# Every linear threshold function has a low-weight approximator

Rocco A. Servedio*
Department of Computer Science
Columbia University
New York, NY 10025
rocco@cs.columbia.edu

## Abstract

*Given any linear threshold function $f$ on $n$ Boolean variables, we construct a linear threshold function $g$ which disagrees with $f$ on at most an $\epsilon$ fraction of inputs and has integer weights each of magnitude at most $\sqrt{n} \cdot 2^{\tilde{O}(1/\epsilon^2)}$. We show that the construction is optimal in terms of its dependence on $n$ by proving a lower bound of $\Omega(\sqrt{n})$ on the weights required to approximate a particular linear threshold function.*

*We give two applications. The first is a* deterministic *algorithm for approximately counting the fraction of satisfying assignments to an instance of the zero-one knapsack problem to within an additive $\pm\epsilon$. The algorithm runs in time polynomial in $n$ (but exponential in $1/\epsilon^2$).*

*In our second application, we show that any linear threshold function $f$ is specified to within error $\epsilon$ by estimates of its Chow parameters (degree 0 and 1 Fourier coefficients) which are accurate to within an additive error of $\pm 1/(n \cdot 2^{\tilde{O}(1/\epsilon^2)})$. This is the first such accuracy bound which is inverse polynomial in $n$ (previous work of Goldberg [12] gave a $1/\mathrm{quasipoly}(n)$ bound), and gives the first polynomial bound (in terms of $n$) on the number of examples required for learning linear threshold functions in the "restricted focus of attention" framework.*

## 1. Introduction

A *linear threshold function*, or LTF, is a Boolean function $f\colon\{-1,1\}^n \to \{-1,1\}$ for which there exist $w = (w_1,\dots,w_n) \in \mathbf{R}^n$ and $\theta \in \mathbf{R}$ such that $f(x) = \mathrm{sgn}\left(\sum_{i=1}^n w_i x_i - \theta\right)$ for all $x \in \{-1,1\}^n$. Linear threshold functions (sometimes referred to in the literature as "threshold gates" or "weighted threshold gates") have been extensively studied since the 1960s [7, 20, 31] and currently play an important role in many areas of theoretical computer science. In complexity theory, complexity classes of fundamental interest such as $\mathsf{TC}^0$ are defined in terms of linear threshold functions, and much effort has been expended on understanding the computational power of single linear threshold gates and shallow circuits composed of these gates (see e.g. [13, 14, 15, 18, 34, 37]). Linear threshold functions also play a central role in computational learning theory and machine learning; many of the most widely used and successful learning techniques such as support vector machines [39], various boosting algorithms [10, 11], and fundamental algorithms such as Perceptron [4, 33] and Winnow [27, 28] are based on linear threshold functions in an essential way. Algorithms which learn linear threshold functions have also proved instrumental in the design of the fastest known learning algorithms for various expressive classes of Boolean functions (see e.g. [24, 25, 29]).

It is not hard to see that any linear threshold function $f\colon\{-1,1\}^n \to \{-1,1\}$ has some representation – in fact infinitely many – in which all the weights $w_i$ are integers. It is of considerable interest in both learning theory and complexity theory (see the references cited above) to understand how large these integer weights must be. Easy counting arguments show that most linear threshold functions over $\{-1,1\}^n$ require integer weights of magnitude $2^{\Omega(n)}$. A classic result of Muroga *et al.* [32] shows that any linear threshold function $f$ over $\{-1,1\}^n$ can be expressed using integer weights $w_1,\dots,w_n$ each satisfying $|w_i| \le 2^{O(n\log n)}$. (This result has since been rediscovered many times, see e.g. [19, 36].) Håstad [16] gave a matching lower bound by exhibiting a particular linear threshold function and proving that any integer representation for it must have weights of magnitude $2^{\Omega(n\log n)}$. Thus the size of weights that are required to (exactly) compute linear threshold functions is now rather well understood.

In this paper we are interested in the size of weights that are required to *approximately* compute linear threshold functions. Let us say that a Boolean function $g$ is an *$\epsilon$-approximator for* $f$ if $\Pr[g(x) \neq f(x)] \le \epsilon$, where the

*Supported in part by NSF CAREER award CCF-0347282, by NSF award CCF-0523664, and by a Sloan Foundation Fellowship.

probability is over a uniform choice of $x$ from $\{-1,1\}^n$. We consider the following:

**Question:** Let $f$ be an arbitrary linear threshold function. If $g$ is an LTF which $\epsilon$-approximates $f$ and has integer weights, how large do the weights of $g$ need to be?

As a first indication that the landscape can change dramatically when we switch from exact to approximate computation, consider the comparison function $COMP(x,y)$ on $2n$ bits which outputs 1 iff $x \geq y$ (viewing $x$ and $y$ as $n$-bit binary numbers). It is not hard to show that $COMP(x,y)$ is a linear threshold function which requires integer weights of magnitude $2^{\Omega(n)}$, but it is also easy to see that $COMP(x,y)$ is $\epsilon$-approximated by a linear threshold function which has only $2\log(1/\epsilon)$ many relevant variables and integer weights each at most $O(1/\epsilon)$.

### 1.1. Our results: approximating linear threshold functions using small weights.

We give a fairly complete answer to the above question. In Section 3 we prove a lower bound by exhibiting a simple linear threshold function $f$ and showing that any $\epsilon$-approximating linear threshold function for $f$ must have some weight of magnitude $\Omega(\sqrt{n})$. Perhaps surprisingly, we also show that $O(\sqrt{n})$ is an *upper* bound on the weights required to approximate any linear threshold function to any constant accuracy $\epsilon > 0$. Our main result is the following, proved in Section 4:

**Theorem 1** *Let $f\colon\{-1,1\}^n \rightarrow \{-1,1\}$ be any linear threshold function. For any $\epsilon > 0$ there is a $\epsilon$-approximating LTF $g$ with integer weights $u_1,\dots,u_n$ which satisfy $\sum_{i=1}^n u_i^2 \leq n\cdot 2^{\tilde{O}(1/\epsilon^2)}$.*

Theorem 1 immediately implies that each individual weight $u_i$ is at most $\sqrt{n}\cdot 2^{\tilde{O}(1/\epsilon^2)}$ in magnitude. It also implies that the sum of the magnitudes of all $n$ weights is at most $n \cdot 2^{\tilde{O}(1/\epsilon^2)}$.

In terms of the dependence on $\epsilon$, the "right" answer is somewhere between $(1/\epsilon)^{\omega(1)}$ (see Section 7) and our upper bound of $2^{\tilde{O}(1/\epsilon^2)}$ from Theorem 1; narrowing this gap is an interesting direction for future work.

### 1.2. Applications.

We give two main applications of Theorem 1. The first, in Section 5, is a *deterministic* algorithm for approximately counting the fraction of satisfying assignments to any linear threshold function (or equivalently, counting the number of solutions to a zero-one knapsack problem) to within additive accuracy $\pm\epsilon$. The algorithm runs in time $\tilde{O}(n^2)\cdot 2^{\tilde{O}(1/\epsilon^2)}$.

The second application is to the problem of reconstructing a linear threshold function from (approximations to) its low-degree Fourier coefficients. Various forms of this problem have been studied since the 1960s (see [6, 5, 3, 12, 23, 42]; we give a detailed description of prior work in Section 6). We show that for any constant $\epsilon > 0$, any linear threshold function $f$ is information-theoretically specified to within error $\epsilon$ by estimates of its degree-0 and degree-1 Fourier coefficients (sometimes known as its Chow parameters) which are accurate to within an additive $\pm 1/O(n)$:

**Theorem 2** *Let $f\colon\{-1,1\}^n \rightarrow \{-1,1\}$ be any linear threshold function. Let $g\colon\{-1,1\}^n \rightarrow \{-1,1\}$ be any Boolean function which satisfies $|\hat{g}(S) - \hat{f}(S)| \leq 1/(n\cdot 2^{\tilde{O}(1/\epsilon^2)})$ for each $S = \emptyset, \{1\}, \{2\}, \dots, \{n\}$. Then $\Pr[f(x) \neq g(x)] \leq \epsilon$.*

This is the first known accuracy bound which is inverse polynomial in $n$ (previous work of Goldberg [12] gave a $1/\mathrm{quasipoly}(n)$ bound). We also show a $1/\Omega(\sqrt{n/\log n})$ bound on the accuracy required. Theorem 2 directly yields the first polynomial bound (in terms of $n$) on the number of examples required for learning linear threshold functions in the "restricted focus of attention" learning framework of [2].

## 2. Preliminaries

For $v \in \mathbf{R}^n$ we write $\|v\|$ to denote $\sqrt{v_1^2 + \cdots + v_n^2}$. We write $u \cdot v$ to denote the inner product $\sum_{i=1}^n u_i v_i$ of two vectors $u, v \in \mathbf{R}^n$.

We will use standard tail bounds on sums of independent random variables, in particular the following form of the Hoeffding bound in which the deviation is bounded in terms of $\|w\|$.

**Theorem 3** *Fix any $0 \neq w \in \mathbf{R}^n$. For any $\gamma > 0$, we have*

$$\Pr_{x\in\{-1,1\}^n}[w\cdot x \geq \gamma\|w\|] \leq e^{-\gamma^2/2} \quad \textit{and}$$

$$\Pr_{x\in\{-1,1\}^n}[w\cdot x \leq -\gamma\|w\|] \leq e^{-\gamma^2/2}.$$

Another useful tool from probability theory is the following theorem, which upper bounds the probability mass that certain sums of independent random variables can have on any small region. The result can be derived from Theorem 2.14 in [35]; a short self-contained proof is given in [38].

**Theorem 4** *Fix any $w \in \mathbf{R}^n$ with $|w_i| \leq 1$ for each $i$. Then for every $\lambda \geq 1$ and $\theta \in \mathbf{R}$, we have*

$$\Pr_{x\in\{-1,1\}^n}[|w\cdot x - \theta| \leq \lambda] \leq 6\lambda/\|w\|.$$

## 3. The lower bound

In this section we exhibit a linear threshold function $f$ and show that any representation with integer weights which computes a good approximator for $f$ must have some weight of magnitude $\Omega(\sqrt{n})$.

