THE CHOW PARAMETERS PROBLEM

RYAN O'DONNELL[*] AND ROCCO A. SERVEDIO[†]

Abstract. In the 2nd Annual FOCS (1961), Chao-Kong Chow proved that every Boolean threshold function is uniquely determined by its degree-0 and degree-1 Fourier coefficients. These numbers became known as the *Chow Parameters.* Providing an algorithmic version of Chow's Theorem—i.e., efficiently constructing a representation of a threshold function given its Chow Parameters—has remained open ever since. This problem has received significant study in the fields of circuit complexity, game theory and the design of voting systems, and learning theory. In this paper we effectively solve the problem, giving a randomized PTAS with the following behavior:

Given the Chow Parameters of a Boolean threshold function f over n bits and any constant $\epsilon > 0$, the algorithm runs in time $O(n^2 \log^2 n)$ and with high probability outputs a representation of a threshold function f' which is ϵ-close to f.

Along the way we prove several new results of independent interest about Boolean threshold functions. In addition to various structural results, these include $\tilde{O}(n^2)$-time learning algorithms for threshold functions under the uniform distribution in the following models:

(i) The Restricted Focus of Attention model, answering an open question of Birkendorf et al.

(ii) An agnostic-type model. This contrasts with recent results of Guruswami and Raghavendra who show NP-hardness for the problem under general distributions.

(iii) The PAC model, with constant ϵ. Our $\tilde{O}(n^2)$-time algorithm substantially improves on the previous best known running time and nearly matches the $\Omega(n^2)$ bits of training data that any successful learning algorithm must use.

Key words. Chow Parameters, threshold functions, approximation, learning theory

AMS subject classifications. 94C10, 06E30, 68Q32, 68R99, 91B12, 91B14, 42C10

1. Introduction.

This paper is concerned with Boolean threshold functions:

Definition 1.1. *A Boolean function* $f : \{-1,1\}^n \to \{-1,1\}$ *is a* threshold function *if it is expressible as* $f(x) = \mathrm{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n)$ *for some real numbers* $w_0, w_1, \ldots, w_n$.

Boolean threshold functions are of fundamental interest in circuit complexity, game theory/voting theory, and learning theory. Early computer scientists studying "switching functions" (i.e., Boolean functions) spent an enormous amount of effort on the class of threshold functions; see for instance the books [10, 26, 36, 48, 38] on this topic. More recently, researchers in circuit complexity have worked to understand the computational power of threshold functions and shallow circuits with these functions as gates; see e.g. [21, 45, 24, 25, 22]. In game theory and social choice theory, where simple cooperative games [42] correspond to monotone Boolean functions, threshold functions (with nonnegative weights) are known as "weighted majority" games and have been extensively studied as models for voting, see e.g. [43, 27, 11, 54]. Finally, in various guises, the problem of learning an unknown threshold function ("halfspace") has arguably been the central problem in machine learning for much of the last two decades, with algorithms such as Perceptron, Weighted Majority, boosting, and support vector machines emerging as central tools in the field.

A beautiful result of C.-K. Chow from the 2nd FOCS conference [9] gives a surprising characterization of Boolean threshold functions: among all Boolean functions,

[*] Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15213 (odonnell@cs.cmu.edu). Supported in part by NSF award CCF-0747250, a CyLab Research Grant, an Okawa Foundation Grant, and a Sloan Foundation Fellowship.

[†] Columbia University, 1214 Amsterdam Avenue, New York, New York, 10027 (rocco@cs.columbia.edu). Supported in part by NSF award CCF-0347282, NSF award CCF-0523664, and a Sloan Foundation Fellowship.

each threshold function $f : \{-1,1\}^n \to \{-1,1\}$ is uniquely determined by the "center of mass" of its positive inputs, $\text{avg}\{x \in \{-1,1\}^n : f(x) = 1\}$, and the number of positive inputs $\#\{x : f(x) = 1\}$. These $n+1$ parameters of f are equivalent, after scaling and additive shifting, to its degree-0 and degree-1 Fourier coefficients (and also, essentially, to its "influences" or "Banzhaf power indices"). We give a formal definition:

DEFINITION 1.2. *Given any Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, its* Chow Parameters[1] *are the rational numbers $\widehat{f}(0), \widehat{f}(1), \ldots, \widehat{f}(n)$ defined by $\widehat{f}(0) = \mathbf{E}[f(x)]$, $\widehat{f}(i) = \mathbf{E}[f(x)x_i]$, for $1 \leq i \leq n$. We also say the* Chow Vector *of f is $\vec{\chi} = \vec{\chi}_f = (\widehat{f}(0), \widehat{f}(1), \ldots, \widehat{f}(n))$.* Throughout this paper the notation $\mathbf{E}[\cdot]$ and $\mathbf{Pr}[\cdot]$ refers to an $x \in \{-1,1\}^n$ chosen uniformly at random. (We note that this corresponds to the "Impartial Culture Assumption" in the theory of social choice [19].) Our notation slightly abuses the standard Fourier coefficient notation of $\widehat{f}(\emptyset)$ and $\widehat{f}(\{i\})$.

Chow's Theorem implies that the following algorithmic problem is in principle solvable:

The Chow Parameters Problem. Given the Chow Parameters $\widehat{f}(0)$, $\widehat{f}(1)$, $\ldots$, $\widehat{f}(n)$ of a Boolean threshold function f, output a representation of f as $f(x) = \text{sgn}(w_0 + w_1x_1 + \cdots w_nx_n)$.

Unfortunately, the proof of Chow's Theorem (reviewed in Section 2.3) is completely nonconstructive and does not suggest any algorithm, much less an efficient one. As we now briefly describe, over the past five decades the Chow Parameters problem has been considered by researchers in a range of different fields.

1.1. Background on the Chow Parameters problem. As far back as 1960 researchers studying Boolean functions were interested in finding an efficient algorithm for the Chow Parameters problem [14]. Electrical engineers at the time faced the following problem: Given an explicit truth table, determine if it can be realized as a threshold circuit and if so, which one. The Chow Parameters are easily computed from a truth table, and Chow's Theorem implies that they give a unique representation for every threshold function. Several heuristics were proposed for the Chow Parameters problem [30, 56, 29, 10], an empirical study was performed to compare various methods [58], and lookup tables were produced mapping Chow Vectors into weights-based representations for each threshold function on six [39], seven [57], and eight [41] bits. Winder provides a good early survey [59]. Generalizations of Chow's Theorem were later given in [7, 46].

Researchers in game theory have also considered the Chow Parameters problem; Chow's Theorem was independently rediscovered by the game theorist Lapidot [34] and subsequently studied in [11, 13, 54, 18]. In the realm of social choice and voting theory the Chow Parameters represent the Banzhaf power indices [43, 2] of the n voters—a measure of each one's "influence" over the outcome. Here the Chow Parameters problem is very natural: Consider designing a voting rule for, say, the European Union. Target Banzhaf power indices are given, usually in proportion to the square-root of the states' populations, and one wishes to come up with a weighted majority voting rule whose power indices are as close to the targets as possible. Researchers in voting theory have recently devoted significant attention to this problem [35, 8], calling it a "fundamental constitutional problem" [16] and in particular considering its computational complexity [51, 1].

[1] Chow's Theorem was proven simultaneously by Tannenbaum [53], but the terminology "Chow Parameters" has stuck.

The Chow Parameters problem also has motivation from learning theory. Ben-David and Dichterman [3] introduced the "Restricted Focus of Attention (RFA)" model to formalize the idea that learning algorithms often have only partial access to each example vector. Birkendorf et al. [5] performed a comprehensive study of the RFA model and observed that the approximation version of the Chow Parameters problem (given approximate Chow Parameters, output an approximating threshold function) is equivalent to the problem of efficiently learning threshold functions under the uniform distribution in the 1-RFA model. (In the 1-RFA model the learner is only allowed to see one bit of each example string in addition to the label; we give details in Section 10.) As the main open question posed in [5], Birkendorf et al. asked whether there is an efficient uniform distribution learning algorithm for threshold functions in the 1-RFA model. This question motivated subsequent research [20, 47] which gave *information-theoretic* sample complexity upper bounds for this learning problem (see Section 3); however no computationally efficient algorithm was previously known.

To summarize, we believe that the range of different contexts in which the Chow Parameters Problem has arisen is evidence of its fundamental status.

