = Y$.

Lemma 2 f is a threshold function of $f_1, \ldots, f_S$, if and only if no positive linear combination of the rows of Y^f equals zero, that is, $\mathbf{q}^T Y^f = 0$ for $\mathbf{q} \geq 0$ implies $\mathbf{q} = 0$.

Proof: It is easy to verify that $f = \text{sgn}(Y\mathbf{w})$ for some $\mathbf{w}$ iff $\mathbf{1} = \text{sgn}(Y^f\mathbf{w})$ or, equivalently, $Y^f\mathbf{w} \geq \mathbf{1}$. Thus, given a function f and a set of input functions represented by the columns of the matrix Y, f can be written as a threshold function of the columns of Y, if and only if there is a feasible solution to the following Linear Program (LP):

$$\begin{aligned} &\text{Minimize } \ 0, \ \text{ subject to} \\ &Y^f\mathbf{w} \ \geq \ \mathbf{1} \end{aligned} \tag{8}$$

The dual of the above LP (see e.g., [22]) is given by:

$$\begin{aligned} &\text{Maximize } \ \mathbf{1}^T\mathbf{q}, \ \text{ subject to} \\ &\mathbf{q}^T Y^f \ = \ 0, \quad \mathbf{q} \geq 0 \end{aligned} \tag{9}$$

The duality theorem of linear programming states that the primal LP (8) is feasible if and only if its dual (9) has a bounded objective function. Now, (9) has a bounded objective function (= 0) if and only if its only feasible solution is $\mathbf{q} = 0$. This is because, if there is a feasible solution $\mathbf{q_0} \neq 0$, then $\alpha\mathbf{q_0}$ is also a feasible solution $\forall\ \alpha > 0$; hence, the objective function $1^T\mathbf{q}$ becomes unbounded. Thus, (8) has a feasible solution if and only if $\mathbf{q}^T Y^f = 0$ for $\mathbf{q} \geq 0$ implies $\mathbf{q} = 0$. □

Theorem 6 Every Boolean function f of n variables (for n even) can be expressed as a threshold function of $2n$ Boolean functions $f_1,\ f_2, \ldots,\ f_{2n}$, such that (1) $C_{ff_i} = 0,\ \ \text{for all} i,\ 1 \leq i \leq 2n$, $i \neq 4$, and (2) $C_{ff_4} = 2^{-(n-1)}$.

Proof: The proof will be developed by constructing a $2^n \times 2n$ (n even) matrix A with ± 1 entries that satisfies the following properties:

1. Every positive linear combination (*plc*) of a nonempty subset of the rows of A has at least one positive coordinate. Thus it follows from the duality theory of Linear Programming (see the proof of Lemma 2) that $A\mathbf{w} \geq \mathbf{1}$, for some $\mathbf{w}$.

2. Every column, except the 4th one, sums to 0 and the 4th column sums to 2, i.e., $\mathbf{1}^T A = [0\ 0\ 0\ 2\ 0 \cdots 0]$.

For any given function f, set $Y^f = A$, and define $Y = (Y^f)^f = A^f$ (i.e., the Y matrix is derived by negating those rows of A that correspond to the -1 entries of the given function f). A proof of Theorem 6 then follows as a direct consequence of Lemma 2 and from the fact that $f = \text{sgn}(Y\mathbf{w})$ for some $\mathbf{w}$, i.e., the given function is a threshold function of the columns of Y.

The construction of the matrix A is recursive and is based on the following 6 by 4 array (partitioned into two sections for reasons to be clarified later); the first two rows will be referred to as r_a and r_b respectively and the bottom four rows will be referred to as r_1, r_2, r_3, and r_4 respectively:

$$\begin{array}{rrrrl} 1 & 1 & -1 & -1 & \quad r_a \\ -1 & -1 & 1 & 1 & \quad r_b \\ & & & & \\ 1 & 1 & 1 & -1 & \quad r_1 \\ 1 & -1 & -1 & 1 & \quad r_2 \\ -1 & 1 & -1 & 1 & \quad r_3 \\ -1 & -1 & 1 & -1 & \quad r_4 \end{array}$$

Following are the properties of the rows of the array:

P1: Rows r_1, r_2, r_3, r_4 sum to 0.

P2: Rows r_a and r_b sum to 0.

P3: Any positive linear combination *plc* of the rows that includes r_b but not r_a has at least one positive coordinate in the last two columns. This is because r_b possesses two 1's in the coordinates 3 and 4 while all other rows $r_1, \ldots, r_4$ contain one 1 and one -1 in those coordinates.