Let $f : \{-1,1\}^{n+1} \to \{-1,1\}$ be defined as $f(x_1,\dots,x_{n+1}) = \mathrm{sgn}(x_1 + \cdots + x_n + nx_{n+1} - n)$. Note that $f(x_1,\dots,x_n,1) = \mathsf{Maj}(x_1,\dots,x_n)$ and $f(x_1,\dots,x_n,-1) = -1$ for all $x$. For convenience we assume that $n \equiv 2 \bmod 4$, but it will be clear that this assumption can be removed WLOG.

Our main result of this section is:

**Theorem 5** *Let $h : \{-1,1\}^{n+1} \to \{-1,1\}$ be any LTF which $\frac{1}{10}$-approximates $f$, and let $\mathrm{sgn}(v_1x_1 + \cdots + v_{n+1}x_{n+1} - \theta)$ be any integer representation for $h$. Then $|v_i| = \Omega(\sqrt{n})$ for some $i$.*

A straightforward application of the Hoeffding bound shows that for any $\epsilon = \Theta(1)$, there is indeed an $\epsilon$-approximating LTF $\mathrm{sgn}(x_1 + \cdots + x_n + Nx_{n+1} - N)$ for $f$ in which $N = \Theta(\sqrt{n})$.

**Proof of Theorem 5:** Let $h_1$ denote the function $h(x_1,\dots,x_n,1) = \mathrm{sgn}(v_1x_1 + \cdots + v_nx_n + v_{n+1} - \theta)$. Since $h$ is an $\frac{1}{10}$-approximator for $f$, we have $\Pr_{x_1,\dots,x_n}[h_1(x) \neq \mathsf{Maj}(x)] \le \frac{1}{5}$.

The following claim will be useful. (Stronger bounds could be given with more effort, but the $n/2$ bound is good enough for our purposes and admits a very simple proof.)

**Claim 1** *The function $h_1$ must depend on at least $n/2$ variables.*

**Proof:** Suppose $h_1$ has $r < n/2$ relevant variables; we will show that then $\Pr_{x_1,\dots,x_n}[h_1 \neq \mathsf{Maj}] > \frac{1}{5}$. For each $\ell = 1,\dots,n$ let $g_\ell : \{-1,1\}^\ell \to \{-1,1\}$ be the Boolean function on variables $x_1,\dots,x_\ell$ which is the closest approximator to $\mathsf{Maj}(x_1,\dots,x_n)$. It follows that

$$\Pr[h_1 \neq \mathsf{Maj}] \ge \Pr[g_r \neq \mathsf{Maj}] \ge \Pr[g_{n/2} \neq \mathsf{Maj}].$$

It is easy to see that each function $g_\ell$ is simply $\mathsf{Maj}(x_1,\dots,x_\ell)$. (On each input $x = (x_1,\dots,x_\ell)$, the value of $g_\ell$ is the bit $b \in \{-1,1\}$ such that the majority of the $2^{n-\ell}$ extensions $(x_1,\dots,x_n)$ of $x$ have $\mathsf{Maj}(x_1,\dots,x_n) = b$; it is easy to check that this bit $b$ is $\mathsf{Maj}(x_1,\dots,x_\ell)$.) We thus have that $\Pr[g_{n/2} \neq \mathsf{Maj}]$

$$\begin{aligned}
&= \Pr[\mathsf{Maj}(x_1,\dots,x_{n/2}) \neq \mathsf{Maj}(x_1,\dots,x_n)] \\
&\ge \Pr[\mathrm{sgn}(x_{n/2+1} + \cdots + x_n) \neq \mathrm{sgn}(x_1 + \cdots + x_{n/2}) \\
&\qquad \& \ |x_{n/2+1} + \cdots + x_n| > |x_1 + \cdots + x_{n/2}|] \\
&= \Pr[\mathrm{sgn}(x_{n/2+1} + \cdots + x_n) \neq \mathrm{sgn}(x_1 + \cdots + x_{n/2})] \cdot \\
&\qquad \Pr[|x_{n/2+1} + \cdots + x_n| > |x_1 + \cdots + x_{n/2}|] \\
&\ge (1/2)(1/2 - o(1)) > 1/5
\end{aligned}$$

where the second equality holds because the signs and magnitudes of the sums are independent (since $n/2$ is odd, each sign is achieved with probability exactly $1/2$). ■

By Claim 1 we may assume WLOG that each of $x_1,\dots,x_{n/2}$ is a relevant variable for $h_1$. Since each $v_i$ is an integer, it follows that each of $|v_1|,\dots,|v_{n/2}|$ is at least 1. Letting $v'$ denote the $n$-dimensional vector $(v_1,\dots,v_n)$, we have that $\|v'\| \ge \sqrt{n/2}$.

Since $h_1$ is a $\frac{1}{5}$-approximator to $\mathsf{Maj}(x_1,\dots,x_n)$ and $\Pr[\mathsf{Maj}(x) = 1] = \frac{1}{2} - o(1)$, we have that $\Pr_{x_1,\dots,x_n}[v_1x_1 + \cdots + v_nx_n + v_{n+1} \ge \theta] \ge 0.295$. Similarly, since $h_{-1}(x) \doteq \mathrm{sgn}(v_1x_1 + \cdots + v_nx_n - v_{n+1} - \theta)$ is a $\frac{1}{5}$-approximator to the constant function $-1$, it must be the case that $\Pr_{x_1,\dots,x_n}[v_1x_1 + \cdots + v_nx_n - v_{n+1} \ge \theta] \le 0.2$. These two inequalities imply that $v_{n+1} > 0$ and that

$$\Pr_{x_1,\dots,x_n}[|v_1x_1 + \cdots + v_nx_n - \theta| \le v_{n+1}] \ge 0.095. \qquad (1)$$

Let $v_{\max}$ denote $\max_{i=1,\dots,n} |v_i|$, let $u_i = v_i/v_{\max}$ for $i = 1,\dots,n$, and let $\lambda = v_{n+1}/v_{\max}$. Suppose first that $\lambda \ge 1$. In this case we can apply Theorem 4 to obtain

$$\begin{aligned}
0.095 \le (1) &= \Pr[|u \cdot x - \theta/v_{\max}| \le \lambda] \\
&\le \frac{6\lambda}{\|u\|} = \frac{6\lambda v_{\max}}{\|v'\|} = \frac{6v_{n+1}}{\|v'\|}
\end{aligned}$$

which implies that $v_{n+1} = \Omega(\sqrt{n})$. On the other hand, if $\lambda < 1$ then again by Theorem 4 we have

$$\begin{aligned}
0.095 \le (1) &= \Pr[|u \cdot x - \theta/v_{\max}| \le \lambda] \\
&\le \Pr[|u \cdot x - \theta/v_{\max}| \le 1] \le \frac{6v_{\max}}{\|v'\|}
\end{aligned}$$

which implies $v_{\max} = \Omega(\sqrt{n})$. So in each case some weight is $\Omega(\sqrt{n})$, and Theorem 5 is proved. ■

## 4. Proof of Theorem 1

Let $\epsilon > 0$ be given and let $f\colon \{-1,1\}^n \to \{-1,1\}$ be any linear threshold function. Without loss of generality we may suppose that $f(x) = \mathrm{sgn}\left(\sum_{i=1}^n w_ix_i - \theta\right)$ where we have $1 = |w_1| \ge |w_2| \ge \cdots \ge |w_n| > 0$.

As in the argument of [38] we consider different cases depending on the value of $\|w\|$. In each case we show how to construct an $\epsilon$-approximating LTF with integer weights that satisfy the claimed bound.

**Case I:** $\|w\| \ge \frac{12}{\epsilon}$. In this case the construction works by rounding the weights to a carefully chosen granularity. We actually prove a stronger version of Theorem 1 in this case by showing that the sum of squared weights for the $\epsilon$-approximator is at most $O(n\ln(1/\epsilon)/\epsilon^2)$ rather than $n \cdot 2^{\tilde{O}(1/\epsilon^2)}$.

Let $\alpha = \frac{\epsilon\|w\|}{6\sqrt{2n\ln(4/\epsilon)}}$. For each $i = 1, \ldots, n$ let $u_i$ be the value obtained by rounding $w_i$ to the nearest integer multiple of $\alpha$. Let $g(x) = \mathrm{sgn}(\sum_{i=1}^n u_i x_i - \theta)$, or equivalently $g(x) = \mathrm{sgn}(\sum_{i=1}^n (u_i/\alpha)x_i - \theta/\alpha)$. We will prove the following lemma:

**Lemma 2** *The linear threshold function* $g(x) = \mathrm{sgn}(\sum_{i=1}^n (u_i/\alpha)x_i - \theta/\alpha)$ *is an* $\epsilon$*-approximator for* $f$ *with integer weights each at most* $O(\sqrt{n\ln(1/\epsilon)})$ *in magnitude. Moreover, the sum of squares of weights is* $O(n\ln(1/\epsilon)/\epsilon^2)$.

**Proof:** For $i = 1, \ldots, n$ let $e_i = w_i - u_i$, so $u \cdot x = w \cdot x - e \cdot x$. Let $\lambda \geq 1$ be such that $\frac{\epsilon}{2} = \frac{6\lambda}{\|w\|}$. We have that $\mathrm{sgn}(u \cdot x - \theta) \neq \mathrm{sgn}(w \cdot x - \theta)$ only if either $|e \cdot x| \geq \lambda$ or $|w \cdot x - \theta| \leq \lambda$. We will show that each of these two events occurs with probability at most $\frac{\epsilon}{2}$ for a random $x$, and consequently $\Pr[g(x) \neq f(x)] \leq \epsilon$.

First we bound $\Pr[|e \cdot x| \geq \lambda]$. We have that $|e_i| \leq \frac{1}{2}\alpha$ for each $i$, so the vector $e = (e_1, \ldots, e_n)$ has $\|e\| \leq \frac{1}{2}\alpha\sqrt{n}$. Observing that $\lambda = \sqrt{2\ln(4/\epsilon)} \cdot \frac{1}{2}\alpha\sqrt{n}$, the Hoeffding bound (Theorem 3) gives

$$\begin{aligned} \Pr[|e \cdot x| \geq \lambda] &\leq \Pr[|e \cdot x| \geq \sqrt{2\ln(4/\epsilon)} \cdot \|e\|] \\ &\leq 2e^{-(\sqrt{2\ln(4/\epsilon)})^2/2} = \epsilon/2. \end{aligned}$$

To bound $\Pr[|w \cdot x - \theta| \leq \lambda]$ we simply apply Theorem 4; this gives us $\Pr[|w \cdot x - \theta| \leq \lambda] \leq \frac{6\lambda}{\|w\|}$, which equals $\frac{\epsilon}{2}$ by our original condition on $w$ in Case I.

