Improved Approximation of Linear Threshold Functions

Ilias Diakonikolas Rocco A. Servedio
Department of Computer Science
Columbia University
New York, NY, U.S.A.
Email: {ilias,rocco}@cs.columbia.edu

***Abstract*—We prove two main results on how arbitrary linear threshold functions $f(x) = \mathrm{sign}(w \cdot x - \theta)$ over the n-dimensional Boolean hypercube can be approximated by simple threshold functions.**

Our first result shows that every n-variable threshold function f is ϵ-close to a threshold function depending only on $\mathrm{Inf}(f)^2 \cdot \mathrm{poly}(1/\epsilon)$ many variables, where $\mathrm{Inf}(f)$ denotes the total influence or average sensitivity of f. This is an exponential sharpening of Friedgut's well-known theorem [Fri98], which states that every Boolean function f is ϵ-close to a function depending only on $2^{O(\mathrm{Inf}(f)/\epsilon)}$ many variables, for the case of threshold functions. We complement this upper bound by showing that $\Omega(\mathrm{Inf}(f)^2 + 1/\epsilon^2)$ many variables are required for ϵ-approximating threshold functions.

Our second result is a proof that every n-variable threshold function is ϵ-close to a threshold function with integer weights at most $\mathrm{poly}(n) \cdot 2^{\tilde{O}(1/\epsilon^{2/3})}$. This is a significant improvement, in the dependence on the error parameter ϵ, on an earlier result of [Ser07] which gave a $\mathrm{poly}(n) \cdot 2^{\tilde{O}(1/\epsilon^2)}$ bound. Our improvement is obtained via a new proof technique that uses strong anti-concentration bounds from probability theory. The new technique also gives a simple and modular proof of the original [Ser07] result, and extends to give low-weight approximators for threshold functions under a range of probability distributions beyond just the uniform distribution.

***Keywords*-Boolean functions; threshold functions; approximation**

I. Introduction

Linear threshold functions (henceforth simply called *threshold functions*) are functions $f : \{-1,1\}^n \to \{-1,1\}$ of the form $f(x) = \mathrm{sign}(w \cdot x - \theta)$ where the weights $w_1, \dots, w_n$ and the threshold θ may be arbitrary real values. Threshold functions are a fundamental type of Boolean function and have played an important role in computer science for decades, see e.g. [Der65], [Mur71], [SRK95]. Recent years have witnessed a flurry of research activity on threshold functions from many perspectives of theoretical computer science, including hardness of learning [FGKP06], [KS08], efficient learning algorithms in various models [Kal07], [OS08], [KKMS08], property testing [MORS09], [GS07], communication complexity and circuit complexity [She07], monotone computation [BW06], derandomization [RS08], [DGJ$^+$09], and more.

Despite their seeming simplicity threshold functions can have surprisingly rich structure, and basic questions about them can be unexpectedly challenging to answer. As one example, a moment's thought shows that every threshold function f can be realized with integer weights $w_1, \dots, w_n$: how large do those integer weights need to be? A fairly straightforward argument gives a bound of $2^{O(n \log n)}$, but while this upper bound was known at least since 1961 [MTT61] and rediscovered several times (e.g. [Hon87], [Rag88]), more than thirty years elapsed before a matching lower bound of $2^{\Omega(n \log n)}$ was finally obtained via a fairly sophisticated construction and proof [Hås94], [AV97].

This paper is about approximating arbitrary threshold functions using "simple" threshold functions, meaning ones that depend on few variables or have small integer weights. We use a natural notion of approximation with respect to the uniform distribution: throughout the paper "h is an ϵ-approximator for f" means that $\mathbf{Pr}[h(x) \neq f(x)] \leq \epsilon$. (All probabilities and expectations over $x \in \{-1,1\}^n$ are taken with respect to the uniform distribution, unless otherwise specified.) We prove two main results about approximating threshold functions, which we motivate and describe below.

A. First main result: optimally approximating threshold functions by juntas.

The *influence* of coordinate i on $f : \{-1,1\}^n \to \{-1,1\}$ is $\mathrm{Inf}_i(f) \stackrel{\mathrm{def}}{=} \mathbf{Pr}[f(x) \neq f(x^{\oplus i})]$, where $x^{\oplus i}$ denotes x with the i-th bit flipped. The *total influence* of f, written $\mathrm{Inf}(f)$, is $\sum_i \mathrm{Inf}_i(f)$; it is a normalized measure of the fraction of edges in the hypercube that are rendered bichromatic by f, and is equal to the "average sensitivity" of f. It is well known (see [FK96] or [BT96] for an explicit proof) that every threshold function has $\mathrm{Inf}(f) \leq \sqrt{n}$, and that the majority function on n variables achieves $\mathrm{Inf}(f) = \Theta(\sqrt{n})$ – and in fact maximizes $\mathrm{Inf}(f)$ over all threshold (or even all unate) functions.

In [Fri98], Friedgut proved the following:

Theorem. *[Fri98] Every Boolean function f is ϵ-approximated by a $2^{O(\mathrm{Inf}(f)/\epsilon)}$-junta, i.e. a function depending only on $2^{O(\mathrm{Inf}(f)/\epsilon)}$ of the n input variables.*

Friedgut's theorem is an important structural result about boolean functions and has been usefully applied in several areas of theoretical computer science, including hardness of approximation [DS05], [CKK+06], [KR08], metric embeddings [KR06], and learning theory [OS07]. In Section II-D we discuss the role of this theorem in a sequence of results on the Fourier representation of Boolean functions.

Friedgut showed that his bound is best possible for general Boolean functions, by giving an explicit family of functions which require $2^{\Omega(\mathrm{Inf}(f)/\epsilon)}$-juntas for any ϵ-approximation. A bound of the form $2^{O(\mathrm{Inf}(f)/\epsilon)}$ is of course nontrivial only if $\mathrm{Inf}(f) \ll \log n$, which is rather small; thus, it is natural to ask whether various restricted classes of functions, such as threshold functions, might admit stronger bounds.

Our first main result is an exponentially stronger version of Friedgut's theorem for threshold functions:

Theorem 1 (First Main Theorem)**.** *Every threshold function f is ϵ-approximated by an $\mathrm{Inf}(f)^2 \cdot \mathrm{poly}(1/\epsilon)$-junta (which is itself a threshold function).*

This bound is essentially optimal; easy examples show that $\Omega(\mathrm{Inf}(f)^2 + 1/\epsilon^2)$ many variables may be required for ϵ-approximation. We conjecture that Theorem 1 extends to degree-d polynomial threshold functions with an exponential dependence on d in the bound, and also conjecture a different extension of Theorem 1 that is inspired by a theorem of Bourgain; see Section II-D.

Techniques. The proof of Friedgut's theorem makes essential use of the Bonami-Gross-Beckner hypercontractive inequality [Bon70], [Gro75], [Bec75]. Our proof of Theorem 1 takes a completely different route and does not use hypercontractivity; instead, the main ingredients are recent Fourier results on threshold functions from [OS08] and a probabilistic construction which is reminiscent of Bruck and Smolensky's randomized construction of polynomial threshold functions [BS92].

In more detail, a key notion in our proof is that of a *regular* threshold function; roughly speaking, this is a threshold function where each of the weights w_i is "small" relative to the 2-norm of the weight vector. Given a regular threshold function $g = \mathrm{sign}(w \cdot x - \theta)$, we use the weights w_i to define a probability distribution over approximators to g (this is done similarly to [BS92]). We show (Lemmas 5 and 6) that a randomly drawn approximator from this distribution has high expected accuracy and does not depend on too many variables (the upper bound is given in terms of the weights w_i and the regularity parameter).

An obvious problem in using this construction to approximate arbitrary threshold functions is that not every threshold function is regular. To get around this, we use a recent result from [OS08] which shows that every threshold function f can be well approximated by a threshold function f' which has two crucial properties: f' is *almost* regular (in the sense that it only has a few "large" weights), and its "small" weights are (appropriately scaled versions of) the influences of the corresponding variables in f. For each restriction ρ that fixes the large-weight variables of f', then, we may use $f'|_\rho$ as the regular threshold function g of the previous paragraph, and we obtain a distribution over approximators to $f'|_\rho$ where the number of relevant variables for each such approximator is at most $\mathrm{Inf}(f)^2 \cdot \mathrm{poly}(1/\epsilon)$. From this, using the probabilistic method, we are able to argue that there is a *single* high-accuracy approximator for f that depends on at most $\mathrm{Inf}(f)^2 \cdot \mathrm{poly}(1/\epsilon)$ variables, as required.

B. Second main result: approximating threshold functions to higher accuracy.

The second main result of this paper is about approximating an arbitrary n-variable threshold function f using a threshold function g with *small integer weights*. Goldberg [Gol06] and Servedio [Ser07] have observed that, because of the $2^{\Omega(n \log n)}$ lower bound [Hås94] on integer weights to exactly represent arbitrary threshold functions, it is not possible in general to construct an ϵ-approximator g with integer weights $\mathrm{poly}(n, 1/\epsilon)$. Servedio [Ser07] gave the first positive result, by showing that for every threshold function f there is an ϵ-approximating threshold function g in which each weight is an integer of magnitude at most $\mathrm{poly}(n) \cdot 2^{\tilde{O}(1/\epsilon^2)}$. This result and the ingredients in its proof have since played an important role in subsequent work on threshold functions, e.g. [OS08], [MORS09], [DGJ+09].

Given the usefulness of [Ser07] and the poor dependence on ϵ in its bound, it is natural to seek a stronger quantitative bound with a better dependence on ϵ; in fact, this was posed as a main open question in [Ser07]. Our second main result makes progress in this direction:

Theorem 2 (Second Main Theorem)**.** *Every n-variable threshold function f is ϵ-approximated by a threshold function $g = \mathrm{sign}(w \cdot x - \theta)$ with $w_1, \ldots, w_n$ all integers of magnitude $n^{3/2} \cdot 2^{\tilde{O}(1/\epsilon^{2/3})}$.*

Another question posed in [Ser07] asked about small integer-weight approximators with respect to other probability distributions beyond just the uniform distribution. As described below, Theorem 2 can be generalized to hold under a range of non-uniform distributions.