1.2. The Chow Parameters problem reframed as an approximation problem. It is unlikely that the Chow Parameters Problem can be solved exactly in polynomial time—note that even checking the correctness of a candidate solution is #P-complete, because computing $\widehat{f}(0)$ is equivalent to counting 0-1 knapsack solutions. Thus, as is implicitly proposed in [5, 1], it is natural to look for a polynomial-time approximation scheme (PTAS). Here we mean an approximation in the following sense:

DEFINITION 1.3. *The* distance *between two Boolean functions* $f, g : \{-1,1\}^n \to \{-1,1\}$ *is* $\mathrm{dist}(f,g) \overset{def}{=} \mathbf{Pr}[f(x) \neq g(x)]$. *If* $\mathrm{dist}(f,g) \leq \epsilon$ *we say that* f *and* g *are* ϵ-close*.*

We would like a PTAS which, given a value ϵ and the Chow Parameters of f, outputs a (representation of a) threshold function f' that is ϵ-close to f. With this relaxed goal of approximating f, one may even tolerate only an approximation of the Chow Parameters of f; this gives us the variant of the problem that Birkendorf et al. considered. (Note that, as we discuss in Section 3, it is in no way obvious that approximate Chow Parameters even *information-theoretically* specify an approximator to f.) In particular the following notion of "approximate" Chow Parameters proves to be most natural:

DEFINITION 1.4. *Let* $f, g : \{-1,1\}^n \to \{-1,1\}$*. We define* $d_{\mathrm{Chow}}(f,g) \overset{def}{=} \sqrt{\sum_{j=0}^{n}(\widehat{f}(j) - \widehat{g}(j))^2}$ *to be the* Chow Distance *between* f *and* g*.*

1.3. Our results. Our main result is an efficient PTAS $\mathcal{A}$ for the Chow Parameters problem which succeeds given approximations to the Chow Parameters. We prove:

MAIN THEOREM. *There is a function* $\kappa(\epsilon) = 2^{-\tilde{O}(1/\epsilon^2)}$ *such that the following holds: Let* $f : \{-1,1\}^n \to \{-1,1\}$ *be a threshold function and let* $0 < \epsilon < 1/2$*. Write* $\vec{\chi}$ *for the Chow Vector of* f *and assume that* $\vec{\alpha}$ *is a vector satisfying* $\|\vec{\alpha} - \vec{\chi}\| \leq \kappa(\epsilon)$*. Then given as input* $\vec{\alpha}$ *and* ϵ *the algorithm* $\mathcal{A}$ *performs* $2^{\mathrm{poly}(1/\kappa(\epsilon))} \cdot n^2 \cdot \log n \cdot \log(\frac{n}{\delta})$ *bit operations and outputs the (weights-based) representation of a threshold function* f^* *which with probability at least* $1 - \delta$ *satisfies* $\mathrm{dist}(f, f^*) \leq \epsilon$*.*

Although the running time dependence on ϵ is doubly-exponential, we emphasize that the polynomial dependence on n is quadratic, independent of ϵ; i.e., $\mathcal{A}$ is an

"EPTAS". Some of our learning applications have only singly-exponential dependence on ϵ.

1.4. Our approach. We briefly describe the two main ingredients of our approach and explain how we combine them to obtain the efficient algorithm $\mathcal{A}$.

First ingredient: small Chow Distance from a threshold function implies small distance. An immediate question that arises when thinking about the Chow Parameters problem is how to recognize whether a candidate solution is a good one. If we are given the Chow Vector $\vec{\chi}_f$ of an unknown threshold function f and we have a candidate threshold function g, we can approximate the Chow Vector $\vec{\chi}_g$ of g by sampling. The following Proposition is easily proved via Fourier analysis in Section 2.3:

Proposition 1.5. *$d_{\text{Chow}}(f,g) \leq 2\sqrt{\text{dist}(f,g)}$.*

This means that if $d_{\text{Chow}}(f,g)$ is large then f and g are far apart. But if $d_{\text{Chow}}(f,g)$ is small, does this necessarily mean that f and g are close?