Thus far we have shown that $g$ is an $\epsilon$-approximating LTF for $f$. It is clear that $g$ has a representation with integer weights each at most $1/\alpha = O(\frac{\sqrt{n\ln(1/\epsilon)}}{\|w\|\epsilon}) = O(\sqrt{n\ln(1/\epsilon)})$, where the second equality uses $\epsilon\|w\| \geq 12$. In fact we can bound the magnitude of the sum of squares of these integer weights. Let $v_i = u_i/\alpha$, so each $v_i$ is an integer and $g(x) = \mathrm{sgn}(v \cdot x - \theta/\alpha)$. Rounding each weight $w_i$ to obtain $u_i$ is easily seen to increase its magnitude by at most a factor of two. Consequently we have that each $|v_i| \leq 2|w_i|/\alpha$, and so we have

$$\begin{aligned} \sum_{i=1}^n v_i^2 &\leq 4(\sum_{i=1}^n w_i^2)/\alpha^2 = 4\|w\|^2 \cdot \frac{72n\ln(4/\epsilon)}{\epsilon^2\|w\|^2} \\ &= O(n\ln(1/\epsilon)/\epsilon^2). \end{aligned}$$

■

**Case II:** $\|w\| < \frac{12}{\epsilon}$. Note that since $|w_1| = 1$, this is equivalent to $w_1^2/(\sum_{j=1}^n w_j^2) > \epsilon^2/144$.

Let us set up some notation. We let $C_1 = 4\ln(4/\epsilon)$, $C_2 = 72\ln(2C_1/\epsilon)$, $\tau = \epsilon^2/144$, and $\ell = \frac{3}{\tau}\ln(C_2/\tau)\ln(4/\epsilon)$. Note that $\ell = \tilde{O}(1/\epsilon^2)$. We assume that $\ell \leq n$; observe that if this is not the case, then Theorem 1 follows trivially from the standard $2^{O(n\log n)}$ weight upper bound of Muroga *et al.*, since we can in fact compute $f$ exactly with weight $2^{O(n\log n)} = 2^{\tilde{O}(1/\epsilon^2)}$.

As in [38] we consider two subcases.

**Case IIa:** $w_k^2/(\sum_{j=k}^n w_j^2) > \epsilon^2/144$ for all $k = 1, \ldots, \ell$. In this case, instead of rounding the weights $w_i$ as we did in Case I, we will truncate the linear threshold function after the first $\ell$ variables and show that the resulting LTF is an $\epsilon$-approximator for $f$. Since this truncated LTF depends on only $\ell$ variables, the standard upper bound of Muroga *et al.* implies that it has an integer representation with each weight at most $2^{O(\ell\log\ell)}$ and hence sum of squared weights also $2^{O(\ell\log\ell)} = 2^{\tilde{O}(1/\epsilon^2)}$.

Let $g(x) = \mathrm{sgn}(w_1x_1 + \cdots + w_\ell x_\ell - \theta)$. Let $W = w_{\ell+1}^2 + \cdots + w_n^2$, and let $\eta = \sqrt{2W\ln(4/\epsilon)}$. We have that $g(x) \neq f(x)$ only if either $|w_{\ell+1}x_{\ell+1} + \cdots + w_nx_n| \geq \eta$ or $|w_1x_1 + \cdots + w_\ell x_\ell - \theta| \leq \eta$. We will show that these events each have probability at most $\frac{\epsilon}{2}$ and thus obtain $\Pr[g(x) \neq f(x)] \leq \epsilon$.

Bounding the first probability is easy; by our choice of $\eta$, the Hoeffding bound gives

$$\Pr[|w_{\ell+1}x_{\ell+1} + \cdots + w_nx_n| \geq \eta] \leq 2e^{-2\ln(4/\epsilon)/2} = \epsilon/2. \quad (2)$$

We now show that $\Pr[|w_1x_1 + \cdots + w_\ell x_\ell - \theta| \leq \eta] \leq \epsilon/2$. Note that since we are in Case IIa, we have $w_\ell^2 > (\epsilon^2/144)\sum_{j=\ell+1}^n w_j^2$ and thus $w_\ell > (\epsilon/12)\sqrt{W} = (\epsilon/12)(\eta/\sqrt{2\ln(4/\epsilon)})$. It therefore suffices to show that

$$\Pr\left[|w_1x_1 + \cdots + w_\ell x_\ell - \theta| \leq \frac{12}{\epsilon}w_\ell\sqrt{2\ln\frac{4}{\epsilon}}\right] \leq \frac{\epsilon}{2}. \quad (3)$$

For $i = 1, \ldots, n$ we will write $W_i$ to denote $\sum_{j=i}^n w_j^2$; note that $W_i = w_i^2 + W_{i+1}$. The following lemma will be useful:

**Lemma 3** *For* $a < b \leq \ell$*, we have*

$$W_b < (1-\tau)^{b-a}W_a < \frac{(1-\tau)^{b-a}}{\tau}w_a^2.$$

**Proof:** Since we are in Case IIa we have $w_a^2 > \tau W_a = \tau w_a^2 + \tau W_{a+1}$, or equivalently $(1-\tau)w_a^2 > \tau W_{a+1}$. Adding $(1-\tau)W_{a+1}$ to both sides gives $(1-\tau)(w_a^2 + W_{a+1}) = (1-\tau)W_a > W_{a+1}$. This implies that $W_b < (1-\tau)^{b-a}W_a$; the second inequality follows from $w_a^2 > \tau W_a$. ■

We divide the weights $w_1, \ldots, w_\ell$ into blocks of consecutive weights as follows. The first block $B_1$ is $\{w_1, \ldots, w_{k_1}\}$ where $k_1$ is the first index such that $W_{k_1+1} < w_1^2/C_2$. Similarly, the $i$-th block $B_i$ is $\{w_{k_{i-1}+1}, \ldots, w_{k_i}\}$ where $k_i$ is the first index such that $W_{k_i+1} < w_{k_{i-1}+1}^2/C_2$.

**Corollary 4** *Each block* $B_i$ *is of length at most* $\frac{1}{\tau}\ln(C_2/\tau)$.

**Proof:** By Lemma 3, the length $|B_i|$ of the $i$-th block must satisfy $1/C_2 \le (1-\tau)^{|B_i|}/\tau$; the corollary follows from this. ■

Recalling that $\ell = \frac{3}{\tau}\ln(C_2/\tau)\ln(4/\epsilon)$, we have that there are at least $3\ln(4/\epsilon)$ many blocks of weights in $w_1,\dots,w_\ell$.

Let us view the choice of a uniform $(x_1,\dots,x_\ell) \in \{-1,1\}^\ell$ as taking place in successive stages, where in the $i$-th stage the variables corresponding to the $i$-th block $B_i$ are chosen. The rest of our analysis in Case IIa will only deal with the first $\ln(4/\epsilon)$ blocks so for the rest of Case IIa we assume that $i \le \ln(4/\epsilon)$.

Immediately after the $i$-th stage, some value – call it $\xi_i$ – has been determined for $w_1x_1+\cdots+w_{k_i}x_{k_i}$. The following lemma shows that if $\xi_i$ is too far from $\theta$, then it is unlikely that the remaining variables $x_{k_i+1},\dots,x_\ell$ will come out in such a way as to make the final sum sufficiently close to $\theta$.

**Lemma 5** *If $|\xi_i-\theta| \ge 2\sqrt{W_{k_i+1}}\sqrt{2\ln(2C_1/\epsilon)}$, then we have*

$$\Pr_{x_{k_i+1},\dots,x_\ell}\left[|w_1x_1+\cdots+w_\ell x_\ell-\theta| \le \frac{12}{\epsilon}w_\ell\sqrt{2\ln\frac{4}{\epsilon}}\right] \le \epsilon/C_1.$$

**Proof:** By the lower bound on $|\xi_i-\theta|$ in the hypothesis of the lemma, it can only be the case that $|w_1x_1+\cdots+w_\ell x_\ell-\theta| \le (12/\epsilon)\sqrt{2\ln(4/\epsilon)}w_\ell$ if

$$\begin{aligned}&|w_{k_i+1}x_{k_i+1}+\cdots+w_\ell x_\ell| \\ \ge\ & 2\sqrt{W_{k_i+1}}\sqrt{2\ln\frac{2C_1}{\epsilon}}-(12/\epsilon)w_\ell\sqrt{2\ln\frac{4}{\epsilon}} \quad (4)\end{aligned}$$

Since $i \le \ln(4/\epsilon)$ and each block is of length at most $\frac{1}{\tau}\ln(C_2/\tau)$ by Corollary 4, we have that $k_i+1 \le \frac{1}{\tau}\ln(C_2/\tau)\ln(4/\epsilon)+1$. Recalling the definition of $\ell$, it follows that $(\ell-(k_i+1))/2 > \frac{1}{\tau}\ln(12/\epsilon)$. Now using Lemma 3, we have that

$$w_\ell \le \sqrt{W_\ell} \le (1-\tau)^{(\ell-(k_i+1))/2}\sqrt{W_{k_i+1}} \le \frac{\epsilon}{12}\sqrt{W_{k_i+1}}.$$

Rearranging this inequality and using $2C_1 \ge 4$, it follows that the RHS of (4) is at least $\sqrt{2\ln(2C_1/\epsilon)}\cdot\sqrt{W_{k_i+1}}$. So to prove the lemma it suffices to bound $\Pr_{x_{k_i+1},\dots,x_\ell}[|w_{k_i+1}x_{k_i+1}+\cdots+w_\ell x_\ell| \ge \sqrt{2\ln(2C_1/\epsilon)}\cdot\sqrt{W_{k_i+1}}]$ by $\epsilon/C_1$. But since $w_{k_i+1}^2+\cdots+w_\ell^2 \le W_{k_i+1}$, the Hoeffding bound implies that this probability is at most $2e^{-(\sqrt{2\ln(2C_1/\epsilon)})^2/2} = \epsilon/C_1$. ■

We now show that regardless of the value $\xi_{i-1}$ immediately *before* the $i$-th stage, immediately *after* the $i$-th stage we will have $|\xi_i-\theta| \le 2\sqrt{W_{k_i+1}}\sqrt{2\ln(2C_1/\epsilon)}$ with probability at most $1/2$ over the choice of values for variables in block $B_i$ in the $i$-th stage.