Theorem 2 is proved using a new approach which we believe may lead to better bounds for a range of problems considered in [OS08], [MORS09], [DGJ+09] which use the approach from [Ser07]. Roughly speaking, the proof in [Ser07] and the applications in [OS08], [MORS09], [DGJ+09] all rely on the fact that for suitable weight vectors w, the random variable $w \cdot x$ (with x uniform over $\{-1, 1\}^n$) can be approximated by a Gaussian. Such approximation provides a great deal of information about $w \cdot x$, but the drawback is that the Gaussian is only a fairly coarse approximator of $w \cdot x$ even for a weight vector as well-behaved as $w = (1, \ldots, 1)$, and this inevitably seems to lead to bounds that are exponential in $1/\epsilon^2$ (as in [Ser07],

[OS08], [DGJ+09]). We now briefly describe how our new approach that yields Theorem 2 gets around this barrier.

Techniques. The main conceptual difference between our new approach and the approach in [Ser07] is this. The proof in [Ser07] starts with an *arbitrary* vector of weights that represent some threshold function; intuitively this could be problematic because these weights may provide an inconvenient representation to work with for the underlying function. In contrast, we focus on the *function itself*, and prove that every threshold function has a "nice" weight vector that represents it. This allows us to exploit strong anti-concentration bounds [Hal77] that apply only under certain assumptions on the weights; we elaborate below.

The notion of *anti-concentration* is an important ingredient in our approach: a random variable has good anti-concentration if it does not assign too much mass to any small interval of the real line. The study of anti-concentration has a rich history in probability theory, see e.g. [DL36], [Kol60], [Ess68], [Rog73], [RV08]. Anti-concentration inequalities for discrete random variables of the form $w \cdot x$ are known to be significantly more delicate than concentration inequalities (i.e. "tail bounds"): while concentration typically depends on the 2-norm of w, anti-concentration depends on the *additive structure* of the coefficients in a subtle way.[1]

We remark that [Ser07] also (implicitly) uses anti-concentration bounds, in particular ones based on Gaussian approximation (that follow from the Berry-Esséen Theorem). In hindsight it can be seen that no stronger anti-concentration bounds can be used in the arguments of [Ser07] because that proof considers *all possible* representations of the form $\mathrm{sign}(w \cdot x - \theta)$, where w ranges over all of $\mathbb{R}^n$. As an example, consider the majority function. For the standard representation as $\mathrm{sign}(\sum_i x_i)$, the anti-concentration bound given by the Berry-Esséen Theorem is the best possible, since an arbitrarily small interval that contains the origin has probability mass $\Omega(1/\sqrt{n})$. On the other hand, it is possible to come up with alternate representations $\mathrm{sign}(w \cdot x)$ for the majority function that have better anti-concentration; this is essentially what our proof does. We prove a structural theorem which states that every threshold function has a representation in which "many" weights are "well-separated;" under this condition on the weights, we obtain strong anti-concentration using a result of of Halász [Hal77]. Finally, we show that strong anti-concentration yields low-weight integer approximation to get our final desired result.

Discussion: Our general approach is both modular and robust. It yields a simple and modular proof of the $\mathrm{poly}(n) \cdot 2^{\tilde{O}(1/\epsilon^2)}$ upper bound from [Ser07] which was proved there via a rather elaborate case analysis. More importantly, the new $\mathrm{poly}(n) \cdot 2^{\tilde{O}(1/\epsilon^{2/3})}$ bound and its proof generalize easily to a wide range of distributions. These include constant-biased product distributions and, using the recent result of [DGJ+09], all K-wise independent distributions for sufficiently large K ($K = \tilde{O}(1/\epsilon^2)$ suffices for ϵ-approximation).

Organization. We prove Theorem 1 in Section II and Theorem 2 in Section III. Because of space constraints our results on non-uniform distributions, as well as some proofs, are deferred to the full version [DS09].

II. Theorem 1: Optimally approximating threshold functions by juntas

This section is structured as follows: in Section II-A we describe a randomized construction of approximators for regular threshold functions. In Section II-B we recall the result from [OS08] that lets us approximate any threshold function by a threshold function that is "almost" regular. In Section II-C we put these pieces together to prove Theorem 1. We give some discussion and conjectures in Sections II-D.

Technical Preliminaries. We assume familiarity with the basic elements of Fourier analysis over $\{-1,1\}^n$. A function $f : \{-1,1\}^n \to \{-1,1\}$ is said to be a "junta on $\mathcal{J} \subseteq [n]$" if f only depends on the coordinates in $\mathcal{J}$. As stated earlier, we say that f is a J-junta, $0 \le J \le n$, if it is a junta on some set of cardinality at most J. For a vector $u \in \mathbb{R}^m$ we write $\|u\|_1$ to denote the L_1 norm of u, i.e. $\|u\|_1 = \sum_{i=1}^m |u_i|$. We write "$X \leftarrow \mathcal{D}$" to indicate that random variable X is distributed according to distribution $\mathcal{D}$.

Finally, we give a precise definition of the notion of a "regular" threshold function:

Definition 3. *Let $f(x) = \mathrm{sign}(w_0 + \sum_{i=1}^n w_i x_i)$ be a threshold function where $\sum_{i=1}^n w_i^2 = 1$. We say that f is τ-regular if $|w_i| \le \tau$ for all $i \in [n]$.*[2]

A. Randomly constructing approximators to regular threshold functions.

Fix $h_\theta(x) = \mathrm{sign}(\theta + \sum_{i=1}^m u_i x_i)$ to be a τ-regular threshold function, so $\sum_{i=1}^m u_i^2 = 1$ and $|u_i| \le \tau$ for all $i \in [m]$. Our notation emphasizes the threshold parameter θ since it will play an important role later.

We begin by defining a distribution $\mathcal{D}$ over linear forms $L(x) = \sum_{i=1}^m c_i x_i$. The distribution $\mathcal{D}$ is defined using the weights u_i similarly to how Bruck and Smolensky [BS92] define a distribution over polynomials using the Fourier coefficients of a Boolean function. A draw of $L(x)$ from $\mathcal{D}$ is obtained as follows: $L(x)$ is first initialized to 0. Then

[1] Roughly speaking, if one forbids more and more additive structure in the w_i's, then one gets better and better anti-concentration; see e.g. [Vu08], [TV08] and Chapter 7 of [TV06].

[2] Strictly speaking, τ-regularity is a property of a particular representation $\mathrm{sign}(w_0 + \sum_{i=1}^n w_i x_i)$ and not of a threshold function f, which could have different representations some of which are τ-regular and some of which are not. The particular representation we are concerned with will always be clear from context. A similar remark holds for Definition 4.

the following is independently repeated $N \stackrel{\text{def}}{=} \Theta(\|u\|_1^2 \cdot \frac{1}{\tau^2} \cdot \ln(1/\tau))$ times: an index $i \in [m]$ is selected with probability $\frac{|u_i|}{\|u\|_1}$, and $\mathrm{sign}(u_i)x_i$ is added to $L(x)$.

Fix any $z \in \{-1,1\}^m$. For $L \leftarrow \mathcal{D}$, we may view $L(z)$ as a sum of N i.i.d. ± 1-valued random variables $Z_1(z), \ldots, Z_N(z)$, where the expectation of each $Z_j(z)$ is $\sum_{i=1}^m \frac{|u_i|}{\|u\|_1}\mathrm{sign}(u_i)z_i = \frac{1}{\|u\|_1}u \cdot z$. We thus have:

$$\mathbf{E}_{L\leftarrow\mathcal{D}}[L(z)] = \sum_{j=1}^N \mathbf{E}[Z_j(z)] = \frac{N}{\|u\|_1}(u \cdot z). \qquad (1)$$

With $\mathcal{D}$ in hand we define a distribution $\mathcal{D}'$ over threshold functions g_θ in the following natural way: to draw a function $g_\theta \leftarrow \mathcal{D}'$ we draw $L \leftarrow \mathcal{D}$ and set

$$g_\theta(x) = \mathrm{sign}(\theta + \frac{\|u\|_1}{N}L(x)). \qquad (2)$$

We would like to show that for $g_\theta \leftarrow \mathcal{D}'$, the probability that $g_\theta(z)$ disagrees with $h_\theta(z)$ is "small," i.e. at most $O(\tau)$. But such a bound cannot hold for every $z \in \{-1,1\}^m$, for if the value of $\theta + u \cdot z$ is arbitrarily close to 0 then the expected value of the argument to sign in (2) may be arbitrarily close to 0. For z such that $\theta + u \cdot z$ is not too close to 0, though, it is possible to argue that $g_\theta(z)$ is incorrect only with small probability (over the draw of $g_\theta \leftarrow \mathcal{D}'$). Moreover, the regularity of h_θ lets us argue that only a small fraction of inputs z have $\theta + u \cdot z$ close to 0, so we can conclude that the expected error of g_θ is low. We now provide the details.