This question has been studied in the learning theory community, in [5] (for threshold functions with small integer weights), [20], and [47]. In Section 3 we show that the answer is yes by proving the following "robust" version of Chow's Theorem:

Theorem 1.6. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any threshold function and let $g : \{-1,1\}^n \to \{-1,1\}$ be any Boolean function such that $d_{\text{Chow}}(f,g) \leq \epsilon$. Then $\text{dist}(f,g) \leq \tilde{O}\left(1/\sqrt{\log(1/\epsilon)}\right)$.*

This is the first result of this nature that is completely independent of n. A key ingredient in the proof of Theorem 1.6 is a new result showing that every threshold function f is extremely close to a threshold function f' for which only a very small fraction of points have small "margin" (see Section 6 for a precise statement). We feel that this and Theorem 1.6 have independent interest as structural results about threshold functions.

Second ingredient: using the Chow Parameters as weights. The second ingredient in our approach is to establish a result, Theorem 7.1, having the following corollary:

Corollary 7.2. *There is an absolute constant $C > 0$ such that the following holds. Let $f(x) = \text{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n)$ be any threshold function, and let H be the set of $1/\epsilon^C$ indices i for which $|w_i|$ is largest.*[2] *Then there exists a threshold function $f'(x) = \text{sgn}(v_0 + v_1x_1 + \cdots + v_nx_n)$ with $\text{dist}(f,f') \leq \epsilon$ in which the weights v_i for $i \in [n] \setminus H$* are the Chow Parameters $\widehat{f}(i)$ themselves.

The heuristic of using the Chow Parameters as possible weights was considered by several researchers in the early '60s (see [59]); however no theorem on the efficacy of this approach was previously known. Our proof of Theorem 7.1 and its robust version Theorem 7.4 rely in part on recent work of Matulef et al. on Property Testing for threshold functions [37].

The algorithm and intuitive explanation. Given these two ingredients, our PTAS $\mathcal{A}$ for the approximate Chow Parameters problem works by constructing a "small" (depending only on ϵ) number of candidate threshold functions. It enumerates "all" (in some sense) possible weight settings for the indices in H, and for each one produces a candidate threshold function by setting the remaining weights equal to the given Chow Parameters. The second ingredient tells us that at least one of these candidate threshold functions must be close to to the unknown threshold function f, and thus

[2] As we discuss at the beginning of Section 7, for any threshold function f the value $|\widehat{f}(i)|$ is equal to $\text{Inf}_i(f)$, the influence of the i-th variable on f. It is well known and easy to show (see e.g. Lemma 7 of [17]) that for a threshold function $f(x) = \text{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n)$, if $\text{Inf}_i(f) > \text{Inf}_j(f)$ then $|w_i| > |w_j|$. So we may equivalently view H as the set of $1/\epsilon^C$ indices i for which $|\widehat{f}(i)|$ is largest.

must have small Chow Distance to f, by Proposition 1.5. Now the first ingredient tells us that *any* threshold function whose Chow Distance to the target Chow Vector is small must itself be close to the target. So the algorithm can estimate each of the candidates' Chow Vectors (this takes $\tilde{O}(n^2)$ time) and output any candidate whose Chow Distance to the target vector is small.

1.5. Consequences in learning theory.

As we show in Section 10, our approach yields a range of new algorithmic results in learning theory. Our Main Theorem directly gives the first poly(n)-time algorithm for learning threshold functions in the uniform distribution 1-RFA model, answering the question of [5]:

THEOREM 1.7. *There is an algorithm which performs* $2^{2^{\tilde{O}(1/\epsilon^2)}} \cdot n^2 \cdot \log n \cdot \log(\frac{n}{\delta})$ *bit operations and properly learns threshold functions to accuracy* ϵ *and confidence* $1-\delta$ *in the uniform distribution 1-RFA model.*

A variant of our algorithm gives a very fast agnostic-type learning algorithm for threshold functions (equivalently, an algorithm for learning Boolean threshold functions from uniformly distributed examples when there is adversarial noise in the labels):

THEOREM 1.8. *Let* g *be any Boolean function and let* $\mathsf{opt} = \min_f \mathbf{Pr}[f(x) \neq g(x)]$ *where the min is over all threshold functions and the probability is uniform over* $\{-1,1\}^n$. *Given an input parameter* $\epsilon > 0$ *and access to independent uniform examples* $(x, g(x))$, *algorithm* $\mathcal{B}$ *outputs the (weights-based) representation of a threshold function* f^* *which with probability at least* $1-\delta$ *satisfies* $\mathbf{Pr}[f^*(x) \neq g(x)] \leq O(\mathsf{opt}^{\Omega(1)}) + \epsilon$. *The algorithm performs* $\mathrm{poly}(1/\epsilon) \cdot n^2 \cdot \log(\frac{n}{\delta}) + 2^{\mathrm{poly}(1/\epsilon)} \cdot n \cdot \log n \cdot \log(\frac{1}{\delta})$ *bit operations.*

