Low-weight Halfspaces for Sparse Boolean Vectors

Philip M. Long
NEC Labs America
plong@nec-labs.com

Rocco A. Servedio
Columbia University
rocco@cs.columbia.edu

ABSTRACT

For $S \subseteq \{0,1\}^n$, a Boolean function $f : S \to \{-1,1\}$ is a *halfspace over* S if there exist $w \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$ such that $f(x) = \text{sign}(w \cdot x - \theta)$ for all $x \in S$. We give bounds on the size of integer weights $w_1, \ldots, w_n \in \mathbb{Z}$ that are required to represent halfspaces over Hamming balls centered at 0^n, i.e. halfspaces over $S = \{0,1\}^n_{\leq k} \stackrel{\text{def}}{=} \{x \in \{0,1\}^n : x_1 + \cdots + x_n \leq k\}$. Such weight bounds for halfspaces over Hamming balls have immediate consequences for the performance of learning algorithms in the increasingly common scenario of learning from very high-dimensional categorical examples which are such that only a small number of features are active in each example.

We give upper and lower bounds on weight both for exact representation (when $\text{sign}(w \cdot x - \theta)$ must equal $f(x)$ for every $x \in S$) and for ε-approximate representation (when $\text{sign}(w \cdot x - \theta)$ may disagree with $f(x)$ for up to an ε fraction of points $x \in S$). Our results show that extremal bounds for exact representation are qualitatively rather similar whether the domain is all of $\{0,1\}^n$ or the Hamming ball $\{0,1\}^n_{\leq k}$, but extremal bounds for approximate representation are qualitatively very different between these two domains.

Categories and Subject Descriptors

F.1.3 [**Theory of Computation**]: Computation by Abstract Devices—*Complexity measures and classes*

General Terms

Theory

Keywords

Linear threshold functions, halfspaces, Boolean functions, Boolean hypercube

ITCS'13, January 9–12, 2012, Berkeley, California, USA.

1. INTRODUCTION AND MOTIVATION

Many machine learning problems involve data with Boolean features. For example, news articles may be classified based on which words they contain, and analogous "bag of visual words" representations have become popular for image classification as well. In many such modern machine learning problems, the set of all possible features is extremely large but each example has only a small number of active features. For example, in a search engine scenario the set of all possible features might be the set of all words that appear in any query, while the active features in a given example might be the much smaller set of words appearing in that query. In such a setting, the space of all possible examples is contained in a Hamming ball centered at the origin $\{0,1\}^n_{\leq k}$ where $k \ll n$. It is intuitively clear that limiting the domain in this way may make learning easier; studying the complexity of learning over such domains can potentially lend insight useful for the design of practical algorithms for learning in such settings.

Let S be a subset of the Boolean hypercube $\{0,1\}^n$. We say that a Boolean function $f : S \to \{-1,1\}$ is a *halfspace over* S if there exist $w \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$ such that $f(x) = \text{sign}(w \cdot x - \theta)$ for all $x \in S$. The pair (w, θ) is an *integer representation* of f if $w \in \mathbb{Z}^n$. The *weight* of an integer representation is $\max_{i=1,\ldots,n} |w_i|$. The weight of a halfspace f over S is the smallest weight of any integer representation which computes f correctly on all $x \in S$.

In this paper we give a detailed study of the weight of halfspaces, both exact and approximate, over *Hamming balls of radius* k, i.e. we study halfspaces over the domain

$$S = \{0,1\}^n_{\leq k} \stackrel{\text{def}}{=} \{x \in \{0,1\}^n : x_1 + \cdots + x_n \leq k\}.$$

The weight of an integer halfspace is closely related to the "margin" by which it classifies examples in S. Since halfspaces play a central role in machine learning, and the margin of a halfspace H is an important measure of the difficulty of learning H, we are naturally motivated to understand the weight of halfspaces over Hamming balls as an initial step toward understanding the impact of sparsity in feature vectors on learning complexity.

Many researchers have studied the weight of halfspaces over the entire Boolean cube (corresponding to taking $S = \{0,1\}^n$), see e.g. [10, 9, 12, 4, 13, 11, 14, 8, 6, 1, 16, 3]. Upper and lower bounds have been obtained both for exact representation as described above, and for a relaxed scenario in which the integer-weight halfspace $\text{sign}(w \cdot x - \theta)$ need only ε-approximate the function f (i.e. we allow $\Pr_{x \in S}[\text{sign}(w \cdot x - \theta) \neq f(x)]$ to be at most ε for some given approximation

parameter $\varepsilon > 0$). We describe these previous results in detail in Section 1.1.

1.1 Previous Work and Our Results

In this section we review prior work on the weight of halfspaces (all of the previous work that we are aware of deals with halfspaces over the entire Boolean cube $\{0,1\}^n$), and state our results for halfspaces over the Hamming ball $\{0,1\}^n_{\leq k}$.

Prior work on exact representation of halfspaces over $\{0,1\}^n$. It has been known at least since the 1960s [10] that every halfspace over $\{0,1\}^n$ has weight at most $n^{O(n)}$ (this fact has been rediscovered several times, see e.g. [7, 12]). Since there are $2^{\Omega(n^2)}$ halfspaces over $\{0,1\}^n$ a counting argument shows that there exist halfspaces over $\{0,1\}^n$ that require weight $2^{\Omega(n)}$, and specific halfspaces that require weight $2^{\Omega(n)}$ have been known for decades [9, 5]. [6] exhibited a specific halfspace that has weight $n^{\Omega(n)}$ and his construction was subsequently refined by [1]. So the weight of exact representations of halfspaces over all of $\{0,1\}^n$ is by now quite well understood.

Our results on exact representation of halfspaces over $\{0,1\}^n_{\leq k}$. We give an essentially complete picture of the weight of halfspaces over Hamming balls $\{0,1\}^n_{\leq k}$ for all values of k. First, it is easy to see that for $k \in \{0,1\}$ every halfspace over $\{0,1\}^n_{\leq k}$ has an integer representation of weight 1. For $k = 2$, by analyzing a greedy construction we show (Theorem 2) that every halfspace over $\{0,1\}^n_{\leq 2}$ has weight $O(n)$, and we observe that a simple explicit halfspace has weight $\Theta(n)$.

Things get more interesting beyond $k = 2$. Using a linear programming argument, we show (Theorem 1) that for every $k \geq 3$, every halfspace over $\{0,1\}^n_{\leq k}$ has an integer representation of weight $(k+2)^{(n+1)/2}$, and we show that already for $k = 3$ there is a simple explicit halfspace for which any integer representation must have weight $2^{\Omega(n)}$. Our main lower bound result for exact representation (Theorem 4) is a general lower bound showing that for every $k \geq 3$, there is an explicit halfspace over $\{0,1\}^n_{\leq k}$ that requires integer weight $k^{\Omega(n)}$. This is established via a construction that carefully combines Håstad's halfspace [6] with a "decision list" type construction. Our lower bound shows that our upper bound on the weight of halfspaces over $\{0,1\}^n_{\leq k}$ is essentially the best possible.

Prior work on approximation over $\{0,1\}^n$. The lower bound of [6] immediately implies that there is an n-variable halfspace f over $\{0,1\}^n$ which is such that any halfspace $\mathrm{sign}(w\cdot x-\theta)$ which satisfies $\Pr_{x\in\{0,1\}^n}[\mathrm{sign}(w\cdot x) \neq f(x)] \leq \varepsilon$ must have weight $1/\varepsilon^{\Omega(\log\log(1/\varepsilon))}$. [16] showed that every n-variable halfspace over $\{0,1\}^n$ can be ε-approximated by a halfspace of weight $\sqrt{n}\cdot 2^{\tilde{O}(1/\varepsilon^2)}$, and showed an $\Omega(\sqrt{n})$ lower bound for constant ε. The upper bound was subsequently improved (as a function of ε) to weight $n^{3/2}\cdot 2^{\tilde{O}(1/\varepsilon^{2/3})}$ by [3], and very recently [2] have further improved the upper bound to $\sqrt{n}\cdot(1/\varepsilon)^{O(\log^2(1/\varepsilon))}$.

Our results on approximation over $\{0,1\}^n_{\leq k}$. We study the weight required to ε-approximate halfspaces over $\{0,1\}^n_{\leq k}$, i.e. given a halfspace f we now allow the integer-weight halfspace $\mathrm{sign}(w\cdot x-\theta)$ to disagree with $f(x)$ on an ε fraction of all points in $\{0,1\}^n_{\leq k}$. (For the informal discussion here k should be viewed as "small" compared to n; precise bounds on k are given in the actual detailed theorem statements.) As our main positive result (Theorem 5), we show that for every halfspace f over $\{0,1\}^n_{\leq k}$ there is a halfspace that ε-approximates f and has integer weights each of which is at most $k^{O(k/\varepsilon)}$, independent of n. This proof combines linear programming arguments with upper bounds on the edge boundary of monotone Boolean-valued functions over the discrete domain $\{1,\ldots,t\}^k$.

As our main negative result (Theorem 7), we show that for any constant $k \geq 3$ there is a simple explicit halfspace f (the "decision list" halfspace, which we denote DL) which is such that any ε-approximator of f over $\{0,1\}^n_{\leq k}$ must have weight $k^{\Omega(1/\varepsilon^{1/(k-1)})}$. This shows that an inverse exponential dependence on $1/\varepsilon$ is necessary in any upper bound.

Finally, we give a detailed analysis of the specific "decision list" halfspace DL and show (Theorem 8) that for this particular function the general weight upper bound of Theorem 5 can be strengthened to $k^{O(k/\sqrt{\varepsilon})}$. This shows that strengthening the analysis of the DL function that is given in Theorem 7 will not be enough to improve that lower bound to match the general upper bound of Theorem 5.

Discussion. Our results show that (as long as $k \geq 3$) the extremal bounds on the weights required for exact representation of halfspaces are fairly similar whether the domain is $\{0,1\}^n_{\leq k}$ or $\{0,1\}^n$; in the former case the "right" weight bound is $k^{\Theta(n)}$, while in the latter case it is $n^{\Theta(n)}$. For ε-approximate representation, though, our results show that there are two interesting qualitative differences between the "right" weight bounds for the two domains. First, our $k^{O(k/\varepsilon)}$ upper bound (independent of n) for $\{0,1\}^n_{\leq k}$ stands in contrast with the $\Omega(\sqrt{n})$ lower bound of [16] for $\{0,1\}^n$; so for Hamming balls *no* dependence on n is necessary in the weights, whereas for the Boolean cube a polynomial dependence is required. Second, our $k^{\Omega(1/\varepsilon^{1/(k-1)})}$ lower bound shows that for any fixed constant k, some halfspaces over $\{0,1\}^n_{\leq k}$ require any ε-approximator to have weights that are *exponential* in $1/\varepsilon$. This is in sharp contrast with the recent [2] upper bound, which shows that over $\{0,1\}^n$ it is always possible to construct an ε-approximating halfspace with integer weights that are only quasipolynomial in $1/\varepsilon$.

Preliminaries. Note that under the correspondence $-1 \leftrightarrow 0$, $1 \leftrightarrow 1$ an integer-weight halfspace $\mathrm{sign}(w\cdot x-\theta)$ over the hypercube $\{-1,1\}^n$ corresponds to an integer-weight halfspace $\mathrm{sign}(2w\cdot x-(\theta+w_1+\cdots+w_n))$ over the hypercube $\{0,1\}^n$. So we may work either over the Hamming ball $\{0,1\}^n_{\leq k} = \{x \in \{0,1\}^n : x_1+\cdots+x_k \leq k\}$ of the 0/1 hypercube, or over the Hamming ball $\{-1,1\}^n_{\leq k} = \{x \in \{-1,1\}^n : x_1+\cdots+x_k \leq -n+2k\}$ of the $+1/-1$ hypercube; weight bounds obtained for one domain will carry over to the other one with at most a factor of 2 difference. Similarly we may also work over $\{-1,1\}^n_{\geq n-k} = \{x \in \{-1,1\}^n : x_1+\cdots+x_k \geq n-2k\}$; sometimes this will be the most convenient.

Some more useful observations: If $\mathrm{sign}(w\cdot x-\theta)$ is a halfspace with integer coefficients over any subset $S \subseteq \{-1,1\}^n$ or $S \subseteq \{0,1\}^n$, then it is easy to see that w.l.o.g. we may modify the threshold θ to be of the form (integer $+\frac{1}{2}$). We also note that if $\mathrm{sign}(w\cdot x-\theta)$ is an integer-weight halfspace with such a threshold that computes a function f

over $\{-1,1\}^n_{\leq k}$, then $\text{sign}(-w \cdot x + \theta)$ is a halfspace of the same weight computing $-f$ over $\{-1,1\}^n_{\leq k}$; so to bound the weight of f over $\{-1,1\}^n_{\leq k}$ it is enough to bound the weight of $-f$.

Finally, we establish some useful notation. We write $[n]$ to denote the set $\{1,\ldots,n\}$. For $i \in [n]$ we write e_i to denote the unit vector in $\mathbb{R}^n$ whose only nonzero entry is a 1 in the i-th coordinate. We let $DL(x)$ denote the "decision list" halfspace over $\{0,1\}^n$ that is defined as follows: $DL(x)$ equals $(-1)^i$, where i is the largest index such that $x_i = 1$. To see that $DL(x)$ is a halfspace, we observe that it can be represented as $DL(x) = \text{sign}(\sum_{i=1}^n (-2)^i x_i)$.

2. UPPER BOUNDS FOR EXACT REPRESENTATION

We start with a general upper bound. The proof (see Appendix A) is a straightforward modification of standard integer weight upper bound arguments for halfspaces over $\{0,1\}^n$ (see e.g. [10, 6]) adapted to the domain $\{0,1\}^n_{\leq k}$.

THEOREM 1. *For $3 \leq k \leq n$, every halfspace over $\{0,1\}^n_{\leq k}$ has weight at most $(k+2)^{(n+1)/2}$.*

We note that the proof given above actually holds for all $k \geq 1$ (not just $k \geq 3$ as in the theorem statement), but much stronger bounds are possible for $k = 1, 2$. For $k = 1$, it is easy to see that every halfspace over $\{0,1\}^n_{\leq 1}$ has an integer representation of weight 1. For $k = 2$ Theorem 1 only gives an upper bound that is exponential in n, but the true upper bound for $k = 2$ is actually linear in n:

THEOREM 2. *Every halfspace f over $\{0,1\}^n_{\leq 2}$ has weight $O(n)$.*

The proof of Theorem 2 is in Appendix B. Before giving the proof, here is some some high-level intuition. Since we are working over $\{0,1\}^n_{\leq 2}$, intuitively in order to set the weight v_j of variable x_j correctly the "only constraint that matters" is how many of the other $n-1$ variables x_i are such that $f(e_i + e_j) = 1$. The proof shows that a suitable greedy approach of setting the weights can satisfy all these constraints taking all the weights to be $O(n)$ in absolute value.