We will use the following notion of the "margin" of an input relative to a threshold function:

Definition 4. *Let $f(x) = \mathrm{sign}(w_0 + \sum_{i=1}^n w_i x_i)$ be a threshold function where the weights are scaled so that $\sum_{i=1}^n w_i^2 = 1$. Given a particular input $z \in \{-1,1\}^n$ we define $\mathrm{marg}(f,z) \stackrel{\text{def}}{=} |w_0 + \sum_{i=1}^n w_i z_i|$.*

Let $\mathrm{MARG}_{\theta,\tau} \stackrel{\text{def}}{=} \{z \in \{-1,1\}^m : \mathrm{marg}(h_\theta, z) \geq \tau\}$ denote the set of points in $\{-1,1\}^m$ with margin at least τ under h_θ. We now show that a random $g_\theta \leftarrow \mathcal{D}'$ has high expected accuracy on each point $z \in \mathrm{MARG}_{\theta,\tau}$:

Lemma 5. *For each $z \in \mathrm{MARG}_{\theta,\tau}$ we have $\mathbf{Pr}_{g_\theta \leftarrow \mathcal{D}'}[h_\theta(z) \neq g_\theta(z)] \leq \tau$. Moreover, each $g_\theta \leftarrow \mathcal{D}'$ is an N-junta.*

Proof: The latter claim is immediate so it suffices to prove the former. Fix any $z \in \mathrm{MARG}_{\theta,\tau}$, so $|\theta + u \cdot z| \geq \tau$. We need to bound from above the probability of the "bad" event (over the random choice of $g_\theta \leftarrow \mathcal{D}'$) that $h_\theta(z) \neq g_\theta(z)$; we refer to this bad event as B.

The key claim is that if B occurs then it must be the case that $|L(z) - \mathbf{E}_{L\leftarrow\mathcal{D}}[L(z)]| \geq \frac{N\tau}{\|u\|_1}$. For suppose that $h_\theta(z) = 1$ and $g_\theta(z) = -1$ (the other case is handled similarly). By definition, we have that $\theta + u \cdot z \geq 0$ and $\theta + (\|u\|_1/N)L(z) < 0$. Since z belongs to $\mathrm{MARG}_{\theta,\tau}$, the first inequality gives that $\theta + u \cdot z \geq \tau$, which implies, via (1), that $\mathbf{E}[L(z)] \geq (N/\|u\|_1)(\tau - \theta)$. The second inequality is equivalent to $L(z) < -\theta N/\|u\|_1$, and consequently we have $\mathbf{E}[L(z)] - L(z) \geq N\tau/\|u\|_1$.

We thus have that $\mathbf{Pr}_{g_\theta\leftarrow\mathcal{D}'}[h_\theta(z) \neq g_\theta(z)] \leq \mathbf{Pr}_{L\leftarrow\mathcal{D}}[|L(z) - \mathbf{E}[L(z)]| \geq \frac{N\tau}{\|u\|_1}]$. Now we again view $L(z)$ as the sum of N i.i.d. $\{-1,1\}$ random variables. The Hoeffding bound yields that $\mathbf{Pr}_{L\leftarrow\mathcal{D}}[|L(z) - \mathbf{E}[L(z)]| \geq \frac{N\tau}{\|u\|_1}]$ is at most

$$2\exp\left(-2\frac{(N\tau/\|u\|_1)^2}{4N}\right) \leq \tau,$$

where the inequality follows by our choice of N. This completes the proof of the lemma. ■

We next note that by the regularity of h_θ, most points in $\{-1,1\}^m$ have a large margin (and hence are covered by Lemma 5):

Lemma 6. $\mathbf{Pr}_{x\in\{-1,1\}^m}[x \notin \mathrm{MARG}_{\theta,\tau}] \leq 4\tau$.

The proof is a direct consequence of regularity via the Berry-Esséen theorem (see [DS09]).

Combining Lemmas 5 and 6, we get the main result of this subsection:

Lemma 7. $\mathbf{E}_{g_\theta\leftarrow\mathcal{D}'}[\mathbf{Pr}_{x\in\{-1,1\}^m}[g_\theta(x) \neq h_\theta(x)]] \leq 5\tau$.

B. Approximating threshold functions using their influences as (almost all of) the weights.

Our next tool is the following theorem on approximating threshold functions. Roughly, it says that every threshold function f can be well approximated by a threshold function f' where all but the $\mathrm{poly}(1/\epsilon)$ largest weights of f' have a special structure: up to sign, they are the values $\mathrm{Inf}_i(f)$. (Recall that for a threshold function f we have $|\widehat{f}(i)| = \mathrm{Inf}_i(f)$.)

Theorem 8. *[Theorem 17 of [OS08]] There is a fixed polynomial $\kappa(\epsilon) = \Theta(\epsilon^{144})$[3] such that the following holds: Let $f(x) = \mathrm{sign}\left(w_0 + \sum_{i\in H} w_i x_i + \sum_{i\in T} w_i x_i\right)$ be a threshold function over head indices H and tail indices T, where $H \stackrel{\text{def}}{=} \{i : |\widehat{f}(i)| \geq \kappa(\epsilon)^2\}$ and T satisfies $\sum_{i\in T} w_i^2 = 1$. Then either:*
(i) f is $O(\epsilon)$-close to a junta over H; or,
(ii) f is $O(\epsilon)$-close to the threshold function $f'(x) = \mathrm{sign}\left(w_0 + \sum_{i\in H} w_i x_i + \sum_{i\in T} \frac{\widehat{f}(i)}{\sigma_T} x_i\right)$, where σ_T denotes $\sqrt{\sum_{i\in T}\widehat{f}(i)^2}$. Moreover, in this case we have $\sigma_T = \Omega(\epsilon^2)$.[4]

Note that $\sum_{i\in T}(\widehat{f}(i)/\sigma_T)^2 = 1$, and since $\sigma_T = \Omega(\epsilon^2)$, for each $i \in T$ we have

$$|\widehat{f}(i)/\sigma_T| < \kappa(\epsilon)^2/\sigma_T \leq O(\epsilon^{288})/\Omega(\epsilon^2) = O(\epsilon^{286}). \qquad (3)$$

[3]See the discussion immediately before Equation (24) of [OS08]; our $\kappa(\epsilon)$ is the $\tau(\epsilon)$ of [OS08].
[4]See Equation (24) of [OS08].

This means that for any restriction ρ fixing the variables in H, the function $f'|_\rho$ is $\mathrm{poly}(\epsilon)$-regular; this is important since it will allow us to apply the results of Section II-A to these restrictions.

C. Proof of Theorem 1.

Now we are ready to prove Theorem 1. We first show that every threshold function f is $O(\epsilon)$-approximated by a $(1+\mathrm{Inf}(f)^2)\cdot\mathrm{poly}(1/\epsilon)$-junta threshold function, and then argue that this yields Theorem 1. For brevity, in the rest of this subsection we write $\mathbb{I}$ for $\mathrm{Inf}(f)$.

Let $0<\epsilon<\frac{1}{2}$ be given and let f be any n-variable threshold function. W.l.o.g. we may consider a representation $f(x)=\mathrm{sign}(w_0+\sum_{i=1}^n w_i x_i)$ in which each $w_i\neq 0$, and by scaling the weights we may further assume that $T=[n]\setminus H$ has $\sum_{i\in T} w_i^2=1$.

We apply Theorem 8 to f. Parseval's identity implies that at most $1/\kappa(\epsilon)^4$ many indices i can have $|\hat{f}(i)|\geq\kappa(\epsilon)^2$, so we have $|H|\leq 1/\kappa(\epsilon)^4=\mathrm{poly}(1/\epsilon)$. In Case (i) we immediately have that f is $O(\epsilon)$-close to a $\mathrm{poly}(1/\epsilon)$-junta, so we suppose that Case (ii) holds, and henceforth argue about the $O(\epsilon)$-approximator f' defined in Case (ii).

We consider all $2^{\mathrm{poly}(1/\epsilon)}$ restrictions ρ obtained by fixing the head variables in H. Our goal is to apply the results of Section II-A to the functions $f'|_\rho$. As noted in Section II-B, for each restriction ρ the resulting function $f'|_\rho$ over the tail variables in T is a $\tau(\epsilon)$-regular threshold function, where $\tau(\epsilon)=O(\epsilon^{286})$ is the function implicit in the RHS of (3) (for brevity we henceforth write τ for $\tau(\epsilon)$). Moreover, all these restrictions are threshold functions defined by the same linear form over the variables in T: they only differ in their threshold values, i.e. the values $\theta_\rho\stackrel{\text{def}}{=}w_0+\sum_{i\in H} w_i\rho_i$.

In keeping with the notation of Section II-A, for each restriction ρ we write h_{θ_ρ} to denote $f'|_\rho$, i.e. $h_{\theta_\rho}(x_T)\stackrel{\text{def}}{=}\mathrm{sign}(\theta_\rho+\sum_{i\in T}u_i x_i)$ where $u_i\stackrel{\text{def}}{=}\frac{\widehat{f}(i)}{\sigma_T}$ and $x_T\stackrel{\text{def}}{=}(x_i)_{i\in T}$. We observe that $\|u\|_1=\sum_{i\in T}|u_i|$ equals

$$\frac{1}{\sigma_T}\sum_{i\in T}|\widehat{f}(i)|\leq\frac{1}{\sigma_T}\sum_{i=1}^{n}|\widehat{f}(i)|\leq\mathbb{I}\cdot\mathrm{poly}(1/\epsilon). \qquad (4)$$

where the last inequality uses $\mathrm{Inf}_i(f)=|\widehat{f}(i)|$ and $\sigma_T=\Omega(\epsilon^2)$. Recalling that N equals $\Theta(\|u\|_1^2\cdot\frac{1}{\tau^2}\cdot\ln(1/\tau))$, we have that N is at most $\mathbb{I}^2\cdot\mathrm{poly}(1/\epsilon)$.

We consider a distribution $\mathcal{D}''$ over threshold functions on $\{-1,1\}^n$ defined as follows: a draw of $g\leftarrow\mathcal{D}''$ is obtained by drawing $L\leftarrow\mathcal{D}$ and setting $g(x)=\mathrm{sign}(w_0+\sum_{i\in H}w_i x_i+\frac{\|u\|_1}{N}\cdot L(x_T))$. For every outcome of $g\leftarrow\mathcal{D}''$, the function g depends on at most $|H|+N=(1+\mathbb{I}^2)\mathrm{poly}(1/\epsilon)$ many variables.