For example, if $\mathsf{opt} = 1/\log(n)$, our algorithm takes time $O(n^2 \cdot \log n \cdot \log(\frac{n}{\delta}))$ and outputs a hypothesis with accuracy $1/\log^{\Omega(1)}(n)$. Thereom 1.8 is in interesting contrast with the algorithm of Kalai et al. [28] which constructs an $(\mathsf{opt} + \epsilon)$-accurate hypothesis but runs in $n^{\mathrm{poly}(1/\epsilon)}$ time (and does not output a threshold function). As we discuss in Section 10, recent hardness results of Guruswami and Raghavendra [23] imply that if P $\neq$ NP there can be no algorithm comparable to ours for learning under arbitrary (as opposed to uniform) distributions over $\{-1,1\}^n$.

Finally, as a corollary of Theorem 1.8, we obtain a uniform-distribution PAC learning algorithm for threshold functions that runs in time $\tilde{O}(n^2)$ for learning to constant accuracy $\epsilon = \Theta(1)$. The fastest previous algorithm we are aware of for learning arbitrary threshold functions in this model (linear programming, using Vaidya [55]) runs in $\tilde{O}(n^{4.5}) \cdot \mathrm{poly}(1/\epsilon)$ time. Thus our algorithm is significantly faster for learning to accuracy $\epsilon = \Theta(1)$, and in fact is faster as long as $\epsilon < 1/(\log n)^c$ for sufficiently small constant $c > 0$. As we explain later, our time bound is very close to the $\Omega(n^2)$ bits of input that any learning algorithm must use.

2. Preliminaries.

2.1. Fourier analysis.

This paper extensively uses the basics of Fourier analysis over the Boolean cube $\{-1,1\}^n$. We give a brief review. We consider functions $f : \{-1,1\}^n \to \mathbb{R}$ (though we often focus on Boolean-valued functions which map to $\{-1,1\}$), and we think of the inputs x to f as being distributed according to the uniform probability distribution. The set of such functions forms a 2^n-dimensional inner product space with inner product given by $\langle f, g \rangle = \mathbf{E}_x[f(x)g(x)]$. The set of functions $(\chi_S)_{S \subseteq [n]}$ defined by $\chi_S(x) = \prod_{i \in S} x_i$ forms a complete orthonormal basis for this space. We will also often write simply x_S for $\prod_{i \in S} x_i$. Given a function

$f : \{-1,1\}^n \to \mathbb{R}$ we define its *Fourier coefficients* by $\widehat{f}(S) = \mathbf{E}_x[f(x)x_S]$, and we have that $f(x) = \sum_S \widehat{f}(S)x_S$.

As an easy consequence of orthonormality we have *Plancherel's identity* $\langle f, g\rangle = \sum_S \widehat{f}(S)\widehat{g}(S)$, which has as a special case *Parseval's identity*, $\mathbf{E}_x[f(x)^2] = \sum_S \widehat{f}(S)^2$. From this it follows that for every $f : \{-1,1\}^n \to \{-1,1\}$ we have $\sum_S \widehat{f}(S)^2 = 1$.

The following definitions are fairly standard in the analysis of Boolean functions:

DEFINITION 2.1. *A function $f : \{-1,1\}^n \to \{-1,1\}$ is said to be a "junta on $J \subset [n]$" if f only depends on the coordinates in J. Typically we think of J as a "small" set in this case.*

DEFINITION 2.2. *We say that $f : \{-1,1\}^n \to \mathbb{R}$ is "τ-regular" if $|\hat{f}(i)| \le \tau$ for all $i \in [n]$.*

The following simple lemma is implicit in [37]; we state and prove it explicitly here for completeness.