**Lemma 6** *For any $\xi_{i-1} \in$ **R**, we have* $\Pr_{x_{k_{i-1}+1},\dots,x_{k_i}}[|\xi_i-\theta| \le 2\sqrt{W_{k_i+1}}\sqrt{2\ln(2C_1/\epsilon)}] \le 1/2$.

**Proof:** Since $\xi_i$ equals $\xi_{i-1}+(w_{k_{i-1}+1}x_{k_{i-1}+1}+\cdots+w_{k_i}x_{k_i})$, we have $|\xi_i-\theta| \le 2\sqrt{W_{k_i+1}}\sqrt{2\ln(2C_1/\epsilon)}$ if and only if the value $w_{k_{i-1}+1}x_{k_{i-1}+1}+\cdots+w_{k_i}x_{k_i}$ lies in the interval $[I_L, I_R]$, where

$$I_L := \theta-\xi_{i-1}-2\sqrt{W_{k_i+1}}\sqrt{2\ln(2C_1/\epsilon)}$$

and

$$I_R := \theta-\xi_{i-1}+2\sqrt{W_{k_i+1}}\sqrt{2\ln(2C_1/\epsilon)},$$

of width $4\sqrt{W_{k_i+1}}\sqrt{2\ln(2C_1/\epsilon)}$.

First suppose that $0 \notin [I_L, I_R]$, i.e. the whole interval has the same sign. If this is the case then $\Pr[w_{k_{i-1}+1}x_{k_{i-1}+1}+\cdots+w_{k_i}x_{k_i} \in [I_L,I_R]] \le \frac{1}{2}$ since by symmetry the value $w_{k_{i-1}+1}x_{k_{i-1}+1}+\cdots+w_{k_i}x_{k_i}$ is equally likely to be positive or negative.

Now suppose that $0 \in [I_L, I_R]$. By definition of $k_i$, we know that $\sqrt{W_{k_i+1}} \le |w_{k_{i-1}+1}|/\sqrt{C_2}$, and consequently the width of $[I_L,I_R]$ is at most $4|w_{k_{i-1}+1}|\sqrt{2\ln(2C_1/\epsilon)}/\sqrt{C_2}$, which is at most $\frac{2}{3}|w_{k_{i-1}+1}|$ by the definition of $C_2$. But now observe that once the value of $x_{k_{i-1}+1}$ is set to either $+1$ or $-1$, this effectively shifts the "target interval," which now $w_{k_{i-1}+2}x_{k_{i-1}+2}+\cdots+w_{k_i}x_{k_i}$ must hit, by a displacement of $w_{k_{i-1}+1}$ to become $[I_L-w_{k_{i-1}+1}x_{k_{i-1}+1}, I_R-w_{k_{i-1}+1}x_{k_{i-1}+1}]$. Since the original interval $[I_L,I_R]$ contained 0 and was of length at most $\frac{2}{3}|w_{k_{i-1}+1}|$, the new interval does not contain 0, and thus again by symmetry we have that the probability (now over the choice of $x_{k_{i-1}+2},\dots,x_{k_i}$) that $w_{k_{i-1}+1}x_{k_{i-1}+1}+\cdots+w_{k_i}x_{k_i}$ lies in $[I_L,I_R]]$ is at most $\frac{1}{2}$. ■

In order to have $|w_1x_1+\cdots+w_\ell x_\ell-\theta| \le (12/\epsilon)\sqrt{2\ln(4/\epsilon)}w_\ell$, it must be the case that either (1) each $|\xi_i-\theta| < 2\sqrt{W_{k_i+1}}\sqrt{2\ln(2C_1/\epsilon)}$ for $i=1,\dots,\ln(4/\epsilon)$ or (2) for some $i \le \ln(4/\epsilon)$ we have $|\xi_i-\theta| \ge 2\sqrt{W_{k_i+1}}\sqrt{2\ln(2C_1/\epsilon)}$ but nonetheless $|w_1x_1+\cdots+w_\ell x_\ell-\theta| < (12/\epsilon)\sqrt{2\ln(4/\epsilon)}w_\ell$. Lemma 6 gives us that the probability of (1) is at most $(1/2)^{\ln(4/\epsilon)} = \epsilon/4$, and Lemma 5 gives us that the probability of (2) is at most $\ln(4/\epsilon)\cdot\epsilon/C_1 = \epsilon/4$. Thus the overall probability that $|w_1x_1+\cdots+w_\ell x_\ell-\theta| \le (12/\epsilon)\sqrt{2\ln(4/\epsilon)}w_\ell$ is at most $\epsilon/2$, and (3) is proved.

**Case IIb:** $w_k^2/(\sum_{j=k}^n w_j^2) \le \epsilon^2/144$ for some $k \in \{1,\dots,\ell\}$. In this case we round the weights $w_k,\dots,w_n$ to obtain an $\epsilon/2$-approximating LTF in which these weights are small integers. We then argue that this LTF is itself $\epsilon/2$-close to an LTF with all small integer weights.

We define weight vectors $u', v' \in \mathbf{R}^n$ as follows: For $i = 1, \ldots, k-1$ let $u'_i = w_i/|w_k|$. For $i = k, \ldots, n$ let $u'_i$ be the value obtained by rounding $w_i/|w_k|$ to the nearest integer multiple of $\alpha' = \frac{(\epsilon/2)\sqrt{w_k^2+\cdots+w_n^2}}{6|w_k|\sqrt{2n\ln(8/\epsilon)}}$. (Note that everywhere $\alpha$ in Case I had an $\epsilon$, now $\alpha'$ has $\epsilon/2$.) Let $v'_i = u'_i/\alpha'$ for all $i = 1, \ldots, n$. Finally let $\theta' = \theta/|w_k|$, and let $g\colon \{-1,1\}^n \to \{-1,1\}$ be the LTF $g(x) = \operatorname{sgn}(u' \cdot x - \theta')$ or equivalently $g(x) = \operatorname{sgn}(v' \cdot x - \theta'/\alpha')$.

We first show that $g$ is an $\frac{\epsilon}{2}$-approximator for $f$ which has "almost all" small integer weights.

**Lemma 7** *The linear threshold function $g(x) = \operatorname{sgn}(v' \cdot x - \theta'/\alpha')$ is an $\frac{\epsilon}{2}$-approximator for $f$. Each weight $v'_i$ for $i \geq k$ is an integer of magnitude $O(\sqrt{n\ln(1/\epsilon)})$, and we have $\sum_{i=k}^n (v'_i)^2 = O(n\ln(1/\epsilon)/\epsilon^2)$.*

**Proof:** Fix any setting $x_1^*, \ldots, x_{k-1}^*$ of the first $k-1$ bits. Let $f_*$ be the linear threshold function on $n-k+1$ variables which is obtained by fixing the first $k-1$ inputs of $f$ to $x_1^*, \ldots, x_{k-1}^*$; note that we may write $f_*(x_k, \ldots, x_n)$ as $\operatorname{sgn}(\sum_{j=k}^n (w_j/|w_k|)x_j - \theta' + \sum_{j=1}^{k-1}(w_j/|w_k|)x_j^*)$. Similarly, let $g_*$ be the LTF on $n-k+1$ variables obtained by fixing the first $k-1$ inputs of $g$ to to $x_1^*, \ldots, x_{k-1}^*$, i.e. $g_*(x_k, \ldots, x_n) = \operatorname{sgn}(\sum_{j=k}^n v'_j x_j - \theta'/\alpha' + \sum_{j=1}^{k-1} v'_j x_j^*)$.

We have that $1 = |w_k/|w_k|| \geq |w_{k+1}/|w_k|| \geq \cdots \geq |w_n/|w_k|| > 0$. Moreover, each weight $v'_i$ for $i \geq k$ is obtained from $w_i/|w_k|$ by rounding to the nearest integer multiple of $\alpha'$ (and then scaling by $\alpha'$ to get integer weights). Since the thresholds of $f_*$ and $g_*$ match up as well (taking into account the scaling by $\alpha'$), we may apply Lemma 2, and conclude that $\Pr_{x_k,\ldots,x_n}[g_* \neq f_*] \leq \frac{\epsilon}{2}$. Since this holds for every restriction $x^* \in \{-1,1\}^{k-1}$, it follows that $\Pr_{x\in\{-1,1\}^n}[g(x) \neq f(x)] \leq \frac{\epsilon}{2}$. The claimed bounds on the weights $v'_i$ for $i \geq k$ follow from Lemma 2. ■

We next show that any linear threshold function which has "almost all" its weights integers whose sum of squares is small (such as $g$) can be $\epsilon/2$-approximated by a linear threshold function with small integer weights.

**Lemma 8** *Let $g\colon \{-1,1\}^n \to \{-1,1\} : g(x) = \operatorname{sgn}(s \cdot x - \mu)$ be a linear threshold function where $s_k, s_{k+1}, \ldots, s_n$ are all integers with $\sum_{j=k}^n s_j^2 \leq N$. Then there is a linear threshold function $g'(x) = \operatorname{sgn}(t \cdot x - \nu)$ which is an $\frac{\epsilon}{2}$-approximator of $g$, where*

*(i) each $t_i$ is an integer;*

*(ii) $|t_i| \leq \sqrt{N\ln(1/\epsilon)} \cdot 2^{O(k\log k)}$ for $i \leq k-1$; and*

*(iii) $\sum_{i=1}^n t_i^2 \leq N \cdot \ln(1/\epsilon) \cdot 2^{O(k\log k)}$.*

Theorem 1 follows in Case IIb by combining Lemmas 7 and 8, recalling that $k \leq \ell = \tilde{O}(1/\epsilon^2)$ and taking $N$ in Lemma 8 to be $O(n\ln(1/\epsilon)/\epsilon^2)$.

**Proof of Lemma 8:** We first observe that by the Hoeffding bound, we have

$$\Pr_{x_k,\ldots,x_n}[|s_k x_k + \cdots + s_n x_n| > \sqrt{2\ln(4/\epsilon)}\sqrt{N}] \leq \epsilon/2.$$

Intuitively, we can thus pretend that $\sum_{j=k}^n s_k x_k$ always has magnitude at most $\sqrt{2\ln(4/\epsilon)}\sqrt{N}$ and this causes us to incur error at most $\epsilon/2$ (we will make this more precise later).