We note that for odd n the decision list halfspace $DL(x) = \text{sign}(\sum_{i=1}^n (-2)^i x_i)$ requires integer weight at least $(n-1)/2$ over $\{0,1\}^n_{\leq 2}$, and thus the $O(n)$ upper bound of Theorem 2 is tight up to a constant factor. To see this, suppose that $\text{sign}(v \cdot x - \theta)$ has integer weights and computes DL correctly over $\{0,1\}^n_{\leq 2}$. By considering inputs of the form e_i where i ranges from 1 to n we see that $v_i \geq \theta$ for i even and $v_i < \theta$ for i odd. By considering inputs of the form $e_i + e_{i+1}$ we see that $v_1 > v_3 > v_5 > \cdots > v_n$ and $v_2 < v_4 < v_6 < \cdots < v_{n-1}$, so there are n distinct integer weights and the largest magnitude weight must be at least $(n-1)/2$ as claimed.

3. LOWER BOUNDS FOR EXACT REPRESENTATION

In this section we give lower bounds on the weight required to exactly represent various halfspaces over $\{-1,1\}^n_{\leq k}$ for $k \geq 3$. We first note that simple counting arguments do not give very good lower bounds (see Appendix C for details).

An exponential lower bound for a simple function. We now observe that even for $k = 3$ a simple halfspace gives an exponential lower bound. In Appendix C we prove:

PROPOSITION 3. *The function $DL(x)$ has weight $2^{\Omega(n)}$ over $\{0,1\}^n_{\leq 3}$.*

Proposition 3 gives an exponential lower bound but in general it does not match the $k^{O(n)}$ upper bound provided by Theorem 1, since the DL function has weight 2^n over the whole boolean cube. As our main lower bound result for exact representation we match the upper bound of Theorem 1 (up to an absolute constant in the exponent) and prove the following:

THEOREM 4. *Let k be an integer of the form $k = 2^\ell$, and let $n = rk + 1$ for some integer r. There is a halfspace G (defined explicitly below) over $\{-1,1\}^n_{\geq n-2k-1}$ for which the weight of any integer representation over $\{-1,1\}^n_{\geq n-2k-1}$ is at least $2^{(n \log k)/2 - O(n)}$, i.e. $k^{\Omega(n)}$.*

We recall that in [6] Håstad gave an explicit halfspace over $\{-1,1\}^k$ and proved that its weight over $\{-1,1\}^k$ is $k^{\Theta(k)}$. Our construction builds on his construction; indeed our $n = (rk+1)$-variable halfspace may be viewed as r copies of Håstad's halfspace "concatenated" in a careful way (the exact meaning of this will be clearer when we describe our construction in detail below).

Setup. First some notational preliminaries: since $k = 2^\ell$ we may view a k-bit string as a function from $\{-1,1\}^\ell$ to $\{-1,1\}$, and for f a k-bit string and $j \in \{0,\ldots,k-1\}$ we write $f(j)$ to denote the j-th coordinate of such a string. For $f, g \in \{-1,1\}^k$ we write (f,g) to denote the inner product $\sum_{j=0}^{k-1} f(j)g(j)$. Note that for $f, g \in \{-1,1\}^k$ we have $|(f,g)| \leq k$.

Following the notation from [6], for $\alpha \subseteq [\ell] = \{1,\ldots,\ell\}$ let φ_α denote the parity function $\varphi_\alpha(x) = \prod_{i \in \alpha} x_i$ over the variables in α. Again following [6], let $\alpha_0, \ldots, \alpha_{k-1}$ be an ordering of subsets of $[\ell]$ such that $|\alpha_i| \leq |\alpha_{i+1}|$ and the symmetric difference $\alpha_i \Delta \alpha_{i+1}$ always satisfies $|\alpha_i \Delta \alpha_{i+1}| \leq 2$. Note that α_0 is the empty set and thus φ_{α_0} is the k-bit string consisting of all 1's, while for each $j = 1, \ldots, k-1$ we have that φ_{α_j} is a k-bit string with exactly half of its entries -1.

Writing $f : \{-1,1\}^\ell \to \{-1,1\}$ in terms of its Fourier representation as $f(j) = \sum_{i=0}^{k-1} \widehat{f}(\alpha_i)\varphi_{\alpha_i}(j)$ we see that

$$(f, \varphi_{\alpha_i}) = k\widehat{f}(\alpha_i),$$

so we may view each inner product (f, φ_{α_i}) as a scaled Fourier coefficient of f.

For $f \in \{-1,1\}^n$ we decompose f by writing it as

$$(b, f^1, \ldots, f^r)$$

where $b \in \{-1,1\}$ and each f^i is a k-bit string. We sometimes refer to f^i as the "i-th block" of f and we write $f^i(j)$ to denote the j-th coordinate of the i-th block of f.

The construction. Let $G : \{-1,1\}^n_{\geq n-2k-1} \to \{-1,1\}$ denote the n-variable function

$$G(b, f^1, \ldots, f^r) \stackrel{\text{def}}{=} \text{sign}\left(b + \sum_{i=1}^{r} \sum_{j=1}^{k-1} (k+1)^{k(i-1)+j} (f^i, \varphi_{\alpha_j})\right).$$

(Note that the inner sum starts with $j = 1$ and not 0; this will be important later.) Since each $(f^i, \varphi_{\alpha_j})$ is a (± 1)-weighted sum of the coordinates of f^i, it is clear that G

is a halfspace with weight at most $k^{O(n)}$. We will show that any integer-weight halfspace for G over $\{-1,1\}^n_{\leq n-2k-1}$ must have some weight that is at least $k^{\Omega(n)}$.

To get some more intuition for the function G, note that for a block f^i we have that $(f^i, \varphi_{\alpha_1}) = \cdots = (f^i, \varphi_{\alpha_{k-1}}) = 0$ if and only if f^i is one of the two inputs $(1,\ldots,1)$ or $(-1,\ldots,-1)$ (this is because the constant $+1$ function and the constant -1 function are the only two Boolean functions that have all nonconstant Fourier coefficients equal to zero). So in words, given an input $f = (b, f^1, \ldots, f^r)$ the value $G(f)$ is obtained as follows: If any block is neither constantly $+1$ nor constantly -1, let i be the largest such block, and output the sign of the Fourier coefficient $\widehat{f^i}(\alpha_j)$ where j is the largest index such that $\widehat{f^i}(\alpha_j)$ is nonzero. Otherwise output the bit b.

The high-level intuition behind the lower bound is as follows. Consider a single block i and fix all other bits in other blocks $i' \neq i$ to be 1. By fixing the bit b appropriately, the function G computes exactly Håstad's halfspace over the k variables in block i. (We recall that Håstad's k-variable halfspace F over a k-bit input string $f \in \{-1,1\}^k$ is $F(f) = \mathrm{sign}((f, \varphi_{\alpha_j}))$ where j is the largest index such that $(f, \varphi_{\alpha_j}) \neq 0$; equivalently,

$$F(f) = \mathrm{sign}\left(\sum_{i=0}^{k-1} (k+1)^i (f, \varphi_{\alpha_i})\right)$$

gives an explicit representation of the halfspace F.) So applying Håstad's weight lower bound for his halfspace F, intuitively the variables in block i should require integer weights growing as $k^{\Omega(k)}$. Since higher blocks dominate lower blocks in G and there are $r = (n-1)/k$ blocks, intuitively a $k^{\Omega(k)}$ growth factor within each of $(n-1)/k$ blocks means that overall the weights should grow as $(k^{\Omega(k)})^{(n-1)/k} = k^{\Omega(n)}$.

Unfortunately, this simple reasoning is not quite right when applied to the actual weights w^i_j of the input variables $f^i(j)$. This is because in Håstad's halfspace all integer coefficients must be large but they do not actually increase by much; in fact, the integer coefficients of all k variables in Håstad's function can be taken to be within a factor of 2 of each other. But as we shall see the reasoning of the previous paragraph is essentially correct when a different representation is used, namely when it is applied to the "Fourier transformed" weights v^i_j that are the coefficients of $(f^i, \varphi_{\alpha_j})$ (see Equation (10) below), and this suffices to give the desired overall weight bound. We will show that the v^i_j's must grow very rapidly, and hence some v^i_j must be large, and consequently some w^i_j must be large.

The analysis. See Appendix D for the actual analysis and proof of Theorem 4.

4. PRELIMINARIES ON APPROXIMATING HALFSPACES OVER HAMMING BALLS

Let f be a halfspace over a domain S. We say that f has an *ε-approximator of weight W over S* if there is an integer vector $(v_1,\ldots,v_n) \in \mathbb{Z}^n$ with $\max_i |v_i| \leq W$ and a threshold $\theta \in \mathbb{R}$ such that

$$\Pr_{x \in S}[\mathrm{sign}(v \cdot x - \theta) \neq f(x)] \leq \varepsilon,$$

where the probability is with respect to a uniform choice of x from S. In the rest of this paper we prove upper and lower bounds on the weight of ε-approximators over the Hamming ball $\{0,1\}^n_{\leq k}$, where k is viewed as "small" compared to n.

Related work. In [3] it was shown that for any fixed $p \in (0,1)$ and any halfspace f over $\{0,1\}^n$, there is an ε-approximating halfspace $\mathrm{sign}(w \cdot x - \theta)$ of weight $n \cdot 2^{\tilde{O}_p(1/\varepsilon^2)}$ for f with respect to the product distribution $\mathcal{D}_p$, i.e.

$$\Pr_{x \sim \mathcal{D}_p}[\mathrm{sign}(w \cdot x - \theta) \neq f(x)] \leq \varepsilon.$$

Here the distribution $\mathcal{D}_p$ is the product distribution over $\{0,1\}^n$ such that each coordinate x_i of a draw from $\mathcal{D}_p$ is independently set to be 1 with probability p. The "$\tilde{O}_p$" in the exponent of the weight bound hides a dependence on p.

For constant $p \in (0, 1/2)$ the distribution $\mathcal{D}_p$ is somewhat similar to the uniform distribution on $\{0,1\}^n_{\leq pn}$ since both distributions are spread much of their weight equally over strings of weight pn. In contrast, we give upper and lower bounds that depend only on k and ε, independent of n, but our bounds require that k be "small" relative to n. Thus the main difference seems to be that the [3] results may be viewed as addressing the case where k is "large" (linear in n) while our results may be viewed as addressing the case where k is "small."

In proving our upper and lower bounds it will often be simpler for us to work with "nice" distributions which are close to the uniform distribution over $\{0,1\}^n_{\leq k}$. In Appendix E we prove some simple observations which we will use in the following sections:

OBSERVATION 1. *Let $\mathcal{D}$ denote the uniform distribution over $\{0,1\}^n_{\leq k}$ and let $\mathcal{D}_1$ denote the uniform distribution over $\{0,1\}^n_{=k}$, the set of all strings with exactly k ones. The total variation distance $\|\mathcal{D} - \mathcal{D}_1\|_1$ between $\mathcal{D}$ and $\mathcal{D}_1$ is at most $4k/n$.*

Moreover, let $\mathcal{D}_2$ denote the distribution over $[k]^n$ defined as follows: a draw of $x \sim \mathcal{D}_2$ is obtained by taking x to be $e_{i_1} + e_{i_2} + \cdots + e_{i_k}$ where each of $i_1,\ldots,i_k$ is drawn independently and uniformly from $[n]$. Then the total variation distance $\|\mathcal{D} - \mathcal{D}_2\|_1$ is at most $(k^2+4k)/n$.

We close this section with the following notation which will be useful later. Let $Z_{n,k}$ denote the set

$$Z_{n,k} = \{x = (x_1,\ldots,x_n) \in \mathbb{Z}^n : x_i \geq 0 \text{ for all } i \text{ and } x_1 + \cdots + x_n = k\}.$$

Let $\Phi : [n]^k \to Z_{n,k}$ denote the mapping $\Phi(a) = \sum_{i=1}^k e_{a_i}$. Thus a draw of $x \sim \mathcal{D}_2$ is obtained by drawing a uniformly from $[n]^k$ and setting $x = \Phi(a)$.

5. UPPER BOUND FOR APPROXIMATING HALFSPACES

In this section we prove our main positive result on approximating halfspaces over $\{0,1\}^n_{\leq k}$ using small weights, which is the following:

THEOREM 5. *Let f be any halfspace over $\{0,1\}^n_{\leq k}$. Let ε, k satisfy $\frac{k^2}{n} \leq c\varepsilon$ where $c > 0$ is a (small) universal constant. Then there is an ε-approximator for f over $\{0,1\}^n_{\leq k}$ that has weight $k^{O(k/\varepsilon)}$.*

As noted in the introduction, it is easy to see that there are halfspaces over the entire Boolean cube $\{0,1\}^n$ that require weight $\Omega(\sqrt{n})$ for ε-approximation even when ε is (say)

1/5; an example of such a halfspace is $\mathrm{sign}(x_1 + x_2 + \cdots + x_{n-1} + nx_n)$ (see [16] for the proof). In contrast, Theorem 5 shows that over Hamming balls of any constant radius, every halfspace can be approximated to any constant accuracy using weights that are independent of n.

Here is some intuition before the formal proof. The proof works by showing that every halfspace can be approximated to within $\varepsilon/2$ with respect to the distribution $\mathcal{D}_2$ (this is sufficient to establish the theorem by Observation 1). To $\varepsilon/2$-approximate an arbitrary halfspace f with respect to $\mathcal{D}_2$, the argument proceeds as follows. After sorting the weights, we first define a collection of $t = O(k/\varepsilon)$ "key coordinates" in $\{1, \ldots, n\}$ (these are just t coordinates which are evenly spaced out in $\{1, \ldots, n\}$). Then we define a set $S \subset Z_{n,k}$ of "key inputs," which are the elements of $Z_{n,k}$ that have nonzero entries only in the key coordinates. Using a linear programming argument, we show that there is a halfspace h' that depends only on the t key coordinates, has weight $k^{O(t)}$, and agrees with f on all key inputs. An additional crucial property of h' is that its weights are sorted in the same order as the weights of f. We then define an n-variable halfspace h by basing the weights of the other $n-t$ non-key coordinates in a natural way on the weights that h assigns to the key coordinates. We use the sortedness of the weights of h' to characterize the error points of h. Finally, we upper bound the error of h by using this characterization together with a simple upper bound on the edge-boundary of monotone Boolean-valued functions over the domain $[t]^k$.