LEMMA 2.3. *Let $f(x) : \{-1,1\}^n \to \{-1,1\}$ be a Boolean threshold function and let $J \subset [n]$ be any subset of coordinates. If f is τ-close to a junta on J, then f is τ-close to a junta on J which is itself a Boolean threshold function.*

Proof. We assume without loss of generality that J is the set $\{1,\ldots,r\}$. It is clear that the junta over $\{-1,1\}^r$ to which f is closest is the function $g(x_1,\ldots,x_r)$ that maps each input $(x_1,\ldots,x_r)$ to the more commonly occuring value of the restricted function $f_{x_1,\ldots,x_r}$ (a function of variables $x_{r+1},\ldots,x_n$). But for $f(x) = \mathrm{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n)$ this more common value will be $\mathrm{sgn}(w_0 + w_1x_1 + \cdots + w_rx_r)$, because for uniform $(x_{r+1},\ldots,x_n) \in \{-1,1\}^{n-r}$ the random variable $w_{r+1}x_{r+1} + \cdots + w_nx_n$ is centered around zero. □

We will also require the following lemma, which gives a lower bound on the degree-1 Fourier weight of any threshold function in terms of its bias:

LEMMA 2.4. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean threshold function and suppose that $1 - |\mathbf{E}[f]| = p$. Then*

$$\sum_{i=1}^{n} \widehat{f}(i)^2 \ge p^2/2.$$

Before giving the proof let us contrast this lemma with some known results. Proposition 2.2 of Talagrand [52] gives a general upper bound $\sum_{i=1}^n \widehat{f}(i)^2 \le O(p^2 \log(1/p))$ for any Boolean function satisfying $1 - |\mathbf{E}[f]| = p$. In [37] it is shown that a slightly stronger bound $\Theta(p^2 \log(1/p))$ holds for threshold functions f that are sufficiently τ-regular. However when we use Lemma 2.4 we will not have regularity (and even if we did, the extra log factor would not end up improving any of our bounds).

Proof. Write $f(x) = \mathrm{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n)$, where we assume without loss of generality that $\sum_{j=1}^n w_j^2 = 1$ and that $w_0 + w_1x_1 + \cdots + w_nx_n \ne 0$ for all $x \in \{-1,1\}^n$. We have

$$\mathbf{E}[f(x)(w\cdot x)] = \sum_{i=1}^{n} \widehat{f}(i)w_i \le \sqrt{\sum_{i=1}^{n} \widehat{f}(i)^2},$$

where the equality is Plancherel's identity and the inequality is Cauchy-Schwarz. On the other hand, using the definition of f we obtain

$$\mathbf{E}[f(x)(w\cdot x)] = \mathbf{E}[\mathbf{1}_{\{|w\cdot x| \ge |w_0|\}} \cdot |w\cdot x|] = p \cdot \mathbf{E}[|w\cdot x| \;\big|\; |w\cdot x| \ge |w_0|].$$

The first equality above holds because each x such that $|w \cdot x| < |w_0|$ can be paired with $-x$; the value of f is the same on these two inputs, so their contributions to the expectation cancel each other out. The second equality above is a routine renormalization using the equality $1 - |\mathbf{E}[f]| = p$.

We now recall the Khintchine inequality with best constant [50], which says that for any $w \in \mathbb{R}^n$ we have $\mathbf{E}[|w \cdot x|] \geq \frac{1}{\sqrt{2}}\|w\|$. Since $\|w\| = 1$ in our setting, we get $\mathbf{E}[|w \cdot x|] \geq= \frac{1}{\sqrt{2}}$, so surely $\mathbf{E}[|w \cdot x| \mid |w \cdot x| \geq |w_0|] \geq 1/\sqrt{2}$. Thus combining all statements yields

$$\sqrt{\sum_{i=1}^{n} \widehat{f}(i)^2} \geq p/\sqrt{2},$$

completing the proof. □

2.2. Mathematical tools. We use the following simple estimate on several occasions:

Fact 2.5. *Suppose A and B are nonnegative and $|A - B| \leq \eta$. Then $|\sqrt{A} - \sqrt{B}| \leq \eta/\sqrt{B}$.*

Proof. $|\sqrt{A} - \sqrt{B}| = \frac{|A-B|}{\sqrt{A}+\sqrt{B}} \leq \frac{\eta}{\sqrt{B}}$. □

We also will need some results from probability theory:

Definition 2.6. *We write Φ for the c.d.f. (cumulative density function) of a standard mean-0, variance-1 Gaussian random variable. We extend the notation by writing $\Phi[a, b]$ to denote $\Phi(b) - \Phi(a)$, allowing $b < a$. Finally, we will use the estimate $|\Phi[a, b]| \leq |b - a|$ without comment.*