It remains only to argue that some g drawn from $\mathcal{D}''$ is $O(\epsilon)$-close to f'. Via the probabilistic method, to do this it suffices to show that $\mathbf{E}_{g\leftarrow\mathcal{D}''}[\mathbf{Pr}_{x\in\{-1,1\}^n}[g(x)\neq f'(x)]]=O(\tau)$ (recall that $\tau\ll\epsilon$). We now do this using the results of Section II-A.

Fix any assignment ρ to the variables in H. By Lemma 7 we have

$$\mathbf{E}_{g\leftarrow\mathcal{D}''}\left[\mathbf{Pr}_{x_T\leftarrow\{-1,1\}^{|T|}}[f'|_\rho(x_T)\neq g|_\rho(x_T)]\right]\leq 5\tau.$$

Averaging over all ρ, we get

$$\mathbf{E}_{g\leftarrow\mathcal{D}''}\left[\mathbf{Pr}_{x\leftarrow\{-1,1\}^n}[f'(x)\neq g(x)]\right]\leq 5\tau$$

which is the desired bound.

So, we have shown that every threshold function f is $O(\epsilon)$-close to a $(1+\mathbb{I}^2)\cdot\mathrm{poly}(1/\epsilon)$-junta; we finish the proof of Theorem 1 by arguing that this implies a $\mathbb{I}^2\cdot\mathrm{poly}(1/\epsilon)$ junta size bound. Let c be an absolute constant such that every f is ϵ-close to a $(1+\mathbb{I}^2)\cdot(1/\epsilon)^c$-junta; we consider different cases based on the size of $\mathbb{I}$. If $\mathbb{I}>1$, then it is clear that $(1+\mathbb{I}^2)(1/\epsilon)^c<2\mathbb{I}^2(1/\epsilon)^c<\mathbb{I}^2(1/\epsilon)^{c+1}$ (using $\epsilon<1/2$). If $\mathbb{I}<\epsilon^2$, since $\sum_{|S|\geq 1}\widehat{f}(S)^2\leq\sum_{|S|\geq 1}|S|\widehat{f}(S)^2=\mathbb{I}$ (the equality is a well-known fact in Fourier analysis of Boolean functions, see e.g. [KKL88]), by Parseval's identity we get that $|\widehat{f}(\emptyset)|\geq 1-\epsilon$. This means that f is ϵ-close to a constant function, which is of course a 0-junta. Finally, if $\epsilon^2\leq\mathbb{I}\leq 1$, then $1+\mathbb{I}^2\leq 2\leq 2\mathbb{I}^2\epsilon^{-4}\leq\mathbb{I}^2\epsilon^{-5}$, so f can be ϵ-approximated by a $\mathbb{I}^2(1/\epsilon)^{c+5}$-junta. So in every case f is ϵ-close to an $\mathrm{Inf}(f)^2\cdot(1/\epsilon)^{c+5}$-junta, and Theorem 1 is proved. □

D. Discussion and Conjectures.

Improved low-weight approximators of threshold functions. Recall the main result of [Ser07]:

Theorem 9. *[Ser07] Every n-variable threshold function f is ϵ-approximated by a threshold function $g=\mathrm{sign}(w\cdot x-\theta)$ with $w_1,\dots,w_n$ all integers satisfying $\sum_{i=1}^n w_i^2\leq n\cdot 2^{\tilde{O}(1/\epsilon^2)}$.*

While a linear dependence on n is the best possible bound which can hold uniformly for all n-variable threshold functions, it is possible to give a sharper bound that depends on f. Applying Theorem 9 to the threshold function junta which is given by Theorem 1, we obtain:

Corollary 10. *Every n-variable threshold function f is ϵ-approximated by a threshold function $g=\mathrm{sign}(w\cdot x-\theta)$ with $w_1,\dots,w_n$ all integers satisfying $\sum_{i=1}^n w_i^2\leq\mathrm{Inf}(f)^2\cdot 2^{\tilde{O}(1/\epsilon^2)}$.*

Since $\mathrm{Inf}(f)^2$ is at most n (but can be much less) for every threshold function f, this strengthens Theorem 9.

A lower bound. We observe that the $\mathrm{Inf}(f)^2\cdot\mathrm{poly}(1/\epsilon)$ upper bound of Theorem 1 is nearly best possible: no strengthening can replace this with a bound smaller than $\Omega(\mathrm{Inf}(f)^2+1/\epsilon^2)$ (see [DS09]).

Extending to degree-d? It is natural to wonder whether Theorem 1 extends to *polynomial* threshold functions (PTFs) of degree d, i.e. Boolean functions $f(x)=\mathrm{sign}(p(x))$ where

p is a degree-d polynomial. We pose the following conjecture which is a broad generalization of Theorem 1:

Conjecture 1. *Every degree-d PTF f is ϵ-approximated by a $(\mathrm{Inf}(f)/\epsilon)^{O(d)}$-junta.*

We suspect that even the $d = 2$ case of Conjecture 1 may be challenging, as the total influence of low-degree polynomial threshold functions does not seem to be well understood.

An exponential sharpening of Bourgain's theorem? By Parseval's identity, every Boolean function f has $\sum_{S\subseteq[n]} \hat{f}(S)^2 = 1$. Since the total influence $\mathrm{Inf}(f)$ equals $\sum_S \hat{f}(S)^2|S|$ and the degree of each monomial x_S is $|S|$, we may interpret $\mathrm{Inf}(f)$ as the "average" Fourier degree of f.

With this point of view, Friedgut's theorem may be viewed as part of a sequence of three results, all of which essentially say that Boolean functions with low degree (in some sense) are close (in some sense) to juntas. The first and earliest of these results is the following theorem of Nisan and Szegedy:

Theorem 11. *[NS94] Every Boolean function with (maximum) Fourier degree k is a $k2^k$-junta.*

This theorem imposes a strong degree condition on f – that it have *zero* Fourier weight above degree k – and gets a strong conclusion, that f is *identical* to a $k2^k$-junta. Next, Friedgut's theorem [Fri98] relaxed both the degree condition on f and the resulting conclusion: if the "average" Fourier degree of f (i.e. $\mathrm{Inf}(f)$) is at most k, then f is ϵ-close to a $2^{O(k/\epsilon)}$-junta. Finally and most recently, Bourgain relaxed the degree condition even further, by showing that if f puts most of its Fourier weight on low-degree monomials, then regardless of where the remaining Fourier weight lies, f must be close to a junta:

Theorem 12. *[Bou02] Every Boolean function f with $\sum_{|S|>k} \hat{f}(S)^2 \leq (\epsilon/k)^{1/2+o(1)}$ is ϵ-close to a $2^{O(k)} \cdot \mathrm{poly}(1/\epsilon)$-junta.*

Let us consider how each junta size bound changes when we restrict our attention to threshold functions in the above theorems. We first observe that the [NS94] bound can be exponentially improved in this case:

Proposition 13. *Every threshold function with (maximum) Fourier degree k is a $(2k-1)$-junta.*

(This follows from the easy fact that any threshold function with r relevant variables contains a subfunction which is an $(\frac{r+1}{2})$-way AND or OR.) Our Theorem 1, of course, tells us that Friedgut's theorem can also be exponentially sharpened if f is a threshold function. This motivates the natural question of whether Bourgain's theorem can be similarly sharpened for threshold functions. We state the following:

Conjecture 2. *Every threshold function f with $\sum_{|S|>k} \hat{f}(S)^2 \leq (\epsilon/k)^{1/2+o(1)}$ is ϵ-close to a $\mathrm{poly}(k/\epsilon)$-junta.*

III. THEOREM 2: APPROXIMATING THRESHOLD FUNCTIONS TO HIGHER ACCURACY.

As outlined in Section I-B, our new approach can be conceptually broken into the following steps:

1) Show that every threshold function has a representation in which many weights are "nice".
2) Use the "niceness" of the weights to establish anti-concentration of $w \cdot x$.
3) Finally, use the anti-concentration of $w \cdot x$ to obtain an approximator with small integer weights.

Note that there is a delicate relationship between the first two steps: the structural result for the weights that is established in the first step must match the necessary conditions for anti-concentration in the second step. The third step is a simple generic lemma translating anti-concentration into low-weight approximation.

The structure of this section is as follows: We first recall the anti-concentration results that we need to implement Step 2 in our above proof template. We then prove the simple lemma that implements Step 3 in our proof template. In Section III-A we give a "warmup" to our main result by using the template to give a clean and modular proof of the main result of [Ser07]. In Section III-B we show how the template yields a variant of Theorem 2 which has an $n^{O(1/\epsilon^{2/3})}$ bound. This subsection includes our main new technical contribution of Section III, a new result on representations of threshold functions, Lemma 22. Roughly speaking, this lemma says that every threshold function has a representation such that many of the differences between consecutive weights are not too small. Finally, in Section III-C we show how this $n^{O(1/\epsilon^{2/3})}$ bound can be improved to fully prove Theorem 2.

All the results of this section can be appropriately generalized to constant-biased product distributions and K-wise independent distributions (but they provably *cannot* be generalized to *every* distribution). Because of space constraints, we give these results in [DS09].