Proof of Theorem 5. We first note that if $k \in \{0, 1\}$ then there is a weight-1 exact representation of f, so we henceforth assume that $k \geq 2$.

Let $w_1, \ldots, w_n, \theta'$ be a weight representation of f over $\{0,1\}^n_{\leq k}$, so $f(x) = \mathrm{sign}(w \cdot x - \theta')$ for all $x \in \{0,1\}^n_{\leq k}$. We may assume that each w_i is an integer and that θ' is of the form (integer $+1/2$). Additionally, we may assume that the weights are sorted $w_1 \leq \cdots \leq w_n$, since if this is not the case we can rename variables to make this condition hold. We use the representation w, θ' to extend the domain of f to all of $\mathbb{R}^n$, i.e. we define $f(x) = \mathrm{sign}(w \cdot x - \theta')$ for all $x \in \mathbb{R}^n$.

Key coordinates and key inputs. Let $t = O(k/\varepsilon)$. Note that if $t \geq n$ then by Theorem 1 in fact there is an exact representation for f over $\{0,1\}^n_{\leq k}$ that has weight $k^{O(k/\varepsilon)}$; thus we may assume that $t < n$. In fact, by the assumptions on ε, k and n in the statement of the theorem we may assume that $k \leq n/t$; this will be useful later.

We define the set $KC \subset [n], |KC| = t$ of "key coordinates" to be a fixed set

$$KC = \{\mathrm{key}_1 = 1, \mathrm{key}_2, \ldots, \mathrm{key}_t = n\}$$

of values in $[n]$ that are equally spaced as much as possible, i.e. for all $j, j' \in [t-1]$ we have $\mathrm{key}_{j+1} - \mathrm{key}_j = \mathrm{key}_{j'+1} - \mathrm{key}_{j'} \pm 1$.

We next define the set $KI \subset Z_{n,k}$ of "key inputs" as

$$KI = \{x \in Z_{n,k} : \text{ for all } i, \text{ if } x_i > 0 \text{ then } i \in KC\},$$

so $x \in Z_{n,k}$ is a key input if and only if all of its nonzero coordinates are key coordinates.

A low-weight halfspace h that agrees with f on all key inputs. Our next step is to establish the existence of a low-weight halfspace that depends only on the key coordinates and agrees with f on all key inputs. This is done via a linear programming argument quite similar to the proof of Theorem 1.

LEMMA 6. *There is a halfspace $h'(x) = \mathrm{sign}(v' \cdot x - \theta)$ with the following properties: (1) For each $i \notin KC$ we have $v'_i = 0$ (so h depends only on the key coordinates); (2) For each $i \in KC$ we have that v'_i is an integer satisfying $|v'_i| \leq k^{O(t)}$; (3) For each $j \in [t-1]$ we have $v'_{\mathrm{key}_j} \leq v'_{\mathrm{key}_{j+1}}$; and (4) $h'(x) = f(x)$ for every key input $x \in KI$.*

PROOF. We obtained the desired integer weights $(v'_i)_{i \in KC}$ and the threshold θ as the solution to a linear program, which we now describe. Each key input $x \in KI$ defines a linear constraint $f(x) \cdot \left(\sum_{i \in KC} v'_i x_i - \theta\right) \geq 1$ over the $t+1$ variables $(v'_i)_{i \in KC}, \theta$. The linear program additionally contains $t-1$ constraints of the form $v'_{\mathrm{key}_j} \leq v'_{\mathrm{key}_{j+1}}$ for all $j \in [t-1]$. This is a feasible linear program, since taking $v'_i = 2w_i$ for all $i \in KC$, $v'_i = 0$ for all $i \notin KC$, and $\theta = 2\theta'$ is a feasible solution. (To see that this works, observe that for any $x \in KI$ the total value of $w \cdot x$ is entirely contributed by coordinates in KC.) It is clear that any feasible solution satisfies items (1), (3) and (4) of the Lemma, so it remains only to show that there is a feasible solution satisfying the weight bound (2). This follows from the same arguments used in the proof of Theorem 1 with trivial modifications (the fact that there are now $t+1$ unknowns in the linear program leads to the claimed bound of $k^{O(t)}$ rather than $k^{O(n)}$ as was the case in Theorem 1). □

Filling in the other weights. We now define the halfspace h that has weights for all coordinates (not just the key coordinates). The halfspace h is defined as $h(x) = \mathrm{sign}(v \cdot x - \theta)$ in a very natural way as follows: for each key coordinate $i \in KC$ we take $v_i = v'_i$. For each non-key coordinate $i \notin KC$, let j be such that $\mathrm{key}_{j-1} < i < \mathrm{key}_j$, i.e. key_j is the first key coordinate immediately after i; we take $v_i = v'_{\mathrm{key}_j}$. For example, if $v' = (1/4, 0, 0, 1/3, 0, 1/2)$ then $v = (1/4, 1/3, 1/3, 1/3, 1/2, 1/2)$. Note that the weights v_i satisfy $v_1 \leq v_2 \leq \cdots \leq v_n$; this will be useful later.

We will show that this halfspace $h(x)$ is the ε-approximator for f claimed in the theorem statement. It is clear that the weight of h is at most $k^{O(t)} = k^{O(k/\varepsilon)}$ as desired; it remains to show that $\Pr_{x \sim \mathcal{D}_2}[h(x) \neq f(x)] \leq \varepsilon/2$, or equivalently, that at most an $\varepsilon/2$ fraction of points $a \in [n]^k$ have $h(\Phi(a)) \neq f(\Phi(a))$.

Bounding $\Pr_{a \in [n]^k}[h(\Phi(a)) \neq f(\Phi(a))]$. We define a function $\mathrm{up} : [n-1] \to KC$ as follows: $\mathrm{up}(i) = \mathrm{key}_j$ where key_j is the smallest element of KC satisfying $i < \mathrm{key}_j$. Similarly we define $\mathrm{down} : [n-1] \to KC$ as $\mathrm{down}(i) = \mathrm{key}_j$ where key_j is the largest element of KC satisfying $\mathrm{key}_j \leq i$. Each $i \in [n-1]$ has $\mathrm{up}(i) = \mathrm{down}(i) + 1$.

For any $a \in [n-1]^k$ we define the "upper key neighbor" of a and "downward key neighbor" of a as

$$\mathrm{ukn}(a) = (\mathrm{up}(a_1), \ldots, \mathrm{up}(a_k)) \in (KC)^k,$$
$$\mathrm{dkn}(a) = (\mathrm{down}(a_1), \ldots, \mathrm{down}(a_k)) \in (KC)^k$$

respectively. It is easy to see that for each $a \in [n-1]^k$, both $\Phi(\mathrm{ukn}(a))$ and $\Phi(\mathrm{dkn}(a))$ are key inputs. Thus Lemma 6 ensures that $\mathrm{sign}(v \cdot \Phi(\mathrm{ukn}(a)) - \theta) = \mathrm{sign}(w \cdot \Phi(\mathrm{ukn}(a)) - \theta')$ for all $a \in [n-1]^k$, and likewise for $\Phi(\mathrm{dkn}(a))$.

We next observe that by the monotonicity of the weights $v_1, \ldots, v_n$, we have that every $a \in [n-1]^k$ satisfies

$$v \cdot \Phi(\mathrm{dkn}(a)) \leq v \cdot \Phi(a) \leq v \cdot \Phi(\mathrm{ukn}(a)).$$

Consequently if $a \in [n-1]^k$ is such that $\text{sign}(v \cdot \Phi(\text{dkn}(a)) - \theta) = \text{sign}(v \cdot \Phi(\text{ukn}(a)) - \theta)$, then $\text{sign}(v \cdot \Phi(a) - \theta)$ must equal the same value, and hence for such an a we have

$$\begin{aligned}&\text{sign}(w \cdot \Phi(\text{dkn}(a)) - \theta') \\ &= \text{sign}(v \cdot \Phi(\text{dkn}(a)) - \theta) = \text{sign}(v \cdot \Phi(a) - \theta) \\ &= \text{sign}(v \cdot \Phi(\text{ukn}(a)) - \theta) = \text{sign}(w \cdot \Phi(\text{ukn}(a)) - \theta') \quad (1)\end{aligned}$$

By monotonicity of the weights $w_1, \ldots, w_n$ we have that $w \cdot \Phi(\text{dkn}(a)) \leq w \cdot \Phi(a) \leq w \cdot \Phi(\text{ukn}(a))$, so, if (1) holds, all the quantities in (1) above are also equal to $\text{sign}(w \cdot \Phi(a) - \theta')$. Thus we have shown that if $a \in [n-1]^k$ is such that $\text{sign}(v \cdot \Phi(\text{dkn}(a)) - \theta) = \text{sign}(v \cdot \Phi(\text{ukn}(a)) - \theta)$, then $\text{sign}(v \cdot \Phi(a) - \theta) = \text{sign}(w \cdot \Phi(a) - \theta')$, i.e. $h(\Phi(a)) = f(\Phi(a))$. We observe that at most a k/n fraction of all inputs $a \in [n]^k$ have $a_i = n$ for any i; by the conditions on k, ε and n in the statement of the theorem, k/n may be assumed to be at most $\varepsilon/4$. So to finish the proof of the theorem, it suffices to show the following, which we refer to as statement (*):

(*): At most an $\varepsilon/4$ fraction of all points $a \in [n-1]^k$ have $\text{sign}(v \cdot \Phi(\text{dkn}(a)) - \theta) = -1$ and $\text{sign}(v \cdot \Phi(\text{ukn}(a)) - \theta) = 1$.

We first note that for any two elements $i, j \in [t-1]$ we have $|\text{down}^{-1}(\text{key}_i)|, |\text{down}^{-1}(\text{key}_j)| \in \{\lfloor n/t \rfloor, \lfloor n/t \rfloor + 1\}$ and we recall from the bounds on t stated at the beginning of the proof that consequently $|\text{down}^{-1}(i)|, |\text{down}^{-1}(j)| \geq k$. As a result, for any two vectors $(i_1, \ldots, i_k) \in [t-1]^k$ and $(j_1, \ldots, j_k) \in [t-1]^k$, we have that the two sets

$$\{a \in [n-1]^k : \text{down}(a_\ell) = i_\ell \text{ for all } \ell = 1, \ldots, k\} \quad \text{and}$$
$$\{b \in [n-1]^k : \text{down}(b_\ell) = j_\ell \text{ for all } \ell = 1, \ldots, k\}$$

have sizes that differ by at most a multiplicative factor of $(1 + \frac{1}{k})^k < 3$. Hence to establish (*) it suffices to show that at most a $\varepsilon/12$ fraction of all vectors $(i_1, \ldots, i_k) \in [t-1]^k$ have

$$\begin{aligned}&\text{sign}(v \cdot \Phi(\text{key}_{i_1}, \ldots, \text{key}_{i_k}) - \theta) = -1 \\ &\text{and } \text{sign}(v \cdot \Phi(\text{key}_{i_1+1}, \ldots, \text{key}_{i_k+1}) - \theta) = 1.\end{aligned}$$

We define a Boolean-valued function $F : [t-1]^k \to \{-1, 1\}$ as follows:

$$F(i_1, \ldots, i_k) = \text{sign}\left(v \cdot \Phi(\text{key}_{i_1}, \ldots, \text{key}_{i_k}) - \theta\right).$$

The monotonicity of the weights $v_{\text{key}_1}, \ldots, v_{\text{key}_{t-1}}$ implies that F is a monotone non-decreasing function over $[t-1]^k$: if $r, s \in [t-1]^k$ satisfy $r_i \leq s_i$ for all $i \in [k]$ then it cannot be the case that $F(r) = 1$ and $F(s) = -1$. Now we upper bound the desired probability using a union bound:

$$\begin{aligned}&\Pr_{(i_1,\ldots,i_k)\in[t-1]^k}[F(i_1, \ldots, i_k) \neq F(i_1 + 1, \ldots, i_k + 1)] \leq \\ &\quad \Pr_{(i_1,\ldots,i_k)\in[t-1]^k}[F(i_1, \ldots, i_k) \neq F(i_1 + 1, i_2, \ldots, i_k)] + \\ &\quad \Pr_{(i_1,\ldots,i_k)\in[t-1]^k}[F(i_1 + 1, i_2, \ldots, i_k) \\ &\qquad\qquad \neq F(i_1 + 1, i_2 + 1, i_3, \ldots, i_k)] \\ &\qquad\qquad\qquad + \cdots + \\ &\quad \Pr_{(i_1,\ldots,i_k)\in[t-1]^k}[F(i_1 + 1, \ldots, i_{k-1} + 1, i_k) \\ &\qquad\qquad \neq F(i_1 + 1, \ldots, i_k + 1)].\end{aligned}$$

By the monotonicity of F, each of the k probabilities on the RHS is at most $1/(t-1)$ (since fixing all the values of the other $k-1$ coordinates, there can be at most one setting of the remaining free coordinate which causes the value of F to change). For a suitable choice of the hidden constant in $t = O(k/\varepsilon)$, we have that $1/(t-1) \leq \varepsilon/(12k)$. Thus the RHS above is at most $\varepsilon/12$ as desired. This concludes the proof of Theorem 5. □

6. LOWER BOUNDS FOR APPROXIMATING HALFSPACES

Recall that the n-variable halfspace DL is defined as

$$DL(x) = \text{sign}(\sum_{i=1}^{n} (-2)^i x_i).$$

Our main result in this section is a lower bound on the weight of any ε-approximator for DL:

THEOREM 7. *Let $k \geq 3$ and $\varepsilon \geq 4k/n$. Then any ε-approximator for DL over $\{0,1\}^n_{\leq k}$ must have weight at least $k^{\Theta(1)/\varepsilon^{1/(k-1)}-1}$.*

Discussion. It is easy to see that for all ε, the function DL has an ε-approximator over $\{0,1\}^n$ of weight $O(1/\varepsilon)$. So Theorem 7 shows that for a specific natural function, taking k to be constant and letting ε vary, getting an ε-approximator over the Hamming ball $\{0,1\}^n_{\leq k}$ (for k constant) requires weights that are *exponentially* larger than the weights required for ε-approximation over the entire Boolean cube. Theorem 7 is also in sharp contrast with the recent upper bound of [2] which shows that there is always an ε-approximator over the entire Boolean cube which has weight at most quasipoly$(1/\varepsilon)$ (as a function of ε).