We will need the following claim:

**Claim 9** *Fix an integer $R > 0$. Let $\Omega$ denote $\{-1,1\}^{k-1} \times \{-R, -R+1, \ldots, R-1, R\}$. Let $h$ be any linear threshold function over $\Omega$, i.e. for some $w \in \mathbf{R}^k$ and $\theta \in \mathbf{R}$ we have that $h(x) = \operatorname{sgn}(w \cdot x - \theta)$ for all $x \in \Omega$. Then there is a representation of $h$ as $h(x) = \operatorname{sgn}(u \cdot x - \theta)$ in which (a) each $u_i$ is an integer, and (b) $|u_i| \leq R \cdot (k+1)!$ for $i = 1, \ldots, k-1$ and $|u_k| \leq (k+1)!$.*

This claim is an extension of Muroga *et al.*'s classic upper bound on the size of integer weights that are required to express linear threshold functions over the usual domain $\{-1,1\}^n$; we defer its proof until later.

Now the pieces are in place to prove Lemma 8. Let $R = \sqrt{2\ln(4/\epsilon)}\sqrt{N}$. Given the LTF $g(x) = \operatorname{sgn}(s \cdot x - \mu)$, let $h : \Omega \to \{-1,1\}$ be the LTF $h(x) = \operatorname{sgn}(\sum_{j=1}^{k-1} s_i x_i + x_k - \mu)$. By Claim 9, we have that over the domain $\Omega$, $h$ is equivalent to $h(x) = \operatorname{sgn}(\sum_{j=1}^k u_i x_i - \mu)$, where $u_1, \ldots, u_k$ satisfy conditions (a) and (b). Now consider $g' : \{-1,1\}^n \to \{-1,1\}$,

$$g'(x) = \operatorname{sgn}\left(\sum_{i=1}^{k-1} u_i x_i + u_k \left(\sum_{j=k}^{n} s_j x_j\right) - \mu\right).$$

By our observation at the start of the proof, at least a $1 - \frac{\epsilon}{2}$ fraction of all $x \in \{-1,1\}^n$ have $|\sum_{j=k}^n s_j x_j| \leq R$. For each such $x$ we have $g'(x) = h\left(x_1, \ldots, x_{k-1}, \sum_{j=k}^n s_j x_j\right) = g(x)$. Thus $g'$ is an $\frac{\epsilon}{2}$-approximator of $g$ with integer weights $t_1, \ldots, t_n$, where $t_i = u_i$ for $i \leq k-1$ and $t_j = u_k s_j$ for $j \geq k$. Plugging in the bounds on $u_i, u_k, s_j$ from the conditions of Lemma 8 and Claim 9, the proof of Lemma 8 is done. ■

**Proof of Claim 9:** We need only slightly modify known proofs of Muroga *et al.*'s upper bound for LTF weights over $\{-1,1\}^n$. In particular we closely follow the outline of the proof which Håstad gives in Section 3 of [16].

Let $H_0 : \mathbf{R}^k \to \mathbf{R}$ be a linear function $H_0(x) = a \cdot x + t$ which satisfies the following conditions:

1. $\operatorname{sgn}(H_0(x)) = h(x)$ for each $x \in \Omega$.

2. $|H_0(x)| \geq 1$ for each $x \in \Omega$.

3. Among all linear functions which satisfy conditions (1) and (2) above, $H_0$ maximizes the number of $x \in \Omega$ which have $|H_0(x)| = 1$. If there is more than one possible $H_0$ which achieves the maximum number, choose one arbitrarily.

Observe that since $h(x)$ is a linear threshold function over $\Omega$, there exists some linear function satisfying (1) and (2), and thus there does exist some $H_0$ satisfying (1)-(3) above.

As in [16], let $x^{(1)}, \ldots, x^{(r)}$ be the set of points in $\Omega$ with $|H_0(x^{(i)}| = 1$. The argument in [16] now directly implies that $H_0$ is uniquely determined by the equations

$$H_0(x^{(i)}) = h(x^{(i)}) \qquad \text{for } i = 1, \ldots, r.$$

Consequently the coefficients $a_1, \ldots, a_k, t$ of $H_0(x)$ can be obtained by solving a linear system of $k+1$ equations:

$$a_1 x_1^{(i)} + \cdots + a_k x_k^{(i)} + t = h(x^{(i)}) \qquad \text{for } i = 1, \ldots, k+1.$$

For each of these equations the right-hand side is $\pm 1$ as are the first $k-1$ coefficients $x_1^{(i)}, \ldots, x_{k-1}^{(i)}$ (and the coefficient of $t$), whereas the $k$-th coefficient $x_k^{(i)}$ is an integer in $\{-R, \ldots, R\}$.

Cramer's rule now tells us that for $j = 1, \ldots, k$, we have $a_j = \det(M_j)/\det(M)$ for suitable $(k+1) \times (k+1)$ matrices $M_1, \ldots, M_k, M$. More precisely, the matrix $M$ has as its $i$-th row the vector $x^{(i)}$ with a 1 appended as the $(k+1)$-st entry, and the matrix $M_j$ is $M$ but with the $j$-th column replaced by the column vector whose $i$-th entry is $h(x^{(i)})$. Since all entries of $M$ except for the $k$-th column are $\pm 1$ and each element in the $k$-th column is an integer of magnitude at most $R$, we have that $\det(M)$ is an integer of magnitude at most $(k+1)!R$, and the same is true for $\det(M_1), \ldots, \det(M_{k-1})$. The matrix $M_k$ is a $\pm 1$ matrix so it satisfies $|\det(M_k)| \leq (k+1)!$. Now since each of $a_1, \ldots, a_k$ has the same denominator we may clear it throughout and obtain a linear threshold function for $h$ whose $k$ integer weights are $\det(M_1), \ldots, \det(M_k)$. This concludes the proof of Claim 9. ■

### 4.1. Discussion and consequences for monotone formula construction.

The main result of [38] is a proof that any monotone linear threshold function $f$ can be $\epsilon$-approximated by a monotone Boolean AND/OR formula of size $n^{10.6} \cdot 2^{\tilde{O}(1/\epsilon^4)}$. The high-level structure of our proof of Theorem 1 is similar to that of [38] in that the same cases I, IIa and IIb are considered,[1] but there are some significant differences. First, in Case I of [38] the weights are simply rounded to the nearest multiple of $1/n$ rather than the nearest $\alpha = \frac{1}{O(\sqrt{n})}$ (ignoring the dependence on $\epsilon$). Second, our Case IIa is handled using a simpler argument in [38] which only yields $\ell = \tilde{O}(1/\epsilon^4)$ in [38] rather than the $\ell = \tilde{O}(1/\epsilon^2)$ we achieve here. Finally, since the goal in [38] is to construct a monotone formula rather than a low-weight linear threshold function, a different approach is used in that paper to handle Case IIb. (In particular, a recursive tree-based decomposition is used in [38] which yields a Boolean formula but not a linear threshold function.)

Inspection reveals that our new analysis of Case I and our new bound on $\ell$ can be straightforwardly worked into the arguments of [38] to obtain the following quantitative improvement of its main result:

**Corollary 10** *Let $f\colon \{-1,1\}^n \to \{-1,1\}$ be any monotone linear threshold function. There is a monotone Boolean formula of size $n^{5.3} \cdot 2^{\tilde{O}(1/\epsilon^2)}$ which is an $\epsilon$-approximator for $f$.*

(Briefly, the improvement from $n^{10.6}$ to $n^{5.3}$ comes from the fact that now in Case I, we have that the sum $\sum_{i=1}^n |v_i|$ of the integer weights of $g(x)$ is $O(n)$ rather than the $O(n^2)$ bound obtained in [38] by rounding each weight to the nearest $1/n$. This $O(n)$ is then plugged into Valiant's probabilistic construction [41] of monotone formulas of size $O(n^{5.3})$ for the majority function on $n$ variables.)

## 5. Application to deterministic approximate counting

We describe an application of our approach to the problem of approximately counting solutions of the zero-one knapsack problem. In an instance of zero-one knapsack we are given a vector $a = (a_1, \ldots, a_n) \in \mathbf{R}^n$ and a threshold $\theta \in \mathbf{R}$; the goal is to approximately compute the fraction $p$ of points $x \in \{0,1\}^n$ which satisfy the linear threshold function $\mathrm{sgn}(\sum_{i=1}^n a_i x_i - \theta)$. It is not hard to see that we may equivalently consider the domain of the LTF to be $\{-1,1\}^n$ as we have been doing throughout this paper.

The problem of efficiently computing a multiplicative $(1 \pm \epsilon)$-approximation of $p$ has received much attention [9, 21, 22]; the first polynomial-time algorithm was given by Morris and Sinclair [30] using sophisticated Monte Carlo Markov Chain techniques, and more recently a simpler randomized algorithm based on dynamic programming and "dart throwing" was given by Dyer [8].

Our techniques, combined with the dynamic programming idea of Dyer [8], give a simple *deterministic* algorithm for computing an $\epsilon$-accurate *additive* approximation of $p$. (Achieving such an additive approximation is trivial, of course, if randomization is allowed: simply make

[1] Readers familiar with [38] will note that Case IIa of this paper is Case IIb of [38] and vice versa.

$O(1/\epsilon^2)$ random draws from $\{-1,1\}^n$ and output the fraction of satisfying assignments in this sample as an approximation of $p$.) See [40] for work in a similar spirit on deterministically counting the fraction of satisfying assignments to a $k$-DNF to additive accuracy $\pm\epsilon$.

**Theorem 6** *There is a deterministic $\tilde{O}(n^2) \cdot 2^{\tilde{O}(1/\epsilon^2)}$-time algorithm with the following property: given an instance of the zero-one knapsack problem for which the true fraction of satisfying assignments in $\{-1,1\}^n$ is $p$, the algorithm outputs a value $\tilde{p}$ such that $|p - \tilde{p}| \le \epsilon$.*

**Proof:** Given $w_1, \dots, w_n, \theta$, the idea is to efficiently construct a linear threshold function $g(x)$ which $\epsilon$-approximates $f(x) = \mathrm{sgn}(w \cdot x - \theta)$ as in the proof of Theorem 1, and then use dynamic programming to *exactly* count the number of satisfying assignments to $g$.