Anti-concentration of weighted sums of Bernoulli random variables. We start with the formal definition of anti-concentration:

Definition 14. *Let $a \in \mathbb{R}^n$ be a weight–vector and $r \in \mathbb{R}_+$. The* Lévy anti-concentration function of *a is defined as*

$$p_r(a) \stackrel{def}{=} \sup_{v\in\mathbb{R}} \mathbf{Pr}_{x\leftarrow\mathcal{U}}[|a \cdot x - v| \leq r].$$

Thus, the anti-concentration of a weight vector a is an upper bound on the probability that $a \cdot x$ lies in any small interval (of length $2r$). An early and important result on anti-concentration was given by Erdős [Erd45]; improving on an earlier result of Littlewood and Offord [LO43], he proved

Theorem 15 (Erdős). *Let $a = (a_1, \dots, a_k) \in \mathbb{R}^k$, $r \in \mathbb{R}_+$ be such that $|a_i| \geq r$ for all $i \in [k]$. Then $p_r(a) \leq \binom{k}{k/2}/2^k = O(k^{-1/2})$.*

A large body of subsequent work generalized this result in many different ways (see e.g. Chapter 7 of [TV06]); anti-concentration results of this general flavor have come to be known as "Littlewood-Offord theorems." We shall require an extension of Theorem 15 which is due to Halász [Hal77], improving upon Erdős-Moser [Erd65] and Sárközy-Szemerédi [SS65]. While Erdős's theorem gives the best (smallest) possible anti-concentration bound assuming that each *weight* a_i is large, Halász's theorem gives a stronger bound under the stronger assumption that the *difference* between any two weights is large:

Theorem 16 (Halász). *Let $a = (a_1, \dots, a_k) \in \mathbb{R}^k$, $r \in \mathbb{R}_+$ be such that $|a_i - a_j| \geq r$ for all $i \neq j \in [k]$. Then $p_r(a) \leq O(k^{-3/2})$.*

Looking ahead, we note that the "3/2" exponent instead of "1/2" is the key to our improvement from $2^{\tilde{O}(1/\epsilon^2)}$ to $2^{\tilde{O}(1/\epsilon^{2/3})}$.

The last fact about anti-concentration that we shall need is the following simple lemma, which says that if we extend a weight vector a by adding more weights, its anti-concentration can only improve:

Lemma 17 (Extension). *Let $a \in \mathbb{R}^k$ be any k-dimensional weight vector and $r \in \mathbb{R}_+$ be any non-negative real. For any $n > k$, let $a' \in \mathbb{R}^n$ be the vector $(a_1, \dots, a_k, a'_{k+1}, \dots, a'_n)$ where the weights $a'_{k+1}, \dots, a'_n$ may be any real numbers. Then we have $p_r(a') \leq p_r(a)$.*

The proof is by a simple averaging argument, using the fact that for $x \leftarrow \{-1,1\}^n$ uniform random, conditioned on any outcome of the variables $x_{k+1}, \dots, x_n$, the distribution of $x_1, \dots, x_k$ is still uniform.

From anti-concentration to a low-weight approximator. The following simple lemma takes us from anti-concentration to a low-weight approximator. We use it to implement Step 3 in our proof template.

Lemma 18. *Let $g = \mathrm{sign}(\sum_{i=1}^n w_i x_i - \theta)$ be any threshold function. If $p_r(w_1, \dots, w_n) \leq \epsilon$, then there exists a 2ϵ-approximator h for g, where h is a threshold function with integer weights each of magnitude $O(\max_i |w_i| \cdot \sqrt{n \ln(1/\epsilon)}/r)$.*

Proof: Let $\alpha = r/(\sqrt{n \ln(2/\epsilon)})$. For each $i \in [n]$, let u_i be the value obtained by rounding w_i to the nearest integer multiple of α and $v_i = u_i/\alpha \in \mathbb{Z}$. We claim that $h(x) = \mathrm{sign}(\sum_{i=1}^n v_i x_i - \theta/\alpha)$ is the desired approximator. It is clear that $\max_i |v_i| = O(\max_i |w_i|/\alpha)$, so it suffices to show that h is $(\epsilon + \epsilon)$-close to g.

For $i \in [n]$, let $e_i = w_i - u_i$, so that $u \cdot x = w \cdot x - e \cdot x$. We have that $g(x) \neq h(x)$ only if $|e \cdot x| \geq r$ or $|w \cdot x - \theta| \leq r$. We bound from above the probability of each of these events by ϵ. The probability of the second event is bounded by ϵ since $\mathbf{Pr}[|w \cdot x - \theta| \leq r] \leq p_r(w) \leq \epsilon$. For the first event we have $\mathbf{Pr}[|e \cdot x| \geq r] \leq \mathbf{Pr}[|e \cdot x| \geq \|e\|_2 \sqrt{2\ln(2/\epsilon)}] \leq \epsilon$, where the first inequality uses the fact $\|e\|_2 \leq (r/\sqrt{2\ln(2/\epsilon)})$ and the second follows from the Hoeffding bound. ■

A. Warmup: Simple Proof of [Ser07] Main Result.

In this section we give a simple and modular proof of nearly the same bound as the main result of [Ser07], following the proof template from the start of Section III. Let $f : \{-1,1\}^n \to \{-1,1\}$ be any threshold function.

First step: This is provided for us by the following result, which is an immediate consequence of Lemma 14 in [OS08]. Intuitively, this result says that every threshold function has a representation in which the k-th largest weight is not too small compared with the largest weight.[5]

Claim 19. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any threshold function, let $\epsilon > 0$, and let $k \in [n]$. There is an ϵ-approximator $g = \mathrm{sign}(\sum_{i=1}^n w_i x_i - \theta)$ for f with the following property: Suppose (reordering and rescaling weights if necessary) that $1 = |w_1| \geq \cdots \geq |w_n|$. Then $|w_k| \geq 1/(k^k \sqrt{3n \ln(2/\epsilon)})$.*

Second step: We apply Erdős's theorem, Theorem 15, to the weight vector $(w_1, \dots, w_k)$ from Claim 19 (we will fix k later), taking $r = 1/(k^k \sqrt{3n \ln(2/\epsilon)})$ to be the bound from Claim 19. Theorem 15 gives $p_r(w_1, \dots, w_k) \leq O(1/\sqrt{k})$, and the Extension Lemma 17 gives that in fact $p_r(w_1, \dots, w_n) \leq O(1/\sqrt{k})$.

Third step: It remains only to fix $k = \min\{1/\epsilon^2, n\}$ and observe that the h obtained from Lemma 18 is an $O(\epsilon)$-approximator for f. (Note that if $1/\epsilon^2 > n$, then integer weights $2^{\tilde{O}(1/\epsilon^2)}$ suffice to *exactly* represent f by [MTT61].) We have thus proved:

Theorem 20. *Every n-variable threshold function f is ϵ-approximated by a threshold function $h = \mathrm{sign}(v \cdot x - \theta)$ with $v_1, \dots, v_n$ all integers of magnitude $n \cdot 2^{\tilde{O}(1/\epsilon^2)}$.*

This is almost identical to the main result of [Ser07]; the bound of [Ser07] has $\sqrt{n}$ in place of n.

B. Toward Theorem 2: An $n^{O(1/\epsilon^{2/3})}$ bound.

In this section we prove an intermediate result towards our ultimate goal of $\mathrm{poly}(n) \cdot 2^{\tilde{O}(1/\epsilon^{2/3})}$:

[5] We do not repeat the proof of Claim 19 from [OS08] here but we note that the proof is self-contained and rather straightforward; it follows along the lines of [MTT61]'s classic argument to upper bound the weights required to represent any threshold function.

Theorem 21. *Every n-variable threshold function f is ϵ-approximated by a threshold function $h = \mathrm{sign}(v \cdot x - \theta)$ with $v_1, \ldots, v_n$ all integers of magnitude $n^{O(1/\epsilon^{2/3})}$.*

We follow the same high-level proof template as the previous section. Let $f : \{-1,1\}^n \to \{-1,1\}$ be a threshold function. We may assume w.l.o.g. that f depends on all n input variables, and since the claimed bound follows again from [MTT61] if $1/\epsilon^{2/3} > n-2$, we assume $1/\epsilon^{2/3} \leq n-2$.

First step: Our goal now is to apply Halász's anti-concentration bound in Step 2 rather than Erdős's theorem. To do this we need the following new result on representing threshold functions, which intuitively says that every threshold function has a representation using weights such that many of the differences between consecutive weights are not too small compared to the largest weight:

Lemma 22. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a threshold function that depends on all n variables. There is a representation $\mathrm{sign}(\sum_{i=1}^n w_i x_i - \theta)$ for f with the following property: Suppose (reordering and rescaling weights if necessary) that $1 = |w_1| \geq \cdots \geq |w_n| > 0$. For $i \in [n-1]$ let $\Delta_i \overset{\text{def}}{=} |w_i| - |w_{i+1}|$. Then for any $k \in [n-2]$, the k-th biggest element of the (multiset) $\Delta_1, \ldots, \Delta_{n-1}$ is at least $\frac{1}{(2n+2)^{2k+8}}$.*

We pause to contrast this result with an earlier theorem due to Håstad [Hås05] that appeared in [Ser07]. Under the same hypotheses as Lemma 22, the earlier theorem asserted that for any $k \in [n]$ the k-th largest weight w_k satisfies $|w_k| \geq \frac{1}{k!(n+1)}$. The proof of the earlier theorem centers on a careful analysis of a linear program in which the variables are the weights $w_1, \ldots, w_n$ and there are 2^n constraints corresponding to the 2^n points $x \in \{-1,1\}^n$. To prove Lemma 22, we must now analyze a linear program with some additional constraints which, intuitively, ensure that there are "gaps" between the weights.[6] We prove Lemma 22 in Section III-D.

Second step: We take $k = 1/\epsilon^{2/3}$ and consider the k largest differences $\Delta_{i_1} = |w_{i_1}| - |w_{i_1+1}|, \ldots, \Delta_{i_k} = |w_{i_k}| - |w_{i_k+1}|$. Lemma 22 implies that for all $a \neq b \in [k]$ we have $|w_{i_a} - w_{i_b}| \geq r$, for $r = 1/(2n+2)^{2k+8}$. Applying Halász's anti-concentration bound, Theorem 16, we get that $p_r(w_{i_1}, \ldots, w_{i_k}) \leq O(k^{-3/2}) = O(\epsilon)$, and the Extension Lemma 17 further gives $p_r(w_1, \ldots, w_n) = O(\epsilon)$.

Third step:. We simply apply Lemma 18. Recalling that $r = 1/(2n+2)^{\Theta(1/\epsilon^{2/3})}$, Theorem 21 is proved. □

[6]In fact, by considering the majority function one can verify that the 2^n-constraint linear program of the earlier proof is not sufficient; that LP yields a representation in which each w_i is the same and hence the "gaps" Δ_i are all 0.