6.1 Proof Sketch of Theorem 7

Since the proof of Theorem 7 is somewhat involved we give an outline here; please see Appendix F for a detailed proof. At a very high level, the idea is that in order for an LTF $\text{sign}(v \cdot x - \theta)$ to be a good approximator for DL, it should be the case that (roughly speaking) $v_i > 0$ for even i, $v_i < 0$ for odd i, and the magnitudes of the weights $|v_i|$ increase sharply with i; the essence of the proof is to show that if any of these conditions are "badly violated" then $\text{sign}(v \cdot x - \theta)$ must disagree with DL on many inputs.

In more detail, let $\text{sign}(v \cdot x - \theta)$ be an arbitrary integer weight halfspace which is a 2ε-approximator for DL with respect to $\mathcal{D}_1$ (by Observation 1 it suffices to consider such approximators). We first show (Claim 17) that without loss of generality we may assume that the threshold θ is 0 and the weights v_i are positive for even i and negative for odd i. This is not too difficult; the bulk of our work is to show that overall the magnitudes of the weights must increase significantly from smallest to largest, and thus the largest magnitude weight must be very large (since the smallest magnitude weight has magnitude at least 1). To do this, we consider the weights in order of increasing magnitude and consider disjoint "blocks" of the smallest-magnitude weights, the next-smallest-magnitude weights, and so on. We show (Lemma 18) that either there are large weights, or else almost all of the blocks are "pure," meaning that they either consist almost entirely of positive (even-index) weights, or

consist almost entirely of negative (odd-index) weights. Finally, the argument concludes by showing that if almost all of the blocks are "pure" as described above, then in fact the halfspace must err on a significant fraction of all inputs.

7. AN UPPER BOUND FOR APPROXIMATING DECISION LISTS

At this point we have established that every halfspace over $\{0,1\}^n_{\leq k}$ can be ε-approximated using weight $k^{O(k/\varepsilon)}$, and that for the DL halfspace any ε-approximator must use weight $k^{\Theta(1)/\varepsilon^{1/(k-1)}-1}$. It is a natural goal to close the gap between these upper and lower bounds; while we have not yet succeeded in doing this, in Appendix G we give a detailed analysis of the DL halfspace and prove a stronger $k^{O(k/\sqrt{\varepsilon})}$ upper bound for it. This tells us that if the $k^{O(k/\varepsilon)}$ upper bound of Theorem 5 is in fact the "right answer," then any lower bound proof establishing this must use a halfspace other than DL.

THEOREM 8. *Let ε, k, n satisfy $\varepsilon = \omega(k^2/n)$. Then there is an ε-approximator for the function DL over $\{0,1\}^n_{\leq k}$ that has weight $k^{O(k/\sqrt{\varepsilon})}$.*

8. CONCLUSION

We have studied exact and approximate representations of halfspaces over the Hamming ball $\{0,1\}^n_{\leq k}$, giving upper and lower bounds on the weight of such representations. While our upper and lower bounds are fairly close, there are still several open questions that naturally suggest themselves for followup work. In particular, our Theorem 5 gives a weight upper bound of $k^{O(k/\varepsilon)}$ which is independent of n but depends super-exponentially on k; we suspect that it may be possible to improve this dependence on k. Even for fixed k there is a gap between our upper bound, which is exponential in ε^{-1}, and our lower bound, which is exponential in $\varepsilon^{-1/(k-1)}$. It would be interesting to close this gap.

Finally, a broader goal for future work is to explore the implications of our newly established weight bounds on the effectiveness of various margin-based learning algorithms over $\{0,1\}^n_{\leq k}$.

APPENDIX

A. PROOF OF THEOREM 1

Fix f to be any halfspace over $\{0,1\}^n_{\leq k}$. Each point x in $\{0,1\}^n_{\leq k}$ with $f(x) = y \in \{-1,1\}$ provides a linear constraint

$$y(w_1x_1 + \cdots + w_nx_n + w_{n+1}) \geq 1$$

over the weights $w_1, \ldots, w_{n+1}$ which define the halfspace $f(x) = \text{sign}(w_1x_1 + \cdots + w_nx_n + w_{n+1})$. Since f is a halfspace the above system of $\sum_{j=0}^{k}\binom{n}{j}$ linear inequalities over variables $w_1, \ldots, w_{n+1}$ is feasible. A standard result in the theory of linear programming (see e.g. [10, 6]) implies that there is a subset of $n+1$ of the above inequalities which is such that if each inequality is replaced with equality, the resulting set of $n+1$ equalities defines a unique weight vector $(w_1, \ldots, w_{n+1}) \in \mathbb{R}^{n+1}$ which is a feasible solution to the entire set of $\sum_{j=0}^{k}\binom{n}{j}$ inequalities. In other words, there is a representation $\text{sign}(w \cdot x + w_{n+1})$ computing f where $(w_1, \ldots, w_{n+1}) \in \mathbb{R}^{n+1}$ is the solution to a linear system

$$Aw = b$$

where $b \in \{-1,1\}^{n+1}$ and A is an $(n+1)\times(n+1)$ 0/1 matrix in which the first n entries of each row have at most k ones and the last entry is 1. Let $\frac{\det(A_i)}{\det(A)}$ be the expression for a solution w_i using Cramer's rule. Since scaling the components of w by the same constant factor does not affect the behavior of f, setting each $w_i = \det(A_i)$ also works. Fix an arbitrary i, let $B = A_i$, and let $B_1, ..., B_{n+1}$ be the columns of B. Hadamard's inequality gives $\det(B) \leq \prod_{j=1}^{n+1} \|B_j\|$, where $\|B_j\|$ denotes the 2-norm of the B_j viewed as a vector in $\mathbb{R}^{n+1}$. Let ℓ_j be the number of nonzero components in B_j; since these components are all ± 1, we have $\|B_j\| = \sqrt{\ell_j}$, so that $\det(B) \leq \prod_{j=1}^{n+1}\sqrt{\ell_j}$. Each row of B has at most $k+2$ nonzero components, so $\sum_{j=1}^{n+1}\ell_j \leq (k+2)(n+1)$. Since $\prod_{j=1}^{n+1}\sqrt{\ell_j}$ is maximized subject to $\sum_{j=1}^{n+1}\ell_j \leq (k+2)(n+1)$ when $\ell_j = k+2$ for all j, we have $w_i = \det(B) \leq (k+2)^{(n+1)/2}$. So f can be realized using integer weights of at most this magnitude.

B. PROOF OF THEOREM 2

Since f is a halfspace over $\{0,1\}^n_{\leq 2}$, it has some representation as $f(x) = \text{sign}(w \cdot x - \theta)$ where $w_1, \ldots, w_n, \theta$ are real numbers. We will use this representation to construct an integer-weight representation $\text{sign}(v \cdot x - \theta')$ that agrees with f on all points in $\{0,1\}^n_{\leq 2}$ and where each $|v_i| \leq O(n)$.

By negating f if necessary (which does not change the integer weight required for a representation) we may assume that $f(0^n) = -1$. This means that $\text{sign}(-\theta) = -1$ and thus we have $\theta > 0$.

We may suppose without loss of generality that $w_1 < \cdots < w_n$ and all n weights $w_1, \ldots, w_n$ are nonzero (since if the weights do not satisfy these conditions they can be reordered and perturbed to satisfy them). We note that if $w_n < 0$ then every input $x \in \{0,1\}^n_{\leq 2}$ (and indeed every input in $\{0,1\}^n$) has $w \cdot x \leq 0 < \theta$; in this case f is the constant-(-1) function and f trivially has a representation of weight 0. Thus we assume going forth that $w_n > 0$.

Let $\ell \in \{1, \ldots, n\}$ be such that $w_{\ell-1} < 0 < w_\ell$ (so $\ell = 1$ if $w_1 > 0$). Now,

- Let $w' \in \mathbb{R}^{n+1}$ be $(w_1, ..., w_{\ell-1}, 0, w_\ell, ..., w_n)$.
- For each $x \in \{0,1\}^n_{=2}$, let
$$x' \in \{0,1\}^{n+1}_{=2} = (x_1, ..., x_{\ell-1}, 0, x_\ell, ..., x_n).$$
- For each $x \in \{0,1\}^n_{=1}$, let
$$x' \in \{0,1\}^{n+1}_{=2} = (x_1, ..., x_{\ell-1}, 1, x_\ell, ..., x_n).$$
- When $x = 0^n$, let $x' \in \{0,1\}^{n+1}$ be $(0, \ldots, 0)$.

Note that, for all $x \in \{0,1\}^n_{\leq 2}$, $\text{sign}(w' \cdot x' - \theta) = \text{sign}(w \cdot x - \theta)$, and, for all x except 0^n, x' has exactly two ones. Furthermore, if we have a weight vector $v \in \mathbb{R}^{n+1}$ such that $v_\ell = 0$, if we define $\hat{v} \in \mathbb{R}^n$ by $\hat{v} = (v_1, ..., v_{\ell-1}, v_{\ell+1}, v_{n+1})$, then, for all $x \in \{0,1\}^n_{\leq 2}$ and all real θ, we have $\text{sign}(v \cdot x' - \theta) = \text{sign}(\hat{v} \cdot x - \theta)$. So, our problem reduces to the problem of finding a vector $v \in \mathbb{R}^{n+1}$ with small integer weights for which $v_\ell = 0$ and there is a θ' such that

$$\text{sign}(v \cdot x - \theta') = \text{sign}(w' \cdot x - \theta)$$

for all $x \in \{0^{n+1}\} \cup \{0,1\}^{n+1}_{=2}$.

Now let us define an $(n+1) \times (n+1)$ matrix

$$(M(i,j))_{i,j \in \{1,\ldots,n+1\}}$$

with entries in $\{-1,1\}$ as follows. The matrix M will be symmetric, i.e. $M(i,j) = M(j,i)$. It will also be monotone increasing within each row and column, i.e. for each value i, the string $M(i,1)\ldots M(i,n+1)$ will be of the form $(-1)^r(1)^{n+1-r}$ for some $r \in \{0,\ldots,n+1\}$. Here is how M is defined:

- For $\{i,j\} \subset \{1,\ldots,n+1\}$ we have $M(i,j) = 1$ if and only if $\mathrm{sign}(w' \cdot (e_i + e_j)) = 1$.
- Define $M(\ell,\ell) = -1$ (recall that $f(0^n) = -1$), and define the other diagonal values, $M(i,i)$ for $i \neq \ell$, as follows. For $i > 1$ simply set $M(i,i)$ equal to $M(i,i-1)$. For $i = 1$ set $M(1,1)$ equal to $M(1,2)$.

For example, if

$$w = (-3, -5/2, 1, 4/3, 6, 7), \theta = 1/2 \tag{2}$$

then

$$w' = (-3, -5/2, 0, 1, 4/3, 6, 7)$$

and

$$M = \begin{pmatrix} -1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

It is easy to check that, in general, the matrix M is indeed symmetric. By virtue of the fact that $w_1 < \cdots < w_{\ell-1} < 0 < w_{\ell+1} < \cdots < w_n$, we have that M is monotone increasing within each row and column. Finally, the construction of M ensures that it faithfully reflects the structure of f over $\{0,1\}^n_{\leq 2}$, in the following sense. Suppose we can define weights $v_1 \leq ... \leq v_{n+1}$ and a threshold θ' with $v_\ell = 0$ such that

$$M(i,j) = 1 \quad \text{if and only if} \quad v_i + v_j \geq \theta'. \tag{3}$$

Then the halfspace

$$\mathrm{sign}(v_1x_1 + \cdots + v_{\ell-1}x_{\ell-1} + v_{\ell+1}x_{\ell+1} + \cdots + v_{n+1}x_{n+1} - \theta')$$

correctly computes f over $\{0,1\}^{n+1}_{\leq 2}$, and therefore correctly solves the original problem. (In fact (3) is stronger than what is needed – all of the correct classifications are already enforced by the off-diagonal elements, with the exception of 0^n, whose correct classification is enforced by the constraint associated with $M(\ell,\ell)$.)

In the rest of the proof we will construct the desired $v_1,\ldots,v_{n+1}$ satisfying (3) where

$$v_1 \leq \cdots \leq v_{\ell-1} \leq v_\ell = 0 \leq v_{\ell+1} \leq \cdots \leq v_{n+1}, \tag{4}$$

each v_i is an integer, and each v_i satisfies $|v_i| \leq O(n)$.

Going forth the following notation will be useful: we write M_i to denote the i-th row of M, which we view as an $(n+1)$-character string $M(i,1)\ldots M(i,n+1)$ over the alphabet $\{-1,1\}$, and is, of course, the same as the i-th column of M.

We may assume that M is not the $(n+1) \times (n+1)$ identically-(-1) matrix (since if it is then f is the constant-(-1) function over $\{0,1\}^n_{=2}$), so in particular the bottom right entry $M(n+1,n+1)$ equals 1. On the other hand, we know that the $M(\ell,\ell)$ entry is -1. Since M is monotone increasing within each row and column, and is symmetric, the only way that M could have all its rows $M_1 = \cdots = M_{n+1}$ equal to each other is if M were either the identically $+1$ or identically -1 matrix. Since M is neither of these matrices, there are at least two distinct rows in M.

The weights $v_1,\ldots,v_{n+1}$ are constructed in a greedy stage-wise fashion that we now describe. We partition the set $\{1,\ldots,n+1\}$ into $2 \leq A \leq n+1$ intervals $I_1,\ldots,I_A$ in the following way. The interval I_1 is $\{1,\ldots,i_1\}$ where i_1 is such that rows $1,\ldots,i_1$ satisfy $M_1 = \cdots = M_{i_1} \neq M_{i_1+1}$. Let j_1 be such that $M_1 = \cdots = M_{i_1} = (-1)^{n+1-j_1}1^{j_1}$. Then interval I_2 is $\{i_1+1,\ldots,i_2\}$ where similarly i_2 is such that $M_{i_1+1} = \cdots = M_{i_2} \neq M_{i_2+1}$. As before j_2 denotes the value such that $M_{i_1+1} = \cdots = M_{i_2} = (-1)^{n+1-j_2}1^{j_2}$ (note that $j_2 > j_1$). Continuing in this way we get intervals $I_1,\ldots,I_A$ and values $0 \leq j_1 < \cdots < j_A \leq n+1$, where the right endpoint of I_A is $n+1$. If $a < b$ are both in the same interval I_i then our construction will assign the same weight to v_a and v_b.

Returning to the example shown in (2), we have

- $I_1 = \{1,2\}$, $I_2 = \{3\}$, $I_3 = \{4,5\}$, $I_4 = \{6,7\}$, and
- $i_1 = 2$, $i_2 = 3$, $i_3 = 5$, $i_4 = 7$, and
- $j_1 = 2$, $j_2 = 4$, $j_3 = 5$, $j_4 = 7$.