P4: Let S be any subset of the six rows in the array containing at least one row from among r_1, r_2, r_3, and not containing r_4. Then every positive linear combination (*plc*) of S has at least one positive coordinate. This can be proved by enumeration of all possible cases: Assume that S contains the row r_2 but does not contain the row r_4. Then one can verify that at least coordinates 1 or 4 of any *plc* of the rows of S must be positive. Similarly if S contains row r_3 but not r_4 then, coordinates 2 or 4 of any *plc* will be positive; finally if S contains r_1 but not r_4 then coordinates 1, 2, 3, or 4 of any *plc* will be positive.

$$
\begin{array}{lll}
r_1 & r_a & \\
r_1 & r_b & \\
r_1 & r_a & \\
r_1 & r_b & \\
r_2 & r_a & \\
r_2 & r_b & \\
r_2 & r_a & \\
r_2 & r_b & \\
r_3 & r_a & \\
r_3 & r_b & \\
r_3 & r_a & \\
r_3 & r_b & \\
r_4 & r_1 & \leftarrow \text{The last quarter starts here} \\
r_4 & r_2 & \\
r_4 & r_3 & \\
r_4 & r_4 & \leftarrow \text{Change the } -1 \text{ entry in coordinate 4 to } +1
\end{array}
$$

Figure 1: Structure of the matrix A for $n = 4$ ($N = 16$).

Let $N = 2^n$ where n is even; therefore, N is a power of 4. Create an $N \times 4$ array where: each of the first $N/4$ rows is r_1; each of the second $N/4$ rows is r_2; each of the third $N/4$ rows is r_3; and the last $N/4$ rows equal r_4.

Consider any subset S of the rows: if it does not contain any row from the bottom $N/4$ rows, then it doesn't contain r_4 and by property P4, discussed above, any *plc* of S has at least one positive coordinate.

The construction of A can be continued by adding a second set of 4 columns to A (thus defining the first 8 columns) as follows: the first $3N/4$ rows are alternately r_a and r_b; the last $N/4$ rows are divided into four groups of $N/16$ rows each, where the first $N/16$ rows are r_1, the second $N/16$ rows are r_2, the third $N/16$ rows are r_3, and the last $N/16$ rows are r_4. Fig. 1 shows the construction for $n = 4$ ($N = 16$).

A *plc* of the rows of the array that does not contain any row from the bottom $N/16$ rows will always have at least one positive coordinate. This is because: (1) if the rows are chosen only from the top $3N/4$ rows then it follows from the previous discussion that at least one coordinate in the first four coordinates of the *plc* will have a positive coordinate, (2) if the *plc* has at least one row from among other rows (i.e., from row number $3N/4+1$ to row number $(N-N/16)$) then consider the last four coordinates: it is a *plc* of rows that contains at least one row from among r_1, r_2, and r_3 and does not contain r_4; hence, by property P4, at least one coordinate will be positive.

Continue the procedure by adding $\log_4 N$ $(= n/2)$ layers, each of which is 4 columns wide; thus constructing $2n$ columns. By then, any subset S that does not contain the very bottom row has the following property: any *plc* of the rows has at least one positive coordinate. Now change the -1 in the 4th coordinate of the last row to a $+1$; this completes the construction of A. It follows from property P3 that any *plc* that contains the last row of A must have at least one positive coordinate in the first 4 columns.

As for the column sums, properties $P1$ and $P2$ ensure that they are all 0, except for the 4th column that we modified in the last row, which sums to 2. □

Comments:

1. The matrix A in the above construction has identical rows. If one wants to construct an A with distinct rows, then one can add n extra columns to obtain an $2^n \times 3n$ matrix as follows: add n extra columns, such that each row has a unique identifier in these columns. Since we use all 2^n identifiers, each of these n columns sums to 0.

2. Since any *plc* of the rows of A has at least one positive coordinate, it implies that $A\mathbf{w} \geq \mathbf{1}$ for some $\mathbf{w}$. For the matrix A constructed in the proof of the above theorem, a closed form description of such a $\mathbf{w}$ is not apparent.

3. Another decomposition of any given function is outlined in [26, 30], where a closed form description of the corresponding $\mathbf{w}$ can be given. The decomposition has the following properties: Every Boolean function f of n variables can be expressed as a threshold function of $2n$ Boolean functions: $f_1,\ f_2, \ldots,\ f_{2n}$ such that (1) $C_{ff_i} = 0, \ \ \forall\ 1 \leq i \leq 2n-2$, and (2) $C_{ff_{2n-1}} = C_{ff_{2n}} = 2^{-(n-1)}$.

Acknowledgments

The authors would like to thank the referees for their detailed comments, and for strengthening the result presented in Theorem 6.

This work was supported in part by the Joint Services Program at Stanford University (US Army, US Navy, US Air Force) under Contract DAAL03-88-C-0011, the SDIO/IST, managed by the Army Research Office under Contract DAAL03-90-G-0108, and the Department of the Navy, NASA Headquarters, Center for Aeronautics and Space Information Sciences under Grant NAGW-419-S6.

References

[1] E. Allender. A note on the power of threshold circuits. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, pages 580–584, 1989.

[2] P. W. Beame, S. A. Cook, and H. J. Hoover. Log depth circuits for division and related problems. *SIAM Journal on Computing*, 15(4):994–1003, November 1986.