C. Proof of Theorem 2: A $\mathrm{poly}(n) \cdot 2^{\tilde{O}(1/\epsilon^{2/3})}$ bound.

Given a threshold function $f(x) = \mathrm{sign}(w \cdot x - \theta)$ such that $|w_1| \geq \cdots \geq |w_n| > 0$, for $k \in [n]$ we denote by σ_k the quantity $\sqrt{\sum_{i=k}^n w_i^2}$. The analysis in [Ser07] is based on the notion of the "τ-critical index":

Definition 23. *We define the* τ-critical index $\ell(\tau)$ *of a threshold function $f = \mathrm{sign}(w \cdot x - \theta)$ as the smallest index $i \in [n]$ for which $|w_i| \leq \tau \cdot \sigma_i$. If this inequality does not hold for any $i \in [n]$, we define $\ell(\tau) = \infty$.*

We now show how to use Theorem 21 and ideas from [Ser07] to prove Theorem 2. Given $\epsilon > 0$, we proceed by a case analysis, as in [Ser07], based on the value of the ϵ-critical index $\ell \overset{\text{def}}{=} \ell(\epsilon)$. If $\ell > L \overset{\text{def}}{=} \tilde{O}(1/\epsilon^2)$, Case IIa in [Ser07] says that f is ϵ-close to the L-junta g obtained by truncating the smallest $(n-L)$ weights, i.e. $g(x) = \mathrm{sign}(\sum_{i=1}^L w_i x_i - \theta)$. By applying Theorem 21 to g, we obtain an ϵ-approximator h with integer weights of magnitude $L^{O(1/\epsilon^{2/3})} = 2^{\tilde{O}(1/\epsilon^{2/3})}$, which is a 2ϵ-approximator for f. It remains to handle the case $\ell \leq L$. To do this, we use another fact from [Ser07]; that, for every value of ℓ, there exists an ϵ-approximator for f with integer weights of magnitude $\sqrt{n \ln(1/\epsilon)} \cdot 2^{O(\ell \log \ell)}$. If $\ell < K \overset{\text{def}}{=} 2/\epsilon^{2/3}$, this yields an ϵ-approximator with integer weights of magnitude $\sqrt{n} \cdot 2^{\tilde{O}(1/\epsilon^{2/3})}$ and we are done. To handle the case $K \leq \ell \leq L$, we use a combination of Gaussian anti-concentration (for the $n - \ell + 1$ smallest weights) and "Halász-type" anti-concentration (for the largest $\ell - 1$ weights).

Let us proceed with the analysis. We start by rounding the weights $w_\ell, \ldots, w_n$, exactly as in Case IIb in [Ser07], to get an ϵ-approximator $g(x) = \mathrm{sign}(\sum_{i=1}^n v_i x_i - \theta')$ for f with the following properties: (i) For $i \geq \ell$, each v_i is an integer of magnitude $O(\sqrt{n \ln(1/\epsilon)})$ and $\sum_{i=\ell}^n v_i^2 = O(n \ln(1/\epsilon)/\epsilon^2)$; (ii) It holds $|v_1| \geq |v_2| \geq \ldots \geq |v_{\ell-1}| > 1$. Our goal is to establish the existence of an ϵ-approximation h for g with small integer weights. To achieve this, we will use the fact that the "tail" of g has small integer coefficients, i.e. the integer–valued random variable $t(x) \overset{\text{def}}{=} \sum_{i=\ell}^n v_i x_i$ has small support.

Let $R, k > 0$ be integers. Denote by $\Omega(R, k)$ the set $\{\pm 1\}^{k-1} \times \{-R, -R+1, \ldots, R-1, R\}$. Now fix an integer $R_0 = \Theta(\sqrt{n} \ln(1/\epsilon)/\epsilon)$ and denote $\Omega_0 \overset{\text{def}}{=} \Omega(R_0, \ell)$. Consider the threshold function $h : \Omega_0 \to \{\pm 1\}$ defined by $h(y) = \mathrm{sign}(\sum_{i=1}^{\ell-1} v_i y_i + y_\ell - \theta')$, $y \in \Omega_0$. We claim that the threshold function $g' : \{-1,1\}^n \to \{-1,1\}$ defined by $g'(x) = h(x_1, \ldots, x_{\ell-1}, t(x))$ is ϵ-close to g. To see this note that $g'(x)$ equals $g(x)$ whenever $|t(x)| = |\sum_{i=\ell}^n v_i x_i| \leq R_0$, and this holds for a random x with probability $1 - \epsilon$ by a Hoeffding bound (since $R_0 \geq \sqrt{2 \ln(2/\epsilon) \sum_{i=\ell}^n v_i^2}$ by the definition of R_0 and property (i) of g).

At this point we use the following technical generalization of Lemma 22, whose proof is given in [DS09]:

Lemma 24. *Let $h' : \Omega(R,k) \to \{\pm 1\}$ be a threshold function that depends on all k variables. Suppose that $h'(y)$ has a representation as $\mathrm{sign}(\sum_{i=1}^{k} w'_i y_i - \theta')$ such that $|w'_1| \geq |w'_2| \geq \ldots \geq |w'_k| > 0$. There exists an alternate representation of h' as $\mathrm{sign}(\sum_{i=1}^{k} u_i y_i - \theta'')$ satisfying $1 = |u_1| \geq \cdots \geq |u_k| > 0$, with the following property: For $i \in [k-1]$ let $\Delta_i \stackrel{\text{def}}{=} |u_i| - |u_{i+1}|$. Then for any $j \in [k-2]$, the j-th biggest element of the (multiset) $\Delta_1, \ldots, \Delta_{k-1}$ is at least $\frac{1}{(2k+2R)\cdot(2k+2)^{2j+8}}$.*

Applying this lemma to h, i.e. setting $h' = h$, $R = R_0$ and $k = \ell$, and fixing $j \stackrel{\text{def}}{=} 1/\epsilon^{2/3} + 2 < K - 2 \leq \ell - 2$, we get a representation $\mathrm{sign}(\sum_{i=1}^{\ell} u_i y_i - \theta'')$ for h such that the j largest differences $\Delta_{i_1} = |u_{i_1}| - |u_{i_1+1}|, \ldots, \Delta_{i_j} = |u_{i_j}| - |u_{i_j+1}|$ are at least r_0, for $r_0 = \frac{1}{(2\ell+2R_0)\cdot(2\ell+2)^{2j+8}} = (1/\sqrt{n}) \cdot 2^{-\tilde{O}(1/\epsilon^{2/3})}$. (Note that the latter equality uses the fact that $\ell \leq L$.) This yields a set of $j' = 1/\epsilon^{2/3}$ weights $u_{l_1}, \ldots, u_{l_{j'}}$ – not including u_ℓ – whose absolute differences are at least r_0, i.e. for all $a \neq b \in [j']$, we have $|u_{l_a} - u_{l_b}| \geq r_0$.

We are now ready to use our proof template again. The alternate representation for h from above and the definition of g' imply that $g'(x)$ can be represented as $\mathrm{sign}(\sum_{i=1}^{\ell-1} u_i x_i + \sum_{i=\ell}^{n} u'_i x_i - \theta'')$, where $u'_i \stackrel{\text{def}}{=} u_\ell v_i$, $\ell \leq i \leq n$. By Halász's bound, Theorem 16, applied to the weights $u_{l_1}, \ldots, u_{l_{j'}}$, and the Extension Lemma 17 as before, we conclude that $p_{r_0}(u_1, \ldots, u'_n) = O(\epsilon)$. Finally, since the maximum weight in (the new representation for) g' is $O(\sqrt{n \log(1/\epsilon)})$ (as follows from the fact that $|u_i| \leq 1$, $i \in [\ell]$, and property (i) of g), Lemma 18 implies the existence of an $O(\epsilon)$-approximator for g' with integer weights each at most $n^{3/2} \cdot 2^{\tilde{O}(1/\epsilon^{2/3})}$. This concludes the proof of Theorem 2. ☐

D. Proof of Lemma 22

Recall Lemma 22:

Lemma 22. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a threshold function that depends on all n variables. There is a representation $\mathrm{sign}(\sum_{i=1}^{n} w_i x_i - \theta)$ for f with the following property: Suppose (reordering and rescaling weights if necessary) that $1 = |w_1| \geq \cdots \geq |w_n| > 0$. For $i \in [n-1]$ let $\Delta_i \stackrel{\text{def}}{=} |w_i| - |w_{i+1}|$. Then for any $k \in [n-2]$, the k-th biggest element of the (multiset) $\Delta_1, \ldots, \Delta_{n-1}$ is at least $\frac{1}{(2n+2)^{2k+8}}$.*

Proof: Let $f(x)$ be a threshold function. We first consider the case that f is odd, i.e. $f(x) = -f(-x)$ for all $x \in \{-1,1\}^n$; in this case f can be represented with a threshold of zero. Once we have established the result for such threshold functions we will use it to establish the general case.

By symmetry of $\{-1,1\}^n$ we may assume that f is monotone increasing in each coordinate x_i. By reordering coordinates we may assume that $\mathrm{Inf}_1(f) \geq \mathrm{Inf}_2(f) \geq \cdots \geq \mathrm{Inf}_n(f) > 0$ (the final inequality is strict because f depends on all n coordinates).

We consider the set $\mathcal{W} \subseteq \mathbb{R}^n$ of weight vectors $w = (w_1, \ldots, w_n)$ that satisfy the following properties:

1) $w \cdot x \geq 1$ for every $x \in \{-1,1\}^n$ such that $f(x) = 1$. Note that since f is odd these inequalities imply the corresponding inequalities for negative points, $w \cdot x \leq -1$ for every $x \in \{-1,1\}^n$ such that $f(x) = -1$.
2) $w_i - w_{i+1} \geq 1$ for all $i = 1, 2, \ldots, n-1$, and $w_n \geq 1$.