Fix any index $i \in \{1,\ldots,A\}$ and consider any element $a \in I_i$. We have that $M(a, n+2-j_i) = 1$ while $M(a, n+1-j_i) = -1$. The idea of our construction is that we will maintain

$$v_a + v_{n+2-j_i} = C \text{ and } v_a + v_{n+1-j_i} = C - 1, \tag{5}$$

where C is a fixed integer (the same across all $i \in \{1,\ldots,A\}$ and all $a \in I_i$). Together with (4) this ensures that (3) holds as required, taking $\theta' = C - 1/2$.

In the first stage we set the weights "at the ends" and in subsequent stages we "work our way in toward the middle." More precisely, in the first stage we start with the first interval I_1. We take $v_1 = \cdots = v_{i_1} = \alpha$, and consequently to obey the description in the previous paragraph v_{n+2-j_1} must equal $C - \alpha$ (in fact we set all of v_{n+2-j_1} through v_{n+1} to equal $C - \alpha$) and v_{n+1-j_1} must equal $C - 1 - \alpha$ (we will explain how α is set below). Moving on to the second interval I_2, we set $v_{i_1+1} = \cdots = v_{i_2} = \alpha + 1$, and we set all of v_{n+2-j_2} through v_{n+1-j_1} to $C - (\alpha+1)$ (note that this setting of v_{n+1-j_1} is consistent with the way it was set when we were dealing with the first interval I_1).

During the first stage, the set of indices that get their weights set to $C - \alpha$ is exactly I_A. Each of their columns has a 1 in the first row of M, and, since M is monotone, each of their columns has all 1's, and so they all have the same columns. Since M is symmetric, they also have the same rows. However, $n+1-j_1$ is not in I_A, because column $n+1-j_1$ has a -1 in the first row. Similarly, in the second round (if there is a second round), the indices in I_{A-1} are exactly the indices whose values are set to $C - \alpha - 1$.

From here let us divide our analysis into cases, depending on whether A is even or odd.

If A is even, after $A/2$ iterations, all of the weights have been determined, and to ensure that the second of the constraints of (5) holds when $i = A/2$, we need

$$\alpha + A/2 = (C - \alpha - A/2) - 1. \tag{6}$$

If A is odd, in iteration number $(A+1)/2$, we want to set the weights of $I_{(A+1)/2}$ both to $\alpha + (A+1)/2$ and $C - (\alpha + (A+1)/2)$, so

$$\alpha + (A+1)/2 = (C - \alpha - (A+1)/2),$$

which is equivalent to (6). Recall that we also need $v_\ell = 0$. Let t be the index for which $\ell \in I_t$. Since $v_\ell = \alpha + (t-1)$, we need $\alpha = -(t-1)$, and then setting $C = A - 2t - 1$ satisfies all of the constraints.

We have constructed $v_1, \ldots, v_{n+1}$ that satisfy (4) where each v_i is an integer. It follows easily from the construction that each $|v_i|$ is at most $O(n)$, and the theorem is proved.

C. ANALYSIS OF DECISION LISTS

Counting arguments. We first show that straightforward counting arguments do not give good bounds. Let $N(n,k)$ denote $\sum_{i=0}^{k} \binom{n}{i}$, the number of points in $\{-1,1\}^n_{\leq k}$ (note that $N_k \leq (en/k)^k$). Since the VC dimension of halfspaces over $\{-1,1\}^n_{\leq k}$ is known to be $n+1$, the Sauer-Shelah lemma [15, 17] says that there are at most

$$\sum_{j=0}^{n+1} \binom{N_k}{j} \leq \sum_{j=0}^{n+1} \binom{(en/k)^k}{j} \leq \left(\frac{e(en/k)^k}{n+1}\right)^{n+1}$$

halfspaces over $\{-1,1\}^n_{\leq k}$. A standard counting argument says that if there are more than C^n halfspaces over a given domain $S \subseteq \{-1,1\}^n_{\leq k}$, then some halfspace over S must require integer weight $\Omega(C)$. So the strongest weight lower bound that can be obtained from this kind of counting argument is $O((en/k)^k/n)$. This is actually quite weak; we will see that much stronger lower bounds can be obtained for explicit functions.

Proof of Observation 3. Let $\mathrm{sign}(v \cdot x - \theta)$ be a representation of DL over $\{0,1\}^n_{\leq 3}$. As noted in the preliminaries we may assume θ is of the form (integer $+\frac{1}{2}$) so its magnitude is at least $1/2$.

Since $DL(0^n) = \mathrm{sign}(0 - \theta)$ is $+1$ we have that $\theta < 0$. Writing each v_i as $w_i\theta$ we may divide through by $|\theta|$ and re-express $\mathrm{sign}(v \cdot x - \theta)$ as $\mathrm{sign}(w \cdot x + 1)$. Here the w_i's may not be integers, but since $|\theta| \geq 1/2$ it suffices to show that $|w_n| = 2^{\Omega(n)}$.

Since $DL(e_j) = -1$ for j odd we have $w_j < -1$ for j odd, and since $DL(e_{k-1} + e_k) = 1$ for k even we have $w_k \geq -w_{k-1} - 1$ and thus $w_k > 0$ for even k. For even $k \geq 4$, since $DL(e_k + e_{k-1} + e_{k-3}) = 1$ we have

$$|w_k| = w_k \geq -w_{k-1} - w_{k-3} - 1 = |w_{k-1}| + |w_{k-3}| - 1, \tag{7}$$
$$\text{for even } k \geq 4.$$

For odd $k \geq 5$, since $DL(e_k + e_{k-1} + e_{k-3}) = -1$ we have $w_k < -w_{k-1} - w_{k-3} - 1$, and since w_k is negative for odd k and positive for even k this means

$$|w_k| > |w_{k-1}| + |w_{k-3}| + 1 \qquad \text{for odd } k \geq 5 \tag{8}$$

An easy induction using the inequalities (7) and (8) and the initial condition $w_j < -1$ for j odd gives that $|w_n| = 2^{\Omega(n)}$. □

D. PROOF OF THEOREM 4

Our goal is to prove the following which immediately gives Theorem 4. (Throughout this section β denotes the constant $\log_2(3/2)$.)

THEOREM 9. *If*

$$\text{for all } f \in \{-1,1\}^n_{\geq n-2k-1},$$
$$G(f) = \mathrm{sign}\left(\sum_{i=1}^{r}\sum_{j=0}^{k-1} w_j^i f^i(j) + w_0 b - \theta\right) \tag{9}$$

where each w_j^i and w_0 is an integer, then for some $j \in \{0, \ldots, k-1\}$ we have

$$w_j^r \geq (e^{-4k^\beta} 2^{(k \log k)/2-k})^{(n-1)/k}/k.$$

Following [6], the main step is to prove the following:

THEOREM 10. *Suppose that*

$$G(f) = \mathrm{sign}\left(\sum_{i=1}^{r}\sum_{j=0}^{k-1} v_j^i (f^i, \varphi_{\alpha_j}) + v_0 b - t\right)$$
$$\text{for all } f = (b, f^1, \ldots, f^r) \in \{-1,1\}^n_{\geq n-2k-1} \tag{10}$$

where each v_j^i and v_0 is an integer. Then

$$v_{k-1}^r \geq (e^{-4k^\beta} 2^{(k \log k)/2-k})^{(n-1)/k}.$$

To show that Theorem 10 implies Theorem 9 we use the following claim which is a simple consequence of Fourier analysis (see Lemma 2.3 of [6]):

CLAIM 11. *For any $f \in \{-1,1\}^k$ and any $(w_0, \ldots, w_{k-1}) \in \mathbb{R}^k$, setting*

$$v_a = \frac{1}{k}\sum_{j=0}^{k-1} w_j \varphi_{\alpha_a}(j) \quad \text{for each } a \in \{0, \ldots, k-1\},$$

we have that $\sum_{j=0}^{k-1} w_j f(j) = \sum_{a=0}^{k-1} v_a (f, \varphi_{\alpha_a})$

Proof of Theorem 9 using Theorem 10: Suppose that $\{w_j^i\}, w_0, \theta$ satisfy (9). By Claim 11, for all

$$f \in \{-1,1\}^n_{\geq n-2k-1}$$

we have that

$$G(f) = \mathrm{sign}\left(\sum_{i=1}^{r}\sum_{a=0}^{k-1} v_a^i (f^i, \varphi_{\alpha_a}) + w_0 b - \theta\right)$$

where $v_a^i = \frac{1}{k}\sum_{j=0}^{k-1} w_j^i \varphi_{\alpha_a}(j)$. We have that $k v_a^i$ is an integer for all i, a and so by Theorem 10 we get that $k v_{k-1}^r \geq (e^{-4k^\beta} 2^{(k \log k)/2-k})^{(n-1)/k}$, i.e.

$$\sum_{j=0}^{k-1} w_j^r \varphi_{\alpha_{k-1}}(j) \geq (e^{-4k^\beta} 2^{(k \log k)/2-k})^{(n-1)/k},$$

which gives Theorem 9 since $|\varphi_{\alpha_{k-1}}(j)| = 1$ for all j. □

D.1 Proof of Theorem 10

Throughout this section $\{v_j^i\}, v_0, t$ are as in (10). Since all weights are integers we may assume that t is of the form integer$+\frac{1}{2}$.

We begin with some straightforward claims that will be useful later.

CLAIM 12. *We have $v_0 \geq 1$. Moreover, for each $i \in [r-1]$ we have $v_0 \geq \sum_{i' \notin \{i,i+1\}, i' \in [r]} v_0^{i'} - t$.*

PROOF. First we observe that for $b \in \{-1,1\}$ we have $G(b, \varphi_{\alpha_0}, \ldots, \varphi_{\alpha_0}) = b$, which follows from the fact that $(\varphi_{\alpha_0}, \varphi_{\alpha_i}) = 0$ for all $i \neq 0$. By (10) this means that we have $\mathrm{sign}(v_0 b + \sum_{i=1}^r v_0^i - t) = b$ so it must be the case that $v_0 > 0$ and since v_0 is an integer this means $v_0 \geq 1$. Furthermore, taking $b = -1$ we find that $v_0 > \sum_{i=1}^r v_0^i - t$.

For the second part of the claim, fix any $i \in [r-1]$ and consider the input $f = (-1, f^1, \ldots, f^r)$ where $f^{i'} = \varphi_{\alpha_0}$ for $i' \notin \{i, i+1\}$ and $f^{i'} = -\varphi_{\alpha_0}$ for $i' \in \{i, i+1\}$. This input f has $2k+1$ bits that are -1 (this is the only place in the proof where we use an input with this many -1 bits) and since $G(f) = -1$ and also

$$G(f) = \mathrm{sign}\left(-v_0 + \sum_{i' \notin \{i,i+1\}, i' \in [r]} v_0^{i'} - v_0^i - v_0^{i+1} - t\right)$$

we have that $v_0 > \sum_{i' \notin \{i,i+1\}, i' \in [r]} v_0^{i'} - v_0^i - v_0^{i+1} - t$. Averaging this with the earlier inequality $v_0 > \sum_{i=1}^r v_0^i - t$ gives the second statement of the claim. □

CLAIM 13. *For every $i \in [r], j \in [k-1]$ we have $v_j^i > v_0$ (in particular, all these weights are positive).*

PROOF. Fix $i \in [r], j \in [k-1]$. For $\varepsilon, b \in \{-1,1\}$ consider the input $f = (b, f^1, \ldots, f^r) \in \{-1,1\}^n$ defined by $f^i = \varepsilon \varphi_{\alpha_j}$ and $f^{i'} = \varphi_{\alpha_0}$ for $i' \neq i$. Since every φ_{α_j} for $j \geq 1$ corresponds to the truth table of a parity function over some nonempty subset of ℓ bits, the string f has either $k/2$ or $k/2+1$ entries that are -1 (depending on whether b is $+1$ or -1). By the definition of G we have $G(f) = \mathrm{sign}(b + (k+1)^{k(i-1)+j}\varepsilon) = \varepsilon$, and, referring to (10), we have $G(f) = \mathrm{sign}(\varepsilon v_j^i + v_0 b + \sum_{i \neq i' \in [r]} v_0^{i'} - t)$. When b is $\mathrm{sign}(\sum_{i \neq i' \in [r]} v_0^{i'} - t)$ and $\epsilon = -b$ this implies that $v_j^i \geq v_0 + |\sum_{i \neq i' \in [r]} v_0^{i'} - t|$ which implies $v_j^i \geq v_0$. □

The proof uses two main lemmas. The first lemma says that weights do not get smaller as we pass from the i-th to the $(i+1)$-st block:

LEMMA 14. *For every $i \in [r-1]$ and every $j \in [k-1]$ we have $v_j^{i+1} \geq v_{k-1}^i$.*

PROOF. Fix $i \in [r-1], j \in [k-1]$. Consider the input $f = (-1, f^1, \ldots, f^r) \in \{-1,1\}^n$ defined by $f^i = -\varphi_{\alpha_{k-1}}$, $f^{i+1} = \varphi_{\alpha_j}$ and $f^{i'} = \varphi_{\alpha_0}$ for $i' \notin \{i, i+1\}$. This f has exactly $k+1$ entries that are -1 and the definition of G implies that $G(f) = 1$. So $G(f) = \mathrm{sign}(v_j^{i+1} - v_{k-1}^i + \sum_{i' \notin \{i,i+1\}, i' \in [r]} v_0^{i'} - v_0 - t) = 1$, which implies that

$$v_j^{i+1} \geq v_{k-1}^i - \sum_{i' \notin \{i,i+1\}, i' \in [r]} v_0^{i'} + v_0 + t \geq v_{k-1}^i,$$

where the final inequality follows from the second statement of Claim 12. □

The crucial lemma for us is Lemma 16, which says that the v_j^i weights grow quite significantly (by a factor of $k^{\Omega(k)}$) from the "beginning" to the "end" of each block i. Because of the way the function G has been set up we will be able to show this by a reduction to a weight lower bound that Håstad proves for his halfspace over $k = 2^\ell$ variables.