Suppose first that $w_1, \dots, w_n$ satisfy Case I of Section 4. Then as in that section we round each weight to the nearest integer multiple of $\alpha$ and divide by $\alpha$ throughout to obtain an $\epsilon$-approximating linear threshold function $g(x) = \mathrm{sgn}(v \cdot x - \theta')$ with integer weights $v_i$ that satisfy $\sum_{i=1}^n |v_i| \le M = O(n \ln(1/\epsilon)/\epsilon^2)$. Let $F(r,s) = |\{x \in \{-1,1\}^r : \sum_{i=1}^r v_i x_i = s\}|$. We can compute $F(r,s)$ for all $1 \le r \le n, -M \le s \le M$ in $O(nM)$ time with dynamic programming, using the initial condition $F(0,0) = 1$ and the relation $F(r+1,s) = F(r, s-v_{r+1}) + F(r, s+v_{r+1})$. The number of satisfying assignments to $g$ is $\sum_{s \ge \theta'} F(n,s)$.

Now suppose that $w_1, \dots, w_n$ satisfy Case IIa. We now take $g(x) = \mathrm{sgn}(w_1 x_1 + \dots + w_\ell x_\ell - \theta)$ to be the truncated LTF analyzed in Case IIb. While the weights $w_1, \dots, w_\ell$ and partial sums $\sum_{i=1}^r w_i x_i$ may not be integers, we can still perform dynamic progamming using the observation that for any $r \le \ell$ there are at most $2^r$ real values $\rho$ such that $F(r,\rho) = |\{x \in \{-1,1\}^r : \sum_{j=1}^r w_i x_i = \rho\}|$ is nonzero. Thus we can compute $F(r,\rho)$ for all $\rho$ which have $F(r,\rho) \ne 0$ as before in overall time $2^{O(\ell)} = 2^{\tilde{O}(1/\epsilon^2)}$.

Finally suppose that $w_1, \dots, w_n$ satisfy Case IIb. In this case we use the linear threshold function $g(x) = \mathrm{sgn}(v' \cdot x - \theta'/\alpha')$ described in Lemma 7. Since $g$ has at most $k-1 \le \ell$ weights which are not integers and the integer weights have total magnitude bounded by $M = O(n \ln(1/\epsilon)/\epsilon^2)$, we now have that for any $1 \le r \le n$ there are at most $O(2^\ell M)$ real values $\rho$ such that $F(r,\rho) = |\{x \in \{-1,1\}^r : \sum_{j=1}^r w_i x_i = \rho\}|$ is nonzero. So we can compute $F(r,\rho)$ for all $\rho$ which have $F(r,\rho) \ne 0$ as before in overall time $\tilde{O}(n 2^\ell M) = \tilde{O}(n^2) \cdot 2^{\tilde{O}(1/\epsilon^2)}$. (The extra log factor comes from sorting the $F(r,\rho)$ values in order of increasing $\rho$ once they have all been computed for each $r$, and performing binary search over this sorted list in the next stage to find each $F(r, \rho \pm v'_{r+1})$ value as required.) ∎

# 6. Approximating an LTF from noisy versions of its low-degree Fourier coefficients

Recall that for a Boolean function $f\colon \{-1,1\}^n \to \{-1,1\}$, the *Fourier coefficients* $\{\hat{f}(S)\}_{S \subseteq [n]}$ of $f$ are the coefficients of the (unique) multilinear polynomial

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S) x_S \qquad \text{where } x_S \text{ denotes } \prod_{i \in S} x_i$$

which agrees with $f$ everywhere on $\{-1,1\}^n$. The *degree* of a Fourier coefficient $\hat{f}(S)$ is the degree $|S|$ of of the corresponding monomial.

In 1961 Chow [6] proved that every linear threshold function is uniquely specified (among all Boolean functions) by its $n+1$ Fourier coefficients of degree 0 and 1; these coefficients are sometimes referred to as the *Chow parameters* of $f$. Following this result (which was later generalized by Bruck [5]), there has been interest in how to algorithmically obtain a weights-based representation $f(x) = \mathrm{sgn}(w \cdot x - \theta)$ of $f$ from its Chow parameters, see e.g. [23, 42]. This seems to be a difficult problem, and we do not address it here.

A related question which has also been studied is the following: suppose we are given noisy rather than exact values of the Chow parameters. How does this affect the precision with which $f$ is (information-theoretically) specified by these parameters? One motivation for studying this question comes from the "1-restricted focus of attention" model in computational learning theory; roughly speaking this is a learning model in which the learner is only allowed to see a single bit $x_i$ of each example $x = (x_1, \dots, x_n)$ used for learning (see [2, 1] for details). As observed by [3, 12], the class of linear threshold functions over $\{-1,1\}^n$ is uniform-distribution information-theoretically learnable from $\mathrm{poly}(n)$ many examples in this framework if and only if any linear threshold function is information-theoretically specified to high accuracy from Chow parameter estimates which are accurate to an additive $\pm 1/\mathrm{poly}(n)$.

With this motivation Birkendorf *et al.* gave the following result:

**Theorem 7 ([3])** *Let $f(x) = \mathrm{sgn}(w_1 x_1 + \dots + w_n x_n - \theta)$ be a linear threshold function with integer weights $w_i$ such that $W = \sum_{i=1}^n |w_i|$. Let $g\colon \{-1,1\}^n \to \{-1,1\}$ be any Boolean function which satisfies $|\hat{g}(S) - \hat{f}(S)| \le \frac{\epsilon}{W}$ for each $S = \emptyset, \{1\}, \{2\}, \dots, \{n\}$. Then $\Pr[f(x) \ne g(x)] \le \epsilon$.*

Theorem 7 gives a strong bound on the precision required in the Chow parameters if $f$ has low weight, but a weak bound for arbitrary LTFs since $W$ may need to be $2^{\Omega(n \log n)}$. Subsequently Goldberg [12] gave an incomparable result which can be rephrased as follows:

**Theorem 8 ([12])** *Let $f$ be any linear threshold function, and let $g\colon \{-1,1\}^n \to \{-1,1\}$ be any Boolean function which satisfies $|\hat{g}(S) - \hat{f}(S)| \leq (\epsilon/n)^{O(\log(n/\epsilon)\log(1/\epsilon))}$ for each $S = \emptyset, \{1\}, \{2\}, \ldots, \{n\}$. Then $\Pr[f(x) \neq g(x)] \leq \epsilon$.*

In contrast, our bound in Theorem 2 has a worse dependence on $\epsilon$ but has a $1/n$ rather than $1/\mathrm{quasipoly}(n)$ dependence on $n$. Theorem 2 yields an affirmative answer (at least for constant $\epsilon$) to the open question of whether arbitrary linear threshold functions can be learned in the uniform distribution 1-RFA model with polynomial sample complexity:

**Corollary 11** *Fix any constant $\epsilon > 0$. There is an algorithm for learning arbitrary linear threshold functions to accuracy $\epsilon$ under the uniform distribution in the 1-restricted focus of attention model, using $\mathrm{poly}(n)$ many examples.*

We prove Theorem 2 in Appendix A.

### 6.1. Lower bounds on required accuracy for Chow parameter estimation.

In this section we sketch a simple (though somewhat indirect) argument which shows that no variant of Theorem 2 in which the bound on $|\hat{g}(S) - \hat{f}(S)|$ is $1/o(\sqrt{n/\log n})$ (as a function of $n$) can be true.

Suppose to the contrary that Theorem 2 held with a bound of the form $1/(o(\sqrt{n/\log n}) \cdot \kappa(\epsilon))$ for some function $\kappa$ that depends only on $\epsilon$. If we fix $\epsilon$ to be a constant such as $1/10$, the bound is simply $1/o(\sqrt{n/\log n})$. Recall that the Fourier coefficient $\hat{f}(S)$ equals $\mathbf{E}[f(x)\chi_S]$ where $\chi_S$ is the Fourier basis function corresponding to $S$. A standard application of the Hoeffding bound shows that a sample of $o(n)$ many uniform random labelled examples $(x, f(x))$ suffices to yield $n+1$ estimates $\alpha(S)$ which all satisfy $|\alpha(S) - \hat{f}(S)| < 1/o(\sqrt{n/\log n})$ with high probability. (The number of examples required is $o(n)$ rather than $o(n/\log n)$ because we need each of the $n+1$ estimates to be correct with high probability.) Now we can do a brute-force search over all Boolean functions to find some $g\colon \{-1,1\}^n \to \{-1,1\}$ whose Chow parameters are all within the desired $1/o(\sqrt{n/\log n})$ additive accuracy of our estimates $\alpha(S)$. The search will eventually find such a function since $f$ is such a function, and by the assumed version of Theorem 2, the $g$ thus obtained will be a $\frac{1}{10}$-approximator of $f$.

Thus, we have seen that the hypothesized Theorem 2 variant implies that there is an algorithm which can learn any linear threshold function $f$ to accuracy $\epsilon = 1/10$, using uniform random examples only, from $o(n)$ examples. However, this contradicts known sample complexity lower bounds in computational learning theory; for instance the results of [26] can be easily used to show that any algorithm which learns linear threshold functions to constant accuracy using uniform random examples over $\{-1,1\}^n$ must use $\Omega(n)$ examples.

## 7. Conclusion

We hope that Theorem 1 may find a range of applications in future work. In computational learning theory, low-weight linear threshold functions are known to be "nice" in several senses; our results suggest that similar properties might sometimes hold for arbitrary linear threshold functions as well. As one example, simple and efficient algorithms are known which can learn low-weight linear threshold functions under noise rates at which no efficient algorithms are known for learning arbitrary linear threshold functions. Can our results (which can be viewed as stating that every linear threshold function is "close to" a low-weight linear threshold function) be used to learn arbitrary linear threshold functions in the presence of higher noise rates?

More concretely, an obvious direction for future work is to improve the asymptotic dependence on $\epsilon$ in our results. As [12] and [38] have observed, Håstad's construction of a linear threshold function which requires integer weights of size $2^{\Omega(n\log n)}$ implies that in general an $\epsilon$-approximating LTF for an arbitrary LTF $f$ may require integer weights of size $(1/\epsilon)^{\Omega(\log\log(1/\epsilon))}$. While this means that it is impossible to obtain an analogue of Theorem 1 with a $\mathrm{poly}(1/\epsilon)$ dependence on $\epsilon$, it may well be possible to improve the current $2^{\tilde{O}(1/\epsilon^2)}$ dependence.