[3] Y. Brandman, A. Orlitsky, and J. Hennessy. A Spectral Lower Bound Technique for the Size of Decision Trees and Two level AND/OR Circuits. *IEEE Transactions on Computers*, 39(2):282–287, February 1990.

[4] J. Bruck. Harmonic analysis of polynomial threshold functions. *SIAM Journal on Discrete Mathematics*, 3(2):168–177, May 1990.

[5] A. K. Chandra, L. Stockmeyer, and U. Vishkin. Constant depth reducibility. *SIAM Journal on Computing*, 13(2):423–439, May 1984.

[6] C. K. Chow. On The Characterization of Threshold Functions. *Proceedings of the 6th Symposium on Switching Circuit Theory and Logical Design*, pages 34–38, 1961.

[7] T. M. Cover. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers*, EC-14:326–334, June 1965.

[8] M. Goldmann, J. Håstad, and A. Razborov. Majority Gates vs. General Weighted Threshold Gates. In *Proceedings of the 7th Annual Structure in Complexity Theory Conference*, pages 2–13, 1992.

[9] M. Goldmann and M. Karpinski. Simulating threshold circuits by majority circuits. In *Proceedings of the 25th Annual ACM Symposium on Theory of Computing*, pages 551–560, 1993.

[10] S. W. Golomb. *Shift Register Sequences.* Laguna Hills, California: Aegean Park Press, 1982.

[11] A. Hajnal, W. Maass, P. Pudlák, M. Szegedy, and G. Turán. Threshold circuits of bounded depth. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, pages 99–110, 1987. Also appears in *Journal of Computer and System Sciences*, 46(2):129-154, April 1993.

[12] T. Hededus and N. Megiddo. On the geometric separability of Boolean functions. Technical Report RJ 9147, IBM Research Division, December 1992.

[13] M. Krause and S. Waack. Variation ranks of communication matrices and lower bounds for depth two circuits having symmetric gates with unbounded fan-in. In *Proceedings of the 32nd Annual Symposium on Foundations of Computer Science*, pages 777–782, 1991.

[14] R. J. Lechner. Harmonic analysis of switching functions. In A. Mukhopadhyay, editor, *Recent Developments in Switching Theory.* New York, Academic Press, 1971.

[15] P. M. Lewis and C. L. Coates. *Threshold Logic.* New York, Wiley, 1967.

[16] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transforms, and learnability. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, pages 574–579, 1989.

[17] M. Minsky and S. Papert. *Perceptrons.* Cambridge, MA: MIT Press, Expanded edition, 1988.

[18] D. E. Muller. Application of Boolean algebra to switching circuit design and to error detection. *Transactions of the IRE, Professional Group on Electronic Computers*, EC-3(3):6–12, 1954.

[19] S. Muroga. *Threshold Logic and its Applications.* New York, Wiley-Interscience, 1971.

[20] E. I. Nechiporuk. The synthesis of networks from threshold elements. *Problemy Kibernetiki*, 11:49–62, April 1964.

[21] I. Ninomiya. A theory of coordinate representation of switching functions. *Memoirs of the Faculty of Engineering, Nagoya University*, 10:175–190, 1958.

[22] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity.* Englewood Cliffs, NJ: Prentice Hall, 1982.

[23] U. Peled and B. Simeone. Polynomial-Time Algorithms for Regular Set-Covering and Threshold Synthesis. *Discrete Applied Mathematics*, 12(1):57–69, September 1985.

[24] N. Pippenger. The complexity of computations by networks. *IBM Journal of Research and Development*, 31(2):235–243, March 1987.

[25] V. P. Roychowdhury, K.-Y. Siu, and A. Orlitsky, editors. *Theoretical Advances in Neural Computation and Learning.* Boston: Kluwer Academic Publishers, 1994.

[26] V. P. Roychowdhury, K.-Y. Siu, A. Orlitsky, and T. Kailath. A Geometric Approach to Threshold Circuit Complexity. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory (COLT)*, pages 97–111, August 1991.

[27] K.-Y. Siu and J. Bruck. On the Power of Threshold Circuits with Small Weights. *SIAM Journal on Discrete Mathematics*, 4(3):423–435, August 1991.

[28] K.-Y. Siu, J. Bruck, T. Kailath, and T. Hofmeister. Depth Efficient Neural Networks for Division and Related Problems. *IEEE Transactions on Information Theory*, 39(3):946–956, May 1993.

[29] K.-Y. Siu, V. P. Roychowdhury, and T. Kailath. Rational Approximation Techniques for Analysis of Neural Networks. *IEEE Transactions on Information Theory*, 40(2), May 1994.

[30] K.-Y. Siu, V. P. Roychowdhury, and T. Kailath. *Discrete Neural Computation: A Theoretical Foundation.* Englewood Cliffs, NJ: Prentice-Hall, 1994.

[31] A. Yao. On ACC and Threshold Circuits. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, pages 619–627, 1990.