DEFINITION 15. *Let t_0 be the index of the first set in the enumeration of subsets of $[\ell]$ such that α_{t_0} has size 2.*

LEMMA 16. *For every $i \in [r]$ we have*

$$v_{k-1}^i \geq e^{-4k^\beta} 2^{(k \log k)/2 - k} \cdot v_{t_0}^i.$$

PROOF. Fix any $i \in [r]$. Consider the $(k+1)$-variable function defined as

$$\begin{aligned} &A(b, f^i) \\ &\stackrel{\mathrm{def}}{=} G(b, \varphi_{\alpha_0}, \ldots, \varphi_{\alpha_0}, f^i, \varphi_{\alpha_0}, \ldots, \varphi_{\alpha_0}) &(11)\\ &= \mathrm{sign}\left(v_0 b + \sum_{j=0}^{k-1} v_j^i (f^i, \varphi_{\alpha_j}) + \sum_{i \neq i' \in [r]} k v_0^{i'} - t\right) &(12)\end{aligned}$$

where in line (11) f^i appears in the i-th of the r blocks and all other blocks are set to φ_{α_0}. The equality (12) holds because for each $i' \neq i$ we have that $(\varphi_{\alpha_0}, \varphi_{\alpha_j})$ is 0 for $j \neq 0$ and is k for $j = 0$. For every $(b, f^i) \in \{-1,1\}^{k+1}$ the corresponding input to G in (11) has at most $k+1$ variables set to -1, so by the definition of G we have that

$$A(b, f^i) = \begin{cases} F_k(f^i) & \text{if } f \text{ is neither } \varphi_{\alpha_0} \text{ nor } -\varphi_{\alpha_0} \\ b & \text{if } f^i \text{ is either } \varphi_{\alpha_0} \text{ or } -\varphi_{\alpha_0} \end{cases} \quad (13)$$

where F_k is Håstad's function on k variables, $F_k(f^i) = \mathrm{sign}((f^i, \varphi_{\alpha_j}))$ where j is the largest index such that

$$(f, \varphi_{\alpha_j}) \neq 0.$$

Recall that since φ_{α_0} is the constant 1 function, we have $(f^i, \varphi_{\alpha_0}) = \sum_{j=0}^{k-1} f^i(j)$. Thus (13) gives us that

$$A\left(\mathrm{sign}\left(\sum_{j=0}^{k-1} f^i(j)\right), f^i\right) = F_k(f^i) \quad \text{for all } f^i \in \{-1,1\}^k.$$

Now it is clear that flipping the value of b changes the value of $A(b, f^i)$ only if f^i is either φ_{α_0} or $-\varphi_{\alpha_0}$. By (12) this implies that for all $f^i \notin \{\varphi_{\alpha_0}, -\varphi_{\alpha_0}\}$ we must have

$$|v_0| < \left|\sum_{j=0}^{k-1} v_j^i (f^i, \varphi_{\alpha_j}) + \sum_{i \neq i' \in [r]} k v_0^{i'} - t\right|.$$

But this means that the k-variable function $A' : \{-1,1\}^k \to \{-1,1\}$

$$\begin{aligned} A'(f^i) \stackrel{\mathrm{def}}{=} \mathrm{sign}\Bigg(&v_0 \left(\frac{1}{k}\sum_{j=0}^{k-1} f^i(j)\right) \\ &+ \sum_{j=0}^{k-1} v_j^i (f^i, \varphi_{\alpha_j}) + \sum_{i \neq i' \in [r]} k v_0^{i'} - t\Bigg) \end{aligned} \quad (14)$$

must equal $F_k(f^i)$ for all $f^i \in \{-1,1\}^k$, because $\frac{1}{k}\sum_{j=0}^{k-1} f^i(j)$ is always at most 1 in magnitude and equals $\mathrm{sign}\left(\sum_{j=0}^{k-1} f^i(j)\right)$

when f^i is φ_{α_0} or $-\varphi_{\alpha_0}$. Scaling the argument to sign$(\cdot)$ by a factor of k in (14), we get

$$\text{sign}\left(v_0\sum_{j=0}^{k-1} f^i(j) + k\left[\sum_{j=0}^{k-1} v_j^i(f^i,\varphi_{\alpha_j}) + \sum_{i\neq i'\in[r]} kv_0^{i'} - t\right]\right),$$

a halfspace over $\{-1,1\}^k$ that computes precisely Håstad's function F_k. As Håstad notes (Lemma 2.2 of his paper) we may remove the constant term $k(\sum_{i\neq i'\in[r]} kv_0^{i'} - t)$ without changing the function. Recalling again that $(f^i,\varphi_{\alpha_0}) = \sum_{j=0}^{k-1} f^i(j)$, we rewrite the resulting expression for $A'(f^i)$ as

$$A'(f^i) = \text{sign}\left(\sum_{j=0}^{k-1} v_j'(f^i,\varphi_{\alpha_j})\right)$$

where v_j' equals kv_j^i for $j\neq 0$ and equals $kv_0^t + v_0$ for $j=0$. Since these coefficients are all integers, we are in precisely the situation of Håstad's Theorem 2.4. The proof of that theorem explicitly establishes (see the second to last highlighted equation on p. 489) that $v'_{k-1} \geq e^{-4k^\beta} 2^{(k\log k)/2-k}\cdot v'_{t_0}$, and the lemma is proved. □

Applying Lemmas 14 and 16 repeatedly and taking j in Lemma 14 to be t_0 each time, we get that

$$v_{k-1}^r \geq (e^{-4k^\beta}2^{(k\log k)/2-k})^{(n-1)/k} v_{t_0}^1$$

which is at least $(e^{-4k^\beta}2^{(k\log k)/2-k})^{(n-1)/k}$ since $v_{t_0}^1$ is at least 1 by Claim 13. This proves Theorem 10. □

E. PROOF OF OBSERVATION 1

For the first claim, if $k > n/4$ then the claimed bound is trivially true so we assume that $k < n/4$. We recall that $\binom{n}{j-1}/\binom{n}{j} = j/(n-j+1)$, and that this is at most $1/2$ for $j \leq n/4$. So induction gives us that $\binom{n}{k-2} \leq \frac{1}{2}\binom{n}{k-1}$, $\binom{n}{k-3} \leq \frac{1}{4}\binom{n}{k-1}$, and so on, so

$$|\{0,1\}^n_{\leq k-1}| \leq \sum_{j=0}^{k-1}\frac{1}{2^j}\times|\{0,1\}^n_{=k-1}| \leq 2|\{0,1\}^n_{=k-1}|.$$

and hence

$$\frac{|\{0,1\}^n_{\leq k-1}|}{|\{0,1\}^n_{\leq k}|} \leq \frac{2|\{0,1\}^n_{=k-1}|}{|\{0,1\}^n_{\leq k}|} \leq \frac{2|\{0,1\}^n_{=k-1}|}{|\{0,1\}^n_{=k}|}$$
$$= \frac{2\binom{n}{k-1}}{\binom{n}{k}} \leq 4k/n. \tag{15}$$

So, the total variation distance between $\mathcal{D}$ and $\mathcal{D}_1$ is

$$\sum_{x:\mathcal{D}(x)>\mathcal{D}_1(x)}(\mathcal{D}(x)-\mathcal{D}_1(x))$$
$$= \frac{|\{0,1\}^n_{\leq k}| - |\{0,1\}^n_{=k}|}{|\{0,1\}^n_{\leq k}|}$$
$$= \frac{|\{0,1\}^n_{\leq k-1}|}{|\{0,1\}^n_{\leq k}|} \leq 4k/n.$$

For the second claim, let dup be the event that $x_i > 1$ for some i. We have

$$\mathcal{D}_2(\text{dup}) = \sum_{i=1}^{k-1}\frac{i}{n} \leq \frac{k(k-1)}{2n}. \tag{16}$$

Conditioned on the event $(\neg\text{dup})$, the distribution $\mathcal{D}_2$ is identical to $\mathcal{D}_1$. Thus for any event E we have

$$\begin{aligned}
&|\mathcal{D}_1(E) - \mathcal{D}_2(E)| \\
&= |\mathcal{D}_1(E) - \mathcal{D}_2(E\mid\neg\text{dup})\mathcal{D}_2(\neg\text{dup}) \\
&\qquad\qquad -\mathcal{D}_2(E\mid\text{dup})\mathcal{D}_2(\text{dup})| \\
&\leq |\mathcal{D}_1(E) - \mathcal{D}_2(E\mid\neg\text{dup})\mathcal{D}_2(\neg\text{dup})| + \mathcal{D}_2(\text{dup}) \\
&= |\mathcal{D}_1(E) - \mathcal{D}_1(E)\mathcal{D}_2(\neg\text{dup})| + \mathcal{D}_2(\text{dup}) \\
&= \mathcal{D}_1(E)\cdot(1-\mathcal{D}_2(\neg\text{dup})) + \mathcal{D}_2(\text{dup}) \\
&\leq 1 - \mathcal{D}_2(\neg\text{dup}) + \mathcal{D}_2(\text{dup}) = 2\mathcal{D}_2(\text{dup}) \\
&\leq \frac{k(k-1)}{n},
\end{aligned}$$

by (16). So $\|\mathcal{D}_2 - \mathcal{D}_1\|_1 \leq \frac{k(k-1)}{n}$ which together with the first claim and the triangle inequality for variation distance gives the desired bound.

F. PROOF OF THEOREM 7

Let $\varepsilon \geq 4k/n$, and assume that sign$(v\cdot x-\theta)$ is an integer-weight halfspace which is a 2ε-approximator for DL with respect to $\mathcal{D}_1$. (Recall that $\mathcal{D}_1$ is the uniform distribution over $\{0,1\}^n_{=k}$.) We will show that if no $|v_i|$ exceeds $k^{\Theta(1)/\varepsilon^{1/(k-1)}-1}$ then sign$(v\cdot x-\theta)$ cannot be a 2ε-approximator for DL.

We first observe that if $\varepsilon > 1000^{-k}$ then the claimed lower bound holds trivially, so we assume henceforth that $\varepsilon \leq 1000^{-k}$. Note that together with the lower bound on ε in the theorem's premises this means that we may assume $k \leq \log n$; such an upper bound on k will be useful later.

Claim 17. *We may assume without loss of generality that all of the following conditions hold:*

1. $\theta = 0$;
2. *each coordinate v_i is a nonzero integer;*
3. $v_i > 0$ *for i even and* $v_i < 0$ *for i odd.*

Proof. We first show how to obtain conditions (1) and (2) at the cost of only a multiplicative-factor increase of $\Theta(k)$ in the weights (this factor of $\Theta(k)$ corresponds to the "-1" at the end of the exponent of the weight bound of Theorem 7). Then we show how to further obtain condition (3) at the cost only of decreasing n from its original value down to some $n' \in [n/2, n]$ and of increasing ε from its original value by at most a factor of 2.

As noted in the preliminaries we may assume that θ is of the form (integer)$+1/2$. Let $u \in \mathbb{R}^n$ denote the vector $u = (1,\ldots,1)$. It is easy to verify that the halfspace sign$((2kv-2\theta u)\cdot x)$ agrees with sign$(v\cdot x-\theta)$ on every $x\in\{0,1\}^n_{=k}$, because for $x\in\{0,1\}^n_{=k}$ we have

$$(2kv-2\theta u)\cdot x = 2kv\cdot x - 2k\theta = 2k(v\cdot x-\theta).$$

Next, we observe that since $2kv_i$ is even and 2θ is odd, we have that each coordinate of $(2kv-2\theta u)$ is a nonzero integer. Thus we have achieved conditions (1) and (2) at the cost of at most a $\Theta(k)$ multiplicative factor for each weight.

So, let us suppose that sign$(v\cdot x)$ achieves conditions (1) and (2); we now deal with the signs of the weights. Let $P\subseteq[n]$ be the set of positive weights, $P \stackrel{\text{def}}{=} \{i : v_i > 0\}$, and $N = [n]\setminus P$ be the set of negative weights $N = \{i : v_i < 0\}$.

Let $E \subset [n]$ denote the set $\{2, 4, \ldots, 2\lfloor n/2 \rfloor\}$ of even indices and $O = [n] \setminus E$ denote the set of odd indices in $[n]$.

We claim that we have $|N \cap E| \leq \frac{n}{200k}$ and $|P \cap O| \leq \frac{n}{200k}$. To see why this must be true, suppose $|N \cap E| > \frac{n}{200k}$. Then there are at least $\frac{n}{200k} \cdot (\frac{n}{200k} - 1) \cdots (\frac{n}{200k} - (k-1))$ inputs $x \in \{0,1\}^n_{=k}$ of the form $x = e_{i_1} + \cdots + e_{i_k}$ where $i_1, \ldots, i_k$ are distinct and all belong to $N \cap E$. For each such x we have $v \cdot x < 0$ (because all the weights which contribute to $v \cdot x$ are negative) but $DL(x) = 1$ (because all the bits that are set to 1 in x are in even coordinates), and hence $\mathrm{sign}(v \cdot x)$ is in error on each such x. This means that $\mathrm{sign}(v \cdot x)$ has error rate at least

$$\frac{\frac{n}{200k} \cdot \left(\frac{n}{200k} - 1\right) \cdots \left(\frac{n}{200k} - (k-1)\right)}{\binom{n}{k}} > \frac{\left(\frac{n}{200k} - (k-1)\right)^k}{\binom{n}{k}}. \tag{17}$$

From our bounds on ε, k and n, the quantity (17) is $\gg 2\varepsilon$; but this contradicts the assumption that $\mathrm{sign}(v \cdot x)$ is a 2ε-approximator of f over $\{0,1\}^n_{=k}$. Thus it must indeed be the case that $|N \cap E| \leq \frac{n}{200k}$. The same argument works for $P \cap O$. Thus, we have established that indeed $|N \cap E| \leq \frac{n}{200k}$ and $|P \cap O| \leq \frac{n}{200k}$.

So an overwhelming majority of the even i lie in P and an overwhelming majority of the odd i lie in N. Let G' be defined as $G' = (P \cap E) \cup (N \cap O)$; intuitively, G' is the set of "good" indices i for which v_i has the "right" sign. The preceding paragraph gives us that $|G'| \geq (1 - \frac{1}{100k})n$.