The first set of 2^{n-1} constraints says that $\mathrm{sign}(w \cdot x)$ is a valid representation for f (i.e. $f(x) = \mathrm{sign}(w \cdot x)$ for all $x \in \{-1,1\}^n$). The second set of n constraints says that no two weights are precisely the same and moreover all the weights are positive. (These are the new constraints that did not feature in the proof of [Hås05].)

Thus $\mathcal{W}$ is the feasible set of a linear program $\mathcal{LP}$ consisting of $2^{n-1} + n$ inequalities on $w_1, \ldots, w_n$: 2^{n-1} inequalities correspond to points of the hypercube $\{-1,1\}^n$ and n inequalities correspond to the set

$$D_n = \{(1,-1,0,\ldots,0)_{1\times n}, (0,1,-1,0,\ldots,0)_{1\times n}, \ldots, (0,\ldots,1,-1)_{1\times n}, (0,\ldots,0,1)_{1\times n}\}.$$

We claim that the linear program $\mathcal{LP}$ is feasible, or equivalently $\mathcal{W} \neq \emptyset$. Indeed, by simple standard arguments it can be shown that every odd threshold function $f : \{-1,1\}^n \to \{-1,1\}$ has a representation $\mathrm{sign}(w \cdot x)$ such that (i) for all $x \in \{-1,1\}^n$, it holds $\mathrm{sign}(w \cdot x) \neq 0$, and (ii) every partial sum of the weights is distinct, i.e. for all $I \neq J \subseteq [n]$ it holds $\sum_{i \in I} w_i \neq \sum_{i \in J} w_j$. The latter in particular implies that $w_1 \neq w_2 \neq \ldots \neq w_n$. Now recall that $\mathrm{Inf}_1(f) \geq \mathrm{Inf}_2(f) \geq \ldots \geq \mathrm{Inf}_n(f) > 0$ and that f is monotone increasing in all its coordinates. It is well known and easy to show (see e.g. [FP04]) that there is a representation $\mathrm{sign}(w \cdot x)$ of such a threshold function that satisfies $w_1 \geq w_2 \geq \ldots w_n > 0$. Therefore, we can scale the weights so that all the constraints in the linear program $\mathcal{LP}$ are simultaneously satisfied.

Having established that $\mathcal{W} \neq \emptyset$, we select a weight vector $w^* \in \mathcal{W}$ that maximizes the number of *tight* inequalities (i.e. satisfied with equality) in $\mathcal{LP}$. If more than one weight vector satisfies a maximum number of tight inequalities, we choose one arbitrarily. At this point, we invoke the following crucial claim:

Claim 25. *There exists a set of n points $y^{(1)}, \ldots, y^{(n)} \in f^{-1}(1) \cup D_n$ such that w^* is the* unique solution *of the linear system: $\{w \cdot y^{(i)} = 1 \mid i = 1, 2, \ldots, n\}$. (Henceforth, we shall denote this system by $(*)$.)*

The proof of the claim is essentially the same as in the proof of Muroga *et al.*'s [MTT61] classic upper bound on the size of integer weights that are required to express LTF's over $\{-1,1\}^n$ (see [DS09]).

Note that $(*)$ is a system of n linear equations in the variables $w_1, \ldots, w_n$ where each coefficient of each variable in the equations is $-1, 0$ or 1 and the right-hand side of each equation is 1. Since our goal is to prove a statement about the magnitude of the differences $w_i - w_{i+1}$, $i = 1, 2, \ldots, n-1$, we define an appropriate set of n new variables and rewrite $(*)$. In particular, we define the set of variables $\delta_1, \ldots, \delta_n$ as follows:

$$\delta_n = w_n, \qquad \delta_i = w_i - w_{i+1} \quad \text{for} \quad i = 1, \ldots, n-1.$$

This is equivalent to

$$w_n = \delta_n, \qquad w_i = \delta_i + \cdots + \delta_n \quad \text{for} \quad i = 1, \ldots, n-1.$$

We let δ denote $[\delta_1, \ldots, \delta_n]$. By rewriting $(*)$, we get an *equivalent* system $(**)$ of n equations in variables $\delta_1, \ldots, \delta_n$ where the coefficients of each variable in each equation are integers in the range $[-n, n]$ and all the right-hand sides remain 1. Hence, the linear system $(**)$ has the unique *strictly* positive solution

$$\delta_n^* = w_n^*, \qquad \delta_i^* = w_i^* - w_{i+1}^* \quad \text{for} \quad i = 1, \ldots, n-1.$$

At this point we reorder the variables δ_i in decreasing order of magnitude of the δ_i^*'s. We thus get a new set of variables $\tau_1, \ldots, \tau_n$ such that

$$\tau_i^* = i\text{-th largest of } \{\delta_1^*, \ldots, \delta_n^*\},$$

breaking ties arbitrarily. We similarly denote $\tau = [\tau_1, \ldots, \tau_n]$.

So $(**)$ is now a system of n equations in variables $\{\tau_i\}_{i \in [n]}$, where the coefficients of each variable in each equation are integers in the range $[-n, n]$ and all the right-hand sides are still 1. The values $\tau_1^*, \ldots, \tau_n^*$ in the unique solution of this system are strictly positive and ordered in decreasing order of magnitude. Let us write

$$\alpha_{j1}\tau_1 + \alpha_{j2}\tau_2 + \ldots + \alpha_{jn}\tau_n = 1$$

for the j-th equation where $\alpha_{ij},\ i, j \in [n]$ are integers in $[-n, n]$. It is not difficult to see that the above system is equivalent to the following system of n equations in $\tau_1, \ldots, \tau_n$:

$$\alpha_{j1}\tau_1 + \alpha_{j2}\tau_2 + \ldots + \alpha_{jn}\tau_n = \alpha_{11}\tau_1 + \alpha_{12}\tau_2 + \ldots + \alpha_{1n}\tau_n$$
$$\text{for} \quad j = 2, 3, \ldots, n, \qquad \text{and} \quad \tau_n = \tau_n^*.$$

Each of the first $n-1$ equations is homogeneous and can be rewritten as $\tau \cdot z^{(j)} = 0$, where $z^{(j)}$ is a vector whose entries are integers in $[-2n, 2n]$. So we have that $\tau^* = [\tau_1^*, \ldots, \tau_n^*]$ is the unique solution to a linear system:

$$Z\tau = b \tag{5}$$

where Z is a non-singular $n \times n$ matrix with entries that are integers in $[-2n, 2n]$ and with last row $(0, \ldots, 0, 1)$, and b is $[0, 0, \ldots, 0, \tau_n^*]$.

Recall that $\tau_1^* \geq \cdots \geq \tau_n^* > 0$. We now show that each τ_k^* is somewhat large compared to w_1^*. The case $k = 1$ is easy: since $\sum_{i=1}^n \tau_i^* = w_1^*$, we have $\tau_1^* \geq w_1^*/n$.

Fix any $k \in \{2, \ldots, n\}$. After possibly reordering the rows of Z, the $(k-1)$-dimensional vector $[1, 0, \ldots, 0]$ can be expressed as a linear combination $a_1 R_1 + \cdots + a_{k-1} R_{k-1}$ where R_i is the i-th row of the $(k-1) \times (k-1)$ upper left submatrix of Z. Since all entries in Z are integers in $[-2n, 2n]$, Cramer's Rule implies that each $|a_i|$ is at most the maximum determinant of any $(k-1) \times (k-1)$ matrix with all entries in $[-2n, 2n]$; this is easily seen to be at most $(k-1)!(2n)^{k-1}$. It follows that there is a linear combination of the first $k-1$ equations of (5) which yields

$$\tau_1 = \sum_{j=k}^{n} \gamma_j^k \tau_j \tag{6}$$

where each $|\gamma_j^k|$ is at most $(k-1) \cdot (2n) \cdot (k-1)!(2n)^{k-1} \leq (2(k-1)n)^k$. From (6), setting $\tau = \tau^*$ and recalling that the τ_i^*'s are positive and ordered by magnitude, we now get $\tau_1^* \leq (n-k+1)(\max_j |\gamma_j^k|)\tau_k^*$ which implies

$$\tau_k^* \geq \frac{\tau_1^*}{(2(k-1)n)^k(n-k+1)} \geq \frac{\tau_1^*}{(2n)^{2k+1}} \tag{7}$$

Observing that $\sum_{i=1}^n \tau_i^* = w_1^*$, we have $\tau_1^* \geq w_1^*/n$, which implies

$$\tau_k^* \geq \frac{w_1^*}{(2n)^{2k+2}}.$$

Finally, we observe that for $k \in [n-1]$, the k-th biggest element of the multiset $\Delta_1, \ldots, \Delta_{n-1}$ (see the lemma statement) is at least τ_{k+1}. (It is either τ_{k+1} or τ_k depending on whether or not $\delta_n^* = w_n^*$ is among the k largest elements of $\{\delta_1^*, \ldots, \delta_n^*\}$.) Renormalizing so that the largest weight is 1, we have shown that for odd f, the k-th biggest element of the multiset $\Delta_1, \ldots, \Delta_{n-1}$ is at least $\frac{1}{(2n)^{2k+4}}$. This completes the proof of Lemma 22 for the case that f is odd.

We now treat the case where f is not odd, i.e. f has a nonzero threshold. We do this by considering the threshold function $g : \{-1, 1\}^{n+1} \to \{-1, 1\}$ which has zero threshold, n weights the same as f, and an $(n+1)$-st weight which is the threshold of f. The result for the zero-threshold case shows that g has a representation $\text{sign}(w_1 x_1 + \cdots + w_n x_n + w_{n+1} x_{n+1})$ where $|w_1| \geq \cdots \geq |w_{n+1}|$, and letting $\Delta_i = |w_i| - |w_{i+1}|$ for $i \in [n]$, the k-th biggest element of $\Delta_1, \ldots, \Delta_n$ is at least $\frac{|w_1|}{(2n+2)^{2k+4}}$ for any $k \in [n]$.