Another goal is to obtain stronger bounds on the accuracy which is required in the Chow parameters in order to specify an arbitrary linear threshold function $f$ to accuracy $\epsilon$. Can the gap between our $1/O(n)$ bound and the $1/\Omega(\sqrt{n/\log n})$ bound given in Section 6.1 be closed?

**Acknowledgments.** We thank Johan Håstad for his kind permission to include the proof of Theorem 9 in this paper. We also thank Adam Klivans for many stimulating discussions on these topics without which this paper would never have been written.

## References

[1] S. Ben-David and E. Dichterman. Learnability with restricted focus of attention guarantees noise-tolerance. In *Proceedings of the 5th International Workshop on Algorithmic Learning Theory*, pages 248–259, 1994.

[2] S. Ben-David and E. Dichterman. Learning with restricted focus of attention. *Journal of Computer and System Sciences*, 56(3):277–298, 1998.

[3] A. Birkendorf, E. Dichterman, J. Jackson, N. Klasner, and H. Simon. On restricted-focus-of-attention learnability of boolean functions. *Machine Learning*, 30:89–123, 1998.

[4] H. Block. The Perceptron: a model for brain functioning. *Reviews of Modern Physics*, 34:123–135, 1962.

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

[6] C. Chow. On the characterization of threshold functions. In *Proceedings of the Symposium on Switching Circuit Theory and Logical Design*, pages 34–38, 1961.

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

[8] M. Dyer. Approximate counting by dynamic programming. In *Proceedings of the 35th Annual Symposium on Theory of Computing (STOC)*, pages 693–699, 2003.

[9] M. Dyer, A. Frieze, K. Rannan, A. Kapoor, L. Perkovic, and U. Vazirani. A mildly exponential time algorithm for approximating the number of solutions to a multidimensional knapsack problem. *Combinatorics, Probability and Computing*, 2:271–284, 1993.

[10] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995.

[11] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[12] P. Goldberg. Estimating a boolean perceptron from its average satisfying assignment: A bound on the precision required. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory*, pages 116–127, 2001.

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

[14] M. Goldmann and M. Karpinski. Simulating threshold circuits by majority circuits. *SIAM Journal on Computing*, 27(1):230–246, 1998.

[15] A. Hajnal, W. Maass, P. Pudlak, M. Szegedy, and G. Turan. Threshold circuits of bounded depth. *Journal of Computer and System Sciences*, 46:129–154, 1993.

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

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

[18] T. Hofmeister. A note on the simulation of exponential threshold weights. In *Computing and Combinatorics, Second Annual International Conference (COCOON)*, pages 136–141, 1996.

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

[20] S. Hu. *Threshold Logic*. University of California Press, 1965.

[21] M. Jerrum and A. Sinclair. *The Markov Chain Monte Carlo method: an approach to approximate counting and integration*, pages 482–520. PWS Publishing, 1997.

[22] R. Kannan. Markov chains and polynomial time algorithms. In *Proceedings of the 35th Symposium on Foundations of Computer Science*, pages 656–671, 1994.

[23] P. Kaszerman. A gemoetric test-synthesis procedure for a threshold device. *Information and Control*, 6(4):381–398, 1963.

[24] A. Klivans, R. O'Donnell, and R. Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer & System Sciences*, 68(4):808–840, 2004. Preliminary version in *Proc. of FOCS'02*.

[25] A. Klivans and R. Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. In *Proceedings of the Thirty-Third Annual Symposium on Theory of Computing*, pages 258–265, 2001.

[26] S. Kulkarni, S. Mitter, and J. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11:23–35, 1993.

[27] N. Littlestone. Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[28] N. Littlestone. Redundant noisy attributes, attribute errors, and linear-threshold learning using Winnow. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, pages 147–156, 1991.

[29] W. Maass and G. Turan. *How fast can a threshold gate learn?*, pages 381–414. MIT Press, 1994.

[30] B. Morris and A. Sinclair. Random walks on truncated cubes and sampling 0-1 knapsack solutions (preliminary version). In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 230–240, 1999.

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

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

[33] A. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

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

[35] V. V. Petrov. *Limit theorems of probability theory*. Oxford Science Publications, Oxford, England, 1995.

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

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

[38] R. Servedio. Monotone Boolean Formulas can Approximate Monotone Linear Threshold Functions. *Discrete Applied Mathematics*, 142(1-3):181–187, 2004.

[39] J. Shawe-Taylor and N. Cristianini. *An introduction to support vector machines*. Cambridge University Press, 2000.

[40] L. Trevisan. A note on approximate counting for $k$-dnf. In *Proceedings of the Eighth International Workshop on Randomization and Computation*, pages 417–426, 2004.

[41] L. Valiant. Short monotone formulae for the majority function. *Journal of Algorithms*, 5:363–366, 1984.

[42] R. Winder. Chow parameters in threshold logic. *Journal of the ACM*, 18(2):265–289, 1971.

## A. Proof of Theorem 2

Let $\epsilon > 0$ be given and let $f\colon \{-1,1\}^n \to \{-1,1\}$ be any linear threshold function. We may suppose that $f(x) =$

$\operatorname{sgn}(F(x))$ where $F(x) = \sum_{i=1}^n w_i x_i - \theta$ with $1 = |w_1| \geq |w_2| \geq \cdots \geq |w_n| \geq 0$; note that wlog we have $|\theta| \leq \sum_{i=1}^n |w_i|$.

Fix any $g\colon \{-1,1\}^n \to \{-1,1\}$ where for $S = \emptyset, \{1\}, \ldots, \{n\}$ we have $|\hat{g}(S) - \hat{f}(S)| \leq 1/M$ with $M = n \cdot 2^{\tilde{O}(1/\epsilon^2)}$. Let $D$ denote $\{x \in \{-1,1\}^n : g(x) \neq f(x)\}$ and $\tau$ denote $|D|/2^n$. We will show that $\tau \leq \epsilon$ and thus establish Theorem 2.

We have

$$\begin{aligned}
\mathbf{E}[|F(x)|] &= \mathbf{E}[fF] = \sum_{S \subseteq [n]} \hat{f}(S)\hat{F}(S) \\
&= \sum_{|S| \leq 1} \hat{f}(S)\hat{F}(S) \\
&= \hat{f}(\emptyset)(-\theta) + \sum_{i=1}^n \hat{f}(\{i\}) w_i \\
&\leq \hat{g}(\emptyset)(-\theta) + \hat{g}(\{1\})w_1 + \cdots + \hat{g}(\{n\})w_n \\
&\quad + (|\theta| + \sum_{i=1}^n |w_i|)/M \qquad (5)
\end{aligned}$$

The second equality above is Parseval's identity, the third is because $F$'s only nonzero Fourier coefficients are of degree 0 and 1, and the fourth is by definition of $F$. The inequality (5) is from our assumption on the Fourier coefficients of $g$. Using Parseval again and writing $B$ to denote $(|\theta| + \sum_{i=1}^n |w_i|)/M$, we have

$$\begin{aligned}
(5) &= \sum_{|S| \leq 1} \hat{g}(S)\hat{F}(S) + B = \sum_{S \subseteq [n]} \hat{g}(S)\hat{F}(S) + B \\
&= \mathbf{E}[g(x)F(x)] + B.
\end{aligned}$$

Rearranging, this gives

$$\begin{aligned}
\frac{|\theta| + \sum_{i=1}^n |w_i|}{M} &\geq \mathbf{E}[|F(x)| - g(x)F(x)] \\
&= \frac{2}{2^n} \sum_{x \in D} |F(x)|. \qquad (6)
\end{aligned}$$

Thus far we have followed the proof from [3] (which is itself closely based on [5]), and indeed it is not difficult to complete the proof of Theorem 7 from here. Instead we will use our ideas from Section 4. The approach is to show that only a small number of points in $\{-1,1\}^n$ can have $|F(x)|$ very small, and thus if $|D|$ is large then the right hand side of (6) must be fairly large, which contradicts (6).

**Case I:** $\|w\| \geq \frac{12}{\epsilon}$. Let $\lambda \geq 1$ be such that $\frac{\epsilon}{2} = \frac{6\lambda}{\|w\|}$. By Theorem 4 we have $\Pr[|F(x)| \leq \lambda] \leq \frac{\epsilon}{2}$. Now suppose that $\tau > \epsilon$; this would mean that for at least $\frac{\epsilon}{2}2^n$ points $x \in D$ we have $|F(x)| > \lambda = \epsilon\|w\|/12$. But the bound (6) now gives

$$(|\theta| + \sum_{i=1}^n |w_i|)/M \geq \frac{2}{2^n} \cdot \frac{\epsilon}{2} 2^n \cdot \frac{\epsilon\|w\|}{12} = \frac{\epsilon^2\|w\|}{12}$$

This implies that we must have

$$M \leq \frac{12(|\theta| + \sum |w_i|)}{\epsilon^2\|w\|} \leq \frac{(|\theta| + \sum |w_i|)}{\epsilon} \leq \frac{2n}{\epsilon}.$$

which contradicts the definition of $M$; so case I is proved.

**Case II:** $\|w\| < \frac{12}{\epsilon}$. In this case we will use the following result due to Håstad [17], which gives a bound on the rate at which weights need to decrease (from largest to smallest in magnitude) for any linear threshold function over $\{-1,1\}^n$.

**Theorem 9 (Håstad [17])** *Let $f\colon \{-1,1\}^n \to \{-1,1\}$ be any linear threshold function which depends on all $n$ variables. There is a representation $\operatorname{sgn}(\sum_i w_i x_i - \theta)$ for $f$ which is such that (assuming the weights $w_1, \ldots, w_n$ are ordered by decreasing magnitude $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$) we have $|w_i| \geq \frac{|w_1|}{i!(n+1)}$ for all $i = 2, \ldots, n$.*

We prove Theorem 9 in Section A.1. Note that this implies in general that for any constant $c = O(1)$, the $c$-th largest weight of any LTF need be at most $1/O(n)$ times smaller than the largest weight. More specifically, in our context Theorem 9 lets us assume without loss of generality that the original weights $w_1, \ldots, w_n$ for $f$ satisfy $|w_i| \geq \frac{1}{i!(n+1)}$ for each $i$. This will prove useful in both cases IIa and IIb below.

In the following $\ell = \tilde{O}(1/\epsilon^2)$ as in Section 4.