Viewing the elements of G' as being sorted in increasing order, it may be the case that G' contains multiple consecutive even elements or multiple consecutive odd elements, i.e. we could have $G' = \{1, 3, 5, 7, 8, 10, 11, 14, \ldots\}$ and the first 4 points in G' would all belong to O. Let G be the subset of G' obtained by going through the points of G' from smallest to largest and greedily keeping the first (odd, even, odd, even, ...) points of alternating parity that we encounter (so if G' were as in the above example we would have $G' = \{1, 8, 11, 14, \ldots\}$). For a point i (like 3 in the above example) to be discarded from G', it must be the case that $i - 1$ does not belong to G'. Since at most $\frac{n}{100k}$ points do not belong to G', we have that the number of points in G' that are discarded in constructing G from G' is at most $\frac{n}{100k}$. Thus overall we have that $|G| \geq (1 - \frac{1}{50k})n$. Consequently, of the $\binom{n}{k}$ points in $\{0,1\}^n_{=k}$, at least

$$\prod_{j=0}^{k-1} \left(\left(1 - \frac{1}{50k}\right)n - j\right) > \left(\left(1 - \frac{1}{50k}\right)n - (k-1)\right)^k$$

of them are of the form $x = \sum_{j=1}^k e_{i_j}$ where all k of the distinct indices $i_1, \ldots, i_k$ belong to G. By the upper bound on k given in the statement of the theorem, this is at least half of the points in $\{0,1\}^n_{=k}$. Let us restrict the halfspace $\mathrm{sign}(v \cdot x)$ to the domain $\{0,1\}^G_{=k}$. Even if all the error points of $\mathrm{sign}(v \cdot x)$ were to lie in $\{0,1\}^G_{=k}$, since $\mathrm{sign}(v \cdot x)$ has error rate at most 2ε over $\{0,1\}^n_{=3}$, it must have error rate at most 4ε over $\{0,1\}^G_{=k}$. Moreover, since the points in G (going from smallest to largest) alternate parity (odd, even, odd, even, ...) we have that DL over the domain $\{0,1\}^G$ is completely isomorphic to DL over the domain $\{0,1\}^{|G|}$. Thus it suffices to analyze the halfspace $\mathrm{sign}(v \cdot x)$ over the domain $\{0,1\}^{|G|}_{=k}$. As claimed in the first paragraph of the proof the number of variables has gone down by at most a factor of 2 (from n to $|G|$) and the error bound has at most doubled from 2ε to 4ε, so the claim is proved. □

Using the above claim, for the rest of the proof we assume that the halfspace $\mathrm{sign}(v \cdot x)$ satisfies conditions (1)-(3). Next, as described in the overview at the start of this subsection, we divide the weights into disjoint blocks according to their magnitudes and show that almost all the blocks are "pure" (almost entirely comprised of even-indexed weights, or almost entirely comprised of odd-indexed weights).

Fix $\pi : [n] \to [n]$ to be a permutation which sorts the weights $v_1, \ldots, v_n$ in increasing order of magnitude, i.e. $0 < |v_{\pi(1)}| \leq |v_{\pi(2)}| \leq \cdots \leq |v_{\pi(n)}|$. (If the weights v_i have all distinct magnitudes then there is a unique such permutation π, and otherwise we fix any such π.) Let $b \stackrel{\text{def}}{=} \Theta(1)/\varepsilon^{1/(k-1)}$. If any weight has $|v_i| > (k/2)^{b/1000}$ then we are done, so we assume that each i has $|v_i| \leq (k/2)^{b/1000}$. We partition $[n]$ into b blocks $S_1, \ldots, S_b$ whose sizes are as nearly even as possible, i.e.

$$S_1 = \{\pi(1), \ldots, \pi(|S_1|)\}, \ldots, S_b = \{\pi(n - |S_b| + 1), \ldots, \pi(n)\}$$

where there is a fixed value $s \approx n/b$ such that $|S_i| \in \{s, s+1\}$ for all $1 \leq i \leq b$. Note that S_1 consists of the smallest-magnitude weights, S_2 consists of the next-smallest-magnitude weights, and so on.

We say that a block S_i is *pure* if at least $\frac{999}{1000}$ of the coefficients $(v_j)_{j \in S_i}$ have the same sign; equivalently, S_i is pure if at least this fraction of the elements of S_i have the same parity (almost all are even, or almost all are odd). We say that a pure block is "pure odd" ("pure even") if $\frac{999}{1000}$ of its elements are odd (even). A block which is not pure is said to be *impure*.

We have the following lemma:

LEMMA 18. *At least $\frac{998}{1000}b$ blocks are pure.*

PROOF. We introduce a different notion, that of a block being "narrow," and use this notion to prove the lemma. We show that at least $\frac{999}{1000}$ of all blocks are narrow, and that at most $\frac{1}{1000}$ of all blocks are both narrow and impure; this gives the lemma.

For a block S_j let $R_j \geq 1$ denote the ratio (largest magnitude of any weight in the block)/(smallest magnitude of any weight in the block), i.e. $R_j = |v_{\pi(i_1)}|/|v_{\pi(i_2)}|$ where $\pi(i_1), \pi(i_2) \in S_j$ and $|v_{\pi(i_1)}| \leq |v_{\pi(i')}| \leq |v_{\pi(i_2)}|$ for all $\pi(i') \in S_j$. (Note that this ratio is well defined for all $j = 1, \ldots, b$ because each weight v_i is nonzero.) We say that a block S_j is *narrow* if $R_j \leq k/2$.

We first show that at least $\frac{999}{1000}b$ blocks are narrow. Recall that $|v_{\pi(n)}| \leq (k/2)^{b/1000}$. Since $|v_{\pi(n)}| \geq |v_{\pi(n)}|/|v_{\pi(1)}| \geq \prod_{i=1}^b R_i$ it must be the case that at least $\frac{999}{1000}b$ blocks are narrow, since otherwise we would have $\prod_{i=1}^b R_i > (k/2)^{b/1000}$.

We next claim that if more than $b/1000$ blocks S_i are both narrow and impure then we have $\Pr_{x \in \{0,1\}^n_{=k}}[\mathrm{sign}(v \cdot x) \neq DL(x)] > 2\varepsilon$. To see this, fix any block ℓ that is both narrow and impure. Consider an input $x = \sum_{j=1}^k e_{i_j}$ chosen uniformly from $\{0,1\}^n_{=k}$ conditioned on $i_1, \ldots, i_k$ all belonging to S_ℓ. Some sign – either positive or negative – must constitute the majority of the largest $\frac{1}{2000}$ elements of $\{v_i\}_{i \in S_\ell}$; say that sign is positive. With probability at least $\frac{1}{4000}$ the element v_{i_k} will belong to this positive subset of the $\frac{1}{2000}$ largest elements of $\{v_i\}_{i \in S_\ell}$. On the other hand, the smallest $(1 - \frac{1}{2000})$ of the elements of $\{v_i\}_{i \in S_\ell}$ must also contain at least $\frac{1}{2000} \cdot |S_\ell|$ negative elements, (because S_ℓ is impure), and with probability $\frac{1}{2^{O(k)}}$ the elements $v_{i_1}, \ldots, v_{i_{k-1}}$ will

all belong to this set of negative elements. Thus, under the conditioning on x described above, with probability at least $1/2^{O(k)}$ we have that

$$(-1)^{i_1} = \cdots = (-1)^{i_{k-1}} \neq (-1)^{i_k}, \tag{18}$$

i.e. $i_1, \ldots, i_{k-1}$ all have the same parity (odd or even) but i_k has the opposite parity (even or odd respectively). However, since S_ℓ is narrow, the magnitude of v_{i_k} can be at most $k/2$ times the minimum magnitude of any of $v_{i_1}, \ldots, v_{i_{k-1}}$. Since $k \geq 3$, it follows that we have that $\text{sign}(v \cdot x) = (-1)^{i_1}$; but this is incorrect since $DL(x) = (-1)^{i_k}$ (because i_k is the largest value in $i_1, \ldots, i_k$). Thus, conditioned on $i_1, \ldots, i_k$ all belonging to S_ℓ, we have that x is classified incorrectly by $\text{sign}(v \cdot x)$ with probability at least $1/2^{O(k)}$. The probability (over a random $x \in \{0,1\}^n_{=k}$) that all k coordinates of x belong to S_ℓ is at least $\Theta(1)/b^k$. Assuming that at least $b/1000$ blocks are both narrow and impure, we get that overall the error rate $\Pr_{x \in \{0,1\}^n_{=k}}[\text{sign}(v \cdot x) \neq DL(x)]$ is at least

$$\frac{b}{1000} \cdot \frac{\Theta(1)}{b^k} \cdot \frac{1}{2^{O(k)}},$$

which exceeds 2ε by our choice of b.

From the above paragraph, we may conclude that at most $b/1000$ blocks S_i are both narrow and impure. Since at least $\frac{999}{1000}b$ blocks are narrow, at least $\frac{998}{1000}b$ of the b blocks are both narrow and pure, and Lemma 18 is proved. $\square$

At this point we have shown that at least 998/1000 of the b blocks are pure. Let $\text{pure}_1 < \text{pure}_2 < \cdots < \text{pure}_{b'}$ be the indices of the pure blocks, where from the above lemma we have $b' \geq \frac{998}{1000}b$. To conclude the proof we now show that if there are so many pure blocks then the error of $\text{sign}(v \cdot x)$ must exceed 2ε.

The following terminology will be useful: Given an index $\kappa \in [n-1]$ we define the "up-shift" $up(\kappa)$ to be $up(\kappa) = \kappa+1$. For a set $S \subset [n]$ we define $up(S)$ to be the set

$$up(S) = \{j+1 : j \in S\}.$$

It is clear that $|up(S)| = |S|$ for all S, and that if a ρ fraction of S is even (odd) then a ρ fraction of $up(S)$ is odd (even).

Consider any $\ell \in \{1, \ldots, b'\}$ for which S_{pure_ℓ} is a pure even block. (There are at least $\frac{49}{100}$ such ℓ's, since half of all indices are odd and half are even and 99.8% of all indices belong to a pure block.) We say that S_{pure_ℓ} is *upshift-decreasing* if at least $\frac{45}{100}$ of the elements $j \in S_{\text{pure}_\ell}$ are even and have $up(j) \in S_{\ell'}$ for some $\ell' < \text{pure}_\ell$, and we say that S_{pure_ℓ} is *upshift-increasing* if at least $\frac{45}{100}$ of the elements $j \in S_{\text{pure}_\ell}$ are even and have $up(j) \in S_{\ell'}$ for some $\ell' > \text{pure}_\ell$. Since (at least) 99.9% of the elements $j \in S_{\text{pure}_\ell}$ are even, and thus have $up(j)$ odd, at least 99.8% of the elements $j \in S_{\text{pure}_\ell}$ are even and have $up(j)$ in some block S_k with $k \neq \text{pure}_\ell$, so S_{pure_ℓ} must be either upshift-decreasing or upshift-increasing.

We consider two cases:

Case I: at least half of all pure even blocks S_{pure_ℓ} are upshift-decreasing. In this case, there are at least $\frac{49}{200}b$ pure even upshift-decreasing blocks S_{pure_ℓ}.

For S_{pure_ℓ} a pure even upshift-decreasing block, let

$$G_{\text{pure}_\ell} \subset S_{\text{pure}_\ell}$$

denote the set

$$G_{\text{pure}_\ell} = \{j \in S_{\text{pure}_\ell} : j \text{ is even and } up(j) \in S_{\ell'} \text{ for some } \ell' < \text{pure}_\ell\}$$

so $|G_{\text{pure}_\ell}| \geq \frac{4}{10} \cdot \frac{n}{b}$ (since $|S_{\text{pure}_\ell}| \approx \frac{n}{b}$). Let L_{pure_ℓ} denote the lower half of the elements in G_{pure_ℓ} and $U_{\text{pure}_\ell} = G_{\text{pure}_\ell} \setminus L_{\text{pure}_\ell}$ denote the upper half of the elements (so for every $\alpha \in L_{\text{pure}_\ell}$ and $\beta \in U_{\text{pure}_\ell}$ we have $\alpha < \beta$). We have $|L_{\text{pure}_\ell}|, |U_{\text{pure}_\ell}| \geq \frac{2}{10} \cdot \frac{n}{b}$.

Fix an ℓ such that S_{pure_ℓ} is a pure even upshift-decreasing block. Consider the set of all inputs $x = e_{i_1} + \cdots + e_{i_k} \in \{0,1\}^n_{=k}$ for which $i_1, \ldots, i_{k-1}$ all belong to L_{pure_ℓ} and i_k belongs to $up(U_{\text{pure}_\ell})$. By the cardinality bounds of the previous paragraph there are at least

$$\left(\frac{2}{10} \cdot \frac{n}{b}\right) \cdot \prod_{j=0}^{k-2} \left(\frac{2}{10} \cdot \frac{n}{b} - j\right)$$

possible such outcomes for x, so the probability that a random $x \in \{0,1\}^n_{=k}$ is of this sort is at least $\frac{1}{2^{\Theta(k)}} \cdot \frac{1}{b^k}$. For such an x we have that $v_{i_1}, v_{i_2}, \ldots, v_{i_{k-1}} > 0$ (since $i_1, \ldots, i_{k-1}$ are even), $v_{i_k} < 0$ (since i_k is odd), and $|v_{i_1}|, \ldots, |v_{i_{k-1}}| \geq |v_{i_k}|$ (since i_k belongs to S'_ℓ for some $\ell' < \text{pure}_\ell$ and

$$i_1, \ldots, i_{k-1}$$

all belong to S_{pure_ℓ}). These conditions together give that $\text{sign}(v \cdot (e_{i_1} + \cdots + e_{i_k})) = +1$. But since we have $i_k \in up(U_{\text{pure}_\ell})$ and $i_1, \ldots, i_{k-1} \in L_{\text{pure}_\ell}$, it must be the case that $i_1, \ldots, i_{k-1} < i_k$; since i_k is odd this means $DL(x) = -1$, so $\text{sign}(v \cdot x)$ is incorrect on such x. Taking a union bound across all $\frac{49}{200}b$ possibilities for ℓ that make S_{pure_ℓ} a pure even upshift-decreasing block, we get that overall

$$\Pr_{x \in \{0,1\}^n_{=k}}[\text{sign}(v \cdot x) \neq DL(x)] \geq \frac{49}{200}b \cdot \frac{1}{2^{\Theta(k)}b^k}$$

which is larger than 2ε.

We now turn to

Case II: at least half of all pure even blocks S_{pure_ℓ} are upshift-increasing, so there are at least $\frac{49}{200}b$ pure even upshift-increasing blocks S_{pure_ℓ}. Recall that in a upshift-increasing block, at least $\frac{4}{10}$ of the elements $j \in S_{\text{pure}_\ell}$ are even and have $up(j) \in S_{\ell'}$ for some $\ell' > \text{pure}_\ell$.