We now observe that for $k \in [n-2]$, the k-th biggest gap between the magnitudes of the w_i's that correspond to actual weights of f is at least the $(k+2)$-th biggest element of $\Delta_1, \ldots, \Delta_n$. This holds since at most two of the values $\Delta_j = |w_j| - |w_{j+1}|$ can involve the weight $w_{j^\star}$ which corresponds to the threshold of f, as opposed to one of its actual weights. Since $|w_1|$ is at least as large as the absolute value of the largest actual weight of f, we get that for $k \in [n-2]$, the k-th biggest gap between the magnitudes of the actual weights of f is at least (largest weight of $f)/(2n+2)^{2k+8}$. Renormalizing so that the largest magnitude weight of f is 1, Lemma 22 is proved. ∎

IV. Conclusions and Future Work

We have already discussed directions for future work relating to Theorem 1 in Section II-D. Regarding Theorem 2, we feel that our high-level approach using anti-concentration holds promise for substantial further progress. Significant strengthenings of Halász's anti-concentration bound are known under stronger restrictions on the additive structure of the weights $w_1, \dots, w_n$, see e.g. [Vu08], [TV08]. Can corresponding extensions of Lemma 22 be established, proving that every threshold function admits a representation with weights that have the required structure? Perhaps every threshold function f can be ϵ-approximated using integer weights at most $\mathrm{poly}(n) \cdot 2^{\mathrm{polylog}(1/\epsilon)}$. We hope that further study of our anti-concentration based approach may yield such a bound.

Acknowledgment

We thank Ryan O'Donnell for asking a question that led to Theorem 1. I.D. was supported by NSF grants CCF-0728736, CCF-0525260, and by an Alexander S. Onassis Foundation Fellowship. R.A.S. was supported by NSF grants CCF-0347282, CCF-0523664 and CNS-0716245, and by DARPA award HR0011-08-1-0069.

References

[AV97] N. Alon and V. H. Vu. Anti-Hadamard Matrices, Coin Weighing, Threshold Gates, and Indecomposable Hypergraphs. *Journal of Combinatorial Theory, Series A*, 79(1):133–160, 1997.

[Bec75] W. Beckner. Inequalities in fourier analysis. *Annals of Mathematics*, 102:159–182, 1975.

[Bon70] A. Bonami. Etude des coefficients fourier des fonctiones de $l^p(g)$. *Ann. Inst. Fourier (Grenoble)*, 20(2):335–402, 1970.

[Bou02] J. Bourgain. On the distributions of the fourier spectrum of boolean functions. *Israel J. Math.*, 131:269–276, 2002.

[BS92] J. Bruck and R. Smolensky. Polynomial threshold functions, AC^0 functions and spectral norms. *SIAM Journal on Computing*, 21:33–42, 1992.

[BT96] N. Bshouty and C. Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 1996.

[BW06] A. Beimel and E. Weinreb. Monotone circuits for monotone weighted threshold functions. *Information Processing Letters*, 97(1):12–18, 2006.

[CKK+06] S. Chawla, R. Krauthgamer, R. Kumar, Y. Rabani, and D. Sivakumar. On the hardness of approximating multicut and sparsest-cut. *Computational Complexity*, 15(2):94–114, 2006.

[Der65] M. Dertouzos. *Threshold logic: a synthesis approach*. MIT Press, Cambridge, MA, 1965.

[DGJ+09] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. Manuscript, 2009.

[DL36] W. Doeblin and P. Lévy. Calcul des probabilités. Sur les sommes de variables aléatoires indépendantes 'a dispersions bornées inférieurement. *C.R. Acad. Sci.*, 202:2027–2029, 1936.

[DS05] I. Dinur and S. Safra. On the hardness of approximating minimum vertex-cover. *Annals of Mathematics*, 162(1):439–485, 2005.

[DS09] I. Diakonikolas and R. Servedio. Improved approximation of linear threshold functions. Full version available at http://www.cs.columbia.edu/ rocco/papers/ccc09.html, 2009.

[Erd45] P. Erdős. On a lemma of Littlewood and offord. *Bull. Amer. Math. Soc.*, 51:898–902, 1945.

[Erd65] P. Erdős. Extremal Problems in Number Theory. *Proc. Sympos. Pure Math.*, 8:181–189, 1965.

[Ess68] C.G. Esséen. On the concentration function of a sum of independent random variables. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 9:290–308, 1968.

[FGKP06] V. Feldman, P. Gopalan, S. Khot, and A. Ponnuswami. New results for learning noisy parities and halfspaces. In *Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 563–576, 2006.

[FK96] E. Friedgut and G. Kalai. Every monotone graph property has a sharp threshold. *Proceedings of the AMS*, 124:2993–3002, 1996.

[FP04] A. Fiat and D. Pechyony. Decision trees: More theoretical justification for practical algorithms. In *Algorithmic Learning Theory, 15th International Conference (ALT 2004)*, pages 156–170, 2004.

[Fri98] E. Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):474–483, 1998.

[Gol06] P. Goldberg. A Bound on the Precision Required to Estimate a Boolean Perceptron from its Average Satisfying Assignment. *SIAM Journal on Discrete Mathematics*, 20:328–343, 2006.

[Gro75] L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

[GS07] D. Glasner and R. Servedio. Distribution-free testing lower bounds for basic boolean functions. In *Proc. 11th International Workshop on Randomization and Computation (RANDOM)*, pages 494–508, 2007.

[Hal77] G. Halász. Estimates for the concentration function of combinatorial number theory and probability. *Period. Math. Hungar.*, 8(3):197–211, 1977.

[Hås94] J. Håstad. On the size of weights for threshold gates. *SIAM Journal on Discrete Mathematics*, 7(3):484–492, 1994.

[Hås05] J. Håstad. Personal communication, 2005.

[Hon87] J. Hong. On connectionist models. Technical Report Technical Report 87-012, Dept. of Computer Science, University of Chicago, 1987.

[Kal07] A. Kalai. Learning nested halfspaces and uphill decision trees. In *Proc. of the 20th Annual Conference in on Computational Learning Theory — (COLT 2007)*, pages 378–392, 2007.

[KKL88] J. Kahn, G. Kalai, and N. Linial. The influence of variables on boolean functions. In *Proceedings of the 29th Annual Symposium on Foundations of Computer Science*, pages 68–80, 1988.

[KKMS08] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[Kol60] A.N. Kolmogorov. Sur les propriétés des fonctions de concentration de M. P. Lévy. *Ann. Inst. H. Poincaré*, 16:27–34, 1958-60.

[KR06] Robert Krauthgamer and Yuval Rabani. Improved lower bounds for embeddings into l_1. In *17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1010–1017, 2006.

[KR08] S. Khot and O. Regev. Vertex cover might be hard to approximate to within $2-\epsilon$. *Journal of Computer & System Sciences*, 74(3):335–349, 2008.

[KS08] S. Khot and R. Saket. On hardness of learning intersection of two halfspaces. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 345–354, 2008.

[LO43] J. E. Littlewood and A. C. Offord. On the number of real roots of a random algebraic equation. *III. Rec. Math. [Mat. Sbornik] N.S.*, 12:277–286, 1943.

[MORS09] K. Matulef, R. O'Donnell, R. Rubinfeld, and R. Servedio. Testing halfspaces. In *Proc. 20th Annual Symposium on Discrete Algorithms (SODA)*, 2009.

[MTT61] S. Muroga, I. Toda, and S. Takasu. Theory of majority switching elements. *J. Franklin Institute*, 271:376–418, 1961.

[Mur71] S. Muroga. *Threshold logic and its applications*. Wiley-Interscience, New York, 1971.

[NS94] N. Nisan and M. Szegedy. On the degree of Boolean functions as real polynomials. *Computational Complexity*, 4:301–313, 1994.

[OS07] R. O'Donnell and R. Servedio. Learning monotone decision trees in polynomial time. *SIAM Journal on Computing*, 37(3):827–844, 2007.

[OS08] R. O'Donnell and R. Servedio. The Chow Parameters Problem. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 517–526, 2008.

[Rag88] P. Raghavan. Learning in threshold networks. In *First Workshop on Computational Learning Theory*, pages 19–27, 1988.

[Rog73] B.A. Rogozin. An integral-type estimate for concentration functions of sums of independent random variables. *Dokl. Akad. Nauk SSSR*, 211:1067–1070, 1973.

[RS08] Y. Rabani and A. Shpilka. Explicit construction of a small epsilon-net for linear threshold functions. Available at http://www.cs.technion.ac.il/shpilka/publications.html, 2008.

[RV08] M. Rudelson and R. Vershynin. The Littlewood-Offord Problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

[Ser07] R. Servedio. Every linear threshold function has a low-weight approximator. *Computational Complexity*, 16(2):180–209, 2007.

[She07] A. Sherstov. Halfspace matrices. In *Proc. 22nd Annual IEEE Conference on Computational Complexity (CCC)*, pages 83–95, 2007.

[SRK95] K.-Y. Siu, V.P. Roychowdhury, and T. Kailath. *Discrete Neural Computation: A Theoretical Foundation*. Prentice-Hall, Englewood Cliffs, NJ, 1995.

[SS65] A. Sárközy and E. Szemerédi. Über ein Problem von Erdős und Moser. *Acta Arithmetica*, 11:205–208, 1965.

[TV06] T. Tao and V. Vu. *Additive Combinatorics*. Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 2006.

[TV08] T. Tao and V. Vu. From the Littlewood-Offord problem to the Circular Law: universality of the spectral distribution of random matrices. Available at http://arxiv.org/abs/0810.2994, 2008.

[Vu08] V. Vu. A structural approach to subset-sum problems. Available at http://arxiv.org/abs/0804.3211, 2008.