**Case IIa:** $w_k^2/(\sum_{j=k}^n w_j^2) > \epsilon^2/144$ for all $k = 1, \ldots, \ell$. As in Case IIa of Section 4 we let $W = w_{\ell+1}^2 + \cdots + w_n^2$, but now we set $\eta' = 2\sqrt{W \ln(8/\epsilon)}$ (compare this with the $\eta = \sqrt{W \ln(4/\epsilon)}$ of the earlier proof).

We have that $|F(x)| \leq \eta'/2$ only if either $|w_{\ell+1}x_{\ell+1} + \cdots + w_n x_n| \geq \eta'/2$ or $|w_1 x_1 + \cdots + w_\ell x_\ell - \theta| \leq \eta'$. As in the derivation of equation (2) the Hoeffding bound gives us $\Pr[|w_{\ell+1}x_{\ell+1} + \cdots + w_n x_n| \geq \eta'/2] \leq \epsilon/4$. It remains to bound $\Pr[|w_1 x_1 + \cdots + w_\ell x_\ell - \theta| \leq \eta']$ by $\epsilon/4$; again reasoning as in the earlier section it suffices to show that

$$\Pr[|w_1 x_1 + \cdots + w_\ell x_\ell - \theta| \leq (24/\epsilon)\sqrt{2\ln(8/\epsilon)}w_\ell] \leq \epsilon/4. \qquad (7)$$

Comparing this with Equation (3), we see that the two expressions differ only in constant factors. One can verify that the arguments of Case IIa in Section 4 (with suitably adjusted constants) also yield (7) as desired.

We thus have that $\Pr[|F(x)| \leq \eta'/2] \leq \epsilon/2$. From the definitions of $\eta'$ and $W$ we have that $\eta'/2 \geq \sqrt{W} \geq |w_{\ell+1}|$, so consequently

$$\Pr[|F(x)| \leq |w_{\ell+1}|] \leq \epsilon/2. \qquad (8)$$

Now let us suppose that $\tau > \epsilon$. Reasoning as in Case I, we thus have that at least $\frac{\epsilon}{2}2^n$ many points $x \in D$ have $|F(x)| > |w_{\ell+1}|$. The bound (6) now gives

$$(|\theta| + \sum_{i=1}^n |w_i|)/M \geq \frac{2}{2^n} \cdot \frac{\epsilon}{2} 2^n \cdot |w_{\ell+1}|$$

which is equivalent to

$$M \leq \frac{|\theta| + \sum |w_i|}{\epsilon |w_{\ell+1}|}.$$

Since $|\theta| \leq \sum |w_i|$, we have that

$$\begin{aligned} M \leq \frac{2}{\epsilon} \cdot \frac{\sum_{i=1}^{n} |w_i|}{|w_{\ell+1}|} &\leq \frac{2}{\epsilon} \left( \frac{\ell}{|w_{\ell+1}|} + \frac{\sum_{i=\ell+1}^{n} |w_i|}{|w_{\ell+1}|} \right) \\ &\leq \frac{2}{\epsilon} \left( \frac{\ell}{|w_{\ell+1}|} + n \right) \\ &\leq \frac{2}{\epsilon} \left( \ell \cdot (\ell+1)!(n+1) + n \right) \end{aligned}$$

where the second inequality holds since each of $|w_1|, \ldots, |w_\ell|$ is at most 1, the third inequality holds since each of $|w_{\ell+1}|, \ldots, |w_n|$ is at most $|w_{\ell+1}|$, and the fourth inequality follows from Theorem 9. But recalling that $\ell = \tilde{O}(1/\epsilon^2)$, this upper bound on $M$ contradicts the fact that $M = n \cdot 2^{\tilde{O}(1/\epsilon^2)}$ (for a suitable choice of the hidden polylogarithmic factor in the exponent of the definition of $M$).

**Case IIb:** $w_k^2 / (\sum_{j=k}^{n} w_j^2) \leq \epsilon^2/144$ for some $k \in \{1, \ldots, \ell\}$. For each $i = 1, \ldots, n$ let $v_i$ denote $w_i/|w_k|$, so we have $1 = |v_k| \geq |v_{k+1}| \geq \cdots \geq v_n$. Using Theorem 4 with $\lambda = 1$, we have that for all $\tau \in \mathbf{R}$,

$$\begin{aligned} & \Pr_{x_k, \ldots, x_n} [|w_k x_k + \cdots + w_n x_n - \tau |w_k|| \leq |w_k|] \\ = & \Pr_{x_k, \ldots, x_n} [|v_k x_k + \cdots + v_n x_n - \tau| \leq 1] \\ \leq & \ 6/\sqrt{v_k^2 + \cdots + v_n^2} \\ = & \ 6|w_k|/\sqrt{w_k^2 + \cdots + w_n^2} \leq \epsilon/2 \end{aligned}$$

where the last inequality holds since we are in Case IIb. It follows that for any $\theta \in \mathbf{R}$ we have

$$\begin{aligned} & \Pr_{x_1, \ldots, x_n} [|w_1 x_1 + \cdots + w_n x_n - \theta| \leq |w_k|] \\ = & \Pr_{x_1, \ldots, x_n} [|F(x)| \leq |w_k|] \leq \epsilon/2. \end{aligned}$$

Now an entirely similar argument to that given from equation (8) through the end of Case IIa shows that as in that case, we must have $\tau \leq \epsilon$. This concludes the analysis of all cases, so Theorem 2 is proved. ■

## A.1. Proof of Theorem 9.

We first consider the case in which $f(x) = f(-x)$ for all $x \in \{-1, 1\}^n$, i.e. $f$ can be represented with a threshold of zero. Once we have the result for such $f$ we will use it to prove the result for general $f$.

Let $\mathrm{sgn}(w_1 x_1 + \cdots + w_n x_n)$ be a representation for $f$ which satisfies the conditions

1. $\mathrm{sgn}(w \cdot x) = f(x)$ for each $x \in \{-1, 1\}^n$.
2. $|w \cdot x| \geq 1$ for each $x \in \{-1, 1\}^n$.
3. Among all vectors in $\mathbf{R}^n$ which satisfy conditions (1) and (2) above, $w$ maximizes the number of $x \in \{-1, 1\}^n$ which have $|w \cdot x| = 1$. If there is more than one such $w$, choose one arbitrarily.

The argument in Section 3 of [16] now implies that there is a set $x^{(1)}, \ldots, x^{(n)}$ of $n$ elements of $\{-1, 1\}^n$ such that the coefficients $w_1, \ldots, w_n$ are determined as the unique solution to the system of equations

$$v_1 x_1^{(i)} + \cdots + v_n x_n^{(i)} = f(x^{(i)}) \quad \text{for } i = 1, \ldots, n.$$

This is a system of $n$ equations in the variables $v_1, \ldots, v_n$ where each coefficient is $\pm 1$ and the right-hand side of each equation, $f(x^{(i)})$, is also $\pm 1$. Recall that $f$ depends on all $n$ variables and consequently we have that each $w_i$ – and in particular $w_n$ – is nonzero. Using this fact it is not difficult to see that the above system of equations is equivalent to the following system of $n$ equations in $v_1, \ldots, v_n$:

$$f(x^{(1)})(v_1 x_1^{(1)} + \cdots + v_n x_n^{(1)}) = f(x^{(i)})(v_1 x_1^{(i)} + \cdots + v_n x_n^{(i)})$$

$$\text{for } i = 2, \ldots, n,$$

and

$$v_n = w_n.$$

Each of these first $n-1$ equations has no constant term and (dividing by two and rearranging) can be rewritten as $v \cdot y^{(i)} = 0$, where $y^{(i)}$ is a vector whose entries are all $-1, 0$ or $1$. So we have that $w_1, \ldots, w_n$ is the solution to the system of equations

$$Yv = b$$

where $Y$ is a nonsingular $n \times n$ matrix with $\{-1, 0, 1\}$ entries where the last row is $(0\ 0\ \cdots\ 0\ 1)$ and $b_1 = \cdots = b_{n-1} = 0$, $b_n = w_n$.

We assume that $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$, and now show that $|w_k|$ must be somewhat large compared with $|w_1|$.

After possibly reordering the first $n-1$ equations, we can find a linear combination of the first $k-1$ equations such that the only nonzero coefficient among $v_1, \ldots, v_{k-1}$ belongs to $v_1$, i.e. an equation of the form

$$v_1 = \sum_{j=k}^{n} a_j v_j. \tag{9}$$

Using Cramer's Rule and the fact that any $(k-1) \times (k-1)$ matrix with entries in $\{-1, 0, 1\}$ has determinant at most $(k-1)!$, it is not hard to show that an equality in the form of (9) must exist where each $|a_j| \leq (k-1)!$. But now if

$|w_k| < \frac{|w_1|}{(k-1)!(n-k+1)}$, then it is impossible for $w$ to satisfy (9) since the right-hand side must be too small. This proves that $|w_k| \geq \frac{|w_1|}{(k-1)!(n-k+1)} \geq \frac{|w_1|}{(k-1)!n}$, so we are done in the zero-threshold case.

We can treat the case where $f$ has a nonzero threshold by considering the function $g : \{-1,1\}^{n+1} \rightarrow \{-1,1\}$ which has zero threshold but an $(n+1)$-st weight which is the threshold of $f$. The argument for the zero-threshold case now shows that $g$ has a representation $\mathrm{sgn}(w_1x_1 + \cdots + w_nx_n + w_{n+1}x_{n+1})$ with $|w_1| \geq \cdots \geq |w_{n+1}|$ and $|w_k| \geq \frac{|w_1|}{(k-1)!(n+1)}$; note that one of these $w_i$ weights actually corresponds to the threshold of the original LTF $f$. If $w_1$ is the threshold then $w_2$ is actually the largest weight of $f$ in magnitude and we have $|w_k| \geq \frac{|w_2|}{(k-1)!(n+1)}$. If $w_r$ is the threshold for some $r > 1$ then $w_1$ is indeed the largest of $f$'s weights. In this case, for $k < r$ we have that $f$'s $k$-th biggest weight is $w_k$ which satisfies $|w_k| \geq \frac{|w_1|}{(k-1)!(n+1)}$, whereas for $k > r$ we have that $f$'s $k$-th biggest weight is $w_{k+1}$ which satisfies $|w_{k+1}| \geq \frac{|w_1|}{k!(n+1)}$. So in every case the magnitude of the $k$-th biggest weight is at least $\frac{1}{k!(n+1)}$ times the magnitude of the biggest weight, and Theorem 9 is proved. ■