This analysis of this case is quite similar to Case I; the difference is that we consider a slightly different event. For S_{pure_ℓ} a pure even upshift-increasing block, let $G_{\text{pure}_\ell} \subset S_{\text{pure}_\ell}$ denote the set

$$G_{\text{pure}_\ell} = \{j \in S_{\text{pure}_\ell} : j \text{ is even and } up(j) \in S_{\ell'} \text{ for some } \ell' > \text{pure}_\ell\}$$

[1] so $|G_{\text{pure}_\ell}| \geq \frac{4}{10} \cdot \frac{n}{b}$. As before, let L_{pure_ℓ} denote the lower half of the elements in G_{pure_ℓ} and $U_{\text{pure}_\ell} = G_{\text{pure}_\ell} \setminus L_{\text{pure}_\ell}$ denote the upper half of the elements (so for every $\alpha \in L_{\text{pure}_\ell}$ and $\beta \in U_{\text{pure}_\ell}$ we have $\alpha < \beta$). As before, we have $|L_{\text{pure}_\ell}|, |U_{\text{pure}_\ell}| \geq \frac{2}{10} \cdot \frac{n}{b}$.

Fix an ℓ such that S_{pure_ℓ} is a pure even upshift-increasing block. Consider the set of all inputs $x = e_{i_1} + \cdots + e_{i_k} \in \{0,1\}^n_{=k}$ for which $i_1, \ldots, i_{k-1}$ all belong to $up(L_{\text{pure}_\ell})$ and i_k belongs to U_{pure_ℓ}. [2] As in Case I there are at least

$$\left(\frac{2}{10} \cdot \frac{n}{b}\right) \cdot \prod_{j=0}^{k-2} \left(\frac{2}{10} \cdot \frac{n}{b} - j\right)$$

possible such outcomes for x, so the probability that a random $x \in \{0,1\}^n_{=k}$ is of this sort is at least $\frac{1}{2^{\Theta(k)}} \cdot \frac{1}{b^k}$. For such

[1] Note that in Case I we had "$\ell' < \text{pure}_\ell$" where now we have "$\ell' > \text{pure}_\ell$" in the definition of G_{pure_ℓ}

[2] Note the difference from Case I.

an x we have that $v_{i_k} > 0$ (since i_k is even), $v_{i_1}, \ldots, v_{i_{k-1}} < 0$ (since $i_1, \ldots, i_{k-1}$ are all odd), and $|v_{i_1}|, \ldots, |v_{i_{k-1}}| \geq |v_{i_k}|$ (since $i_1, \ldots, i_{k-1}$ belong to S_{ℓ_1} and S_{ℓ_2} respectively for some $\ell_1, \ell_2 > \text{pure}_\ell$ whereas A belongs to S_{pure_ℓ}). These conditions together give that $\text{sign}(v \cdot (e_{i_1} + \cdots + e_{i_k})) = -1$. But since we have $i_k \in U_{\text{pure}_\ell}$ and $i_1, \ldots, i_{k-1} \in up(L_{\text{pure}_\ell})$ it must be the case that $i_k > i_1, \ldots, i_{k-1}$; since i_k is even this means $DL(x) = +1$, so $\text{sign}(v \cdot x)$ is incorrect on such an x. The rest of the argument (analyzing the probability) proceeds exactly as in Case I: taking a union bound across all $\frac{49}{200}b$ possibilities for ℓ that make S_{pure_ℓ} a pure even upshift-increasing block, we get that overall

$$\Pr_{x \in \{0,1\}^n_{=k}}[\text{sign}(v \cdot x) \neq DL(x)] \geq \frac{49}{200} b \cdot \frac{1}{2^{\Theta(k)} b^k}$$

which is larger than 2ε. We are done in Case II, and done with the proof of Theorem 7.

G. PROOF OF THEOREM 8

Recall that the obvious halfspace representation for DL as

$$\text{sign}(\sum_{i=1}^{n} (-2)^i x_i)$$

has weight 2^n. We first present a simple construction with an easy analysis that gives an $\varepsilon/2$-approximator of weight $k^{O(k/\varepsilon)}$ under distribution $\mathcal{D}_2$ (this yields an ε-approximator over $\{0,1\}^n_{\leq k}$ by Observation 1 and our choice of ε). This of course only recovers the general result of Theorem 5, but then we will sharpen this DL-specific simple construction and analysis to prove the theorem.

We assume that ε is of the form 1/integer and we define $r \stackrel{\text{def}}{=} k/\varepsilon$. Note that $r < n$ by the assumed lower bound on ε.

We partition $[n]$ into r blocks $S_1, \ldots, S_r$ whose sizes are as nearly even as possible, i.e.

$$S_1 = \{1, \ldots, |S_1|\}, \ldots, S_r = \{n - |S_r| + 1, \ldots, n\}$$

where there is a fixed value $s \approx n/r$ such that $|S_i| \in \{s, s+1\}$ for all $1 \leq i \leq r$. For $j \in [n]$ let $bl(j) \in [r]$ denote the index of the block $S_{bl(j)}$ that contains j. For $1 \leq j \leq n$ let $w_j \stackrel{\text{def}}{=} (-1)^j (2k)^{bl(j)}$. It is clear that $\max_{j \in [n]} |w_j| = (2k)^r = (2k)^{k/\varepsilon}$.

We claim that $\text{sign}(w \cdot x)$ is an $O(\varepsilon)$-approximator for $DL(x)$ over $\mathcal{D}_2$. To establish this, consider an input $x = e_{i_1} + \cdots + e_{i_k}$ drawn from $\mathcal{D}_2$, i.e. $(i_1, \ldots, i_k)$ is drawn uniformly from $[n]^k$. Let $b^\star$ denote $\max\{bl(i_1), \ldots, bl(i_k)\}$. Since the weights increase by a factor of $2k$ between successive blocks, it is easy to see that if there is precisely one index $j \in [k]$ for which $bl(i_j) = b^\star$, then $\text{sign}(w \cdot x) = (-1)^{\max\{i_1, \ldots, i_k\}}$ agrees with the value $DL(x)$. So we have that $\Pr_{x \sim \mathcal{D}_2}[\text{sign}(w \cdot x) \neq DL(x)]$ is at most the probability that there are at least two distinct indices $j_1, j_2 \in [k]$ such that $bl(i_{j_1}) = bl(i_{j_2}) = b^\star$. It is clear that for each $\ell \in [r]$, the probability that both (none of $bl(i_1), \ldots, bl(i_k)$ lie in $[\ell + 1, \ldots, r]$) and (at least two of $bl(i_1), \ldots, bl(i_k)$ equal ℓ) is at most

$$O(1) \cdot \left(\frac{\ell}{r}\right)^k \cdot \frac{k^2}{\ell^2}.$$

Summing over $\ell = 1, \ldots, r$ we get that $\Pr_{x \sim \mathcal{D}_2}[\text{sign}(w \cdot x) \neq DL(x)]$ is at most

$$\sum_{\ell=1}^{r} O(1) \cdot \frac{k^2}{r^k} \cdot \ell^{k-2} = O\left(\frac{k}{r}\right) = O(\varepsilon)$$

by our choice of $r = k/\varepsilon$. This concludes the initial simple construction and analysis.

We now build on the above simple construction to prove Theorem 8. The idea is to have the magnitude of the weights increase gradually within each block while keeping the sign of each weight correct as in the earlier construction. This lets us argue that in order for an input to be misclassified, it must have the "top two" bits that are set to 1 being quite close to each other, as well as a third input bit set to 1 that is also close to these top two. This more stringent condition lets us give a stronger bound on the probability of failure, which lets us use smaller weights to achieve an overall failure probability of ε.

We now take $r = k/\sqrt{\varepsilon}$. As before we may assume this is an integer which is less than n. We define r blocks of variables $S_1, \ldots, S_r$ and $bl(\cdot)$ as before.

We define integer weights $w_1, \ldots, w_n$ as follows. For each j the sign of w_j is $(-1)^j$. The magnitude of the weights is defined as follows: first, $|w_1| = (2k)^r$. If the first weight in block S_i (say its index is $\alpha_i + 1$) has $|w_{\alpha_i+1}| = C$, then the magnitudes of weights increase linearly in that block from C to $(2k)C$, i.e. for $j \in \{1, \ldots, |S_i|\}$ we have

$$|w_{\alpha_i + j}| = C + C \cdot \left\lceil \frac{(2k-1) \cdot j}{|S_i|} \right\rceil$$

so the final weight in block S_i has magnitude $|w_{\alpha_i + |S_i|}| = (2k)C$. If the final weight $w_{\alpha_i+|S_i|}$ of block S_i has magnitude $(2k)C$ then the first weight $w_{\alpha_i+|S_i|+1} = w_{\alpha_{i+1}+1}$ of the next block has magnitude $(4k^2)C$ (so there is a factor-of-$(2k)$ increase in the weights between each pair of successive blocks). It is clear that all weights are integers and that the largest one has magnitude $|w_n| \leq (2k)^r \cdot (2k)^{2r} = k^{O(r)}$. The halfspace we consider is $\text{sign}(w \cdot x)$.

Consider an input $x = e_{i_1} + \cdots + e_{i_k}$ drawn from $\mathcal{D}_2$, so $(i_1, \ldots, i_k)$ is drawn uniformly from $[n]^k$. As before let $b^\star$ denote $\max\{bl(i_1), \ldots, bl(i_k)\}$. As before, the only way that it is possible for $\text{sign}(w \cdot x)$ to disagree with $DL(x)$ is if there is some $\ell \in [r]$ such that both (none of $bl(i_1), \ldots, bl(i_k)$ lie in $[\ell + 1, \ldots, r]$) and (at least two of $bl(i_1), \ldots, bl(i_k)$ equal ℓ). (Our subsequent analysis will impose even more conditions that must be satisfied in order for $\text{sign}(w \cdot x)$ to be incorrect on x.)

Fix any $\ell \in [r]$. The probability that both

> none of $bl(i_1), \ldots, bl(i_k)$ lie in $[\ell + 1, \ldots, r]$

and

> at least two of $bl(i_1), \ldots, bl(i_k)$ equal ℓ

is at most k^2 times the probability that both

> none of $bl(i_1), \ldots, bl(i_k)$ lie in $[\ell + 1, \ldots, r]$

and

> $bl(i_1) = bl(i_2) = \ell$;

let us condition on this event. Let us write $i_1 = \alpha_\ell + j_1$ and $i_2 = \alpha_\ell + j_2$; we have that j_1, j_2 are selected independently

and uniformly from $\{1,\ldots,|S_\ell|\} \approx \{1,\ldots,n/r\}$. This means that $||w_{i_1}| - |w_{i_2}||$ is essentially distributed as

$$\left| w_{\alpha_\ell+1} \cdot \frac{(2k-1)\cdot(j_1-j_2)}{|S_\ell|} \right|$$

(we have omitted ceiling operators for readability; it is easy to check that this omission does not significantly affect the subsequent analysis), and consequently x is classified incorrectly only if at least one of the $k-2$ values $(|w_{i_j}|)_{j=3,\ldots,k}$ is at least $\left| w_{\alpha_\ell+1} \cdot \frac{(j_1-j_2)}{|S_\ell|} \right|$, for otherwise the cumulative effect of the other $k-2$ weights would not be large enough to offset the effect of w_{i_1} and w_{i_2}.

Let $c \in \{0,1,\ldots,\}$ be such that

$$|j_1 - j_2|/|S_\ell| \in ((2k)^{-(c+1)}, (2k)^{-c}].$$

Since every possible outcome for $|j_1 - j_2|$ (where j_1, j_2 are drawn independently from $\{1,\ldots,|S_\ell|\}$ has probability at most $O(1)/|S_\ell|$, we have that for each c the value $\Pr[|j_1 - j_2|/|S_\ell| \in ((2k)^{-(c+1)}, (2k)^{-c}]$ is at most $O((2k)^{-c})$. Because the weights increase by a factor of $2k$ between successive blocks, this means that the only way that $|w_{i_j}|$ can be at least $\left| w_{\alpha_\ell+1} \cdot \frac{(j_1-j_2)}{|S_\ell|} \right|$ is if $bl(i_j)$ belongs to $\{\ell - c - 1, \ell - c, \ldots, \ell\}$ (recall that because of our conditioning we have $bl(i_j) \leq \ell$). Because of the conditioning described earlier, for each fixed $j \in \{3,\ldots,k\}$ this occurs with probability $O(1+c)/\ell$. Taking a union bound over $k-2$ different j's, the probability that any $|w_{i_j}|$ is as large as would be necessary to cause an error is at most $O((1+c)k)/\ell$.

Putting all the pieces together and summing over all possible values $\ell = 1,\ldots,r$, we have that

$$\begin{aligned}
&\Pr_{x\sim\mathcal{D}_2}[\mathrm{sign}(w\cdot x) \neq DL(x)] \\
&\leq \sum_{\ell=1}^{r} O(1)\cdot\left(\frac{\ell}{r}\right)^k \frac{k^2}{\ell^2}\cdot\sum_{c=0}^{\infty} O((2k)^{-c})\cdot\frac{O((1+c)k)}{\ell} \\
&= O(1)\cdot\frac{k^3}{r^k}\sum_{\ell=1}^{r}\ell^{k-3}\sum_{c=0}^{\infty}\frac{1+c}{(2k)^c} \\
&= O(1)\cdot\frac{k^2}{r^2}
\end{aligned}$$

which is $O(\varepsilon)$ by our choice of r. The theorem is proved.

H. REFERENCES

[1] 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.

[2] A. De, I. Diakonikolas, V. Feldman, and R. Servedio. Near-optimal solutions for the Chow Parameters Problem and low-weight approximation of halfspaces. *STOC*, 2012.

[3] I. Diakonikolas and R. Servedio. Improved approximation of linear threshold functions. In *Proc. 24th Annual IEEE Conference on Computational Complexity (CCC)*, pages 161–172, 2009.

[4] M. Goldmann, J. Håstad, and A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.

[5] S. Hampson and D. Volper. Linear function neurons: structure and training. *Biological Cybernetics*, 53:203–217, 1986.

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

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

[8] W. Maass and G. Turan. How fast can a threshold gate learn? In *Computational Learning Theory and Natural Learning Systems: Volume I: Constraints and Prospects*, pages 381–414. MIT Press, 1994.

[9] M. Minsky and S. Papert. *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, MA, 1968.

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

[11] P. Orponen. Neural networks and complexity theory. In *Proceedings of the 17th International Symposium on Mathematical Foundations of Computer Science*, pages 50–61, 1992.

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

[13] A. Razborov. On small depth threshold circuits. In *Proceedings of the Third Scandinavian Workshop on Algorithm Theory (SWAT)*, pages 42–52, 1992.

[14] M. Saks. Slicing the hypercube. In Keith Walker, editor, *Surveys in Combinatorics 1993*, pages 211–257. London Mathematical Society Lecture Note Series 187, 1993.

[15] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (Series A)*, 13:145–147, 1972.

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

[17] S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific J. Math.*, 41:247–261, 1972.