The Berry-Esseen theorem is a version of the Central Limit Theorem with explicit error bounds:

Theorem 2.7. *(Berry-Esseen) Let $X_1, \ldots, X_n$ be a sequence of independent random variables satisfying $\mathbf{E}[X_i] = 0$ for all i, $\sqrt{\sum \mathbf{E}[X_i^2]} = \sigma$, and $\sum \mathbf{E}[|X_i|^3] = \rho_3$. Let $S = (X_1 + \cdots + X_n)/\sigma$ and let F denote the c.d.f. of S. Then*

$$\sup_x |F(x) - \Phi(x)| \leq C\rho_3/\sigma^3,$$

where Φ is the c.d.f. of a standard Gaussian random variable, and C is a universal constant. It is known [49] that one can take $C = .7915$.

Corollary 2.8. *Let $x_1, \ldots, x_m$ denote independent ± 1 random bits and let $w_1, \ldots, w_m \in \mathbb{R}$. Write $\sigma = \sqrt{\sum w_i^2}$, and assume $|w_i|/\sigma \leq \tau$ for all i. Then for any interval $[a, b] \subseteq \mathbb{R}$,*

$$\left|\mathbf{Pr}[a \leq w_1 x_1 + \cdots + w_m x_m \leq b] - \Phi([\tfrac{a}{\sigma}, \tfrac{b}{\sigma}])\right| \leq 2\tau.$$

In particular,

$$\mathbf{Pr}[a \leq w_1 x_1 + \cdots + w_m x_m \leq b] \leq \frac{|b - a|}{\sigma} + 2\tau.$$

2.3. Margins, and Chow's Theorem. Having introduced Fourier analysis, we recall and prove Proposition 1.5:

Proposition 1.5. $d_{\text{Chow}}(f, g) \leq 2\sqrt{\text{dist}(f, g)}$.

Proof. For $f, g : \{-1,1\}^n \to \{-1,1\}$ we have

$$\begin{aligned}\mathrm{dist}(f,g) = \frac{1}{4}\mathbf{E}[(f(x)-g(x))^2] &= \frac{1}{4}\sum_{S\subseteq[n]}(\widehat{f}(S)-\widehat{g}(S))^2 \\ &\geq \frac{1}{4}\sum_{j=0}^{n}(\widehat{f}(j)-\widehat{g}(j))^2 = \frac{1}{4}d_{\mathrm{Chow}}(f,g)^2,\end{aligned}$$

where the second equality is Parseval's identity. □

Let us introduce a notion of "margin" for threshold functions:

Definition 2.9. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean threshold function, $f(x) = \mathrm{sgn}(w_0+w_1x_1+\cdots+w_nx_n)$, where the weights are scaled so that $\sum_{j\geq 0} w_j^2 = 1$. Given a particular input $x \in \{-1,1\}^n$ we define* $\mathrm{marg}(f,x) = |w_0 + w_1x_1 + \cdots + w_nx_n|$.[3]

Remark 2.10. *The usual notion of "margin" from learning theory also involves scaling the* data points *x so that $\|x\| \leq 1$ for all x. Thus we have that the learning theoretic margin of f on x is* $\mathrm{marg}(f,x)/\sqrt{n}$.

We now present a proof of Chow's theorem from 1961:

Theorem 2.11. *(Chow.) Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean threshold function and let $g : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function such that $\widehat{g}(j) = \widehat{f}(j)$ for all $0 \leq j \leq n$. Then $g = f$.*

Note that another way of phrasing this is: "If f is a Boolean threshold function, g is a Boolean function, and $d_{\mathrm{Chow}}(f,g) = 0$, then $\mathrm{dist}(f,g) = 0$." Our Theorem 1.6 gives a "robust" version of this statement.

Proof. Write $f(x) = \mathrm{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n)$, where the weights are scaled so that $\sum_{j=0}^n w_j^2 = 1$. We may assume without loss of generality that $\mathrm{marg}(f,x) \neq 0$ for all x. (Otherwise, first perturb the weights slightly without changing f.) Now we have

$$\begin{aligned}0 &= \sum_{j=0}^{n} w_j(\widehat{f}(j)-\widehat{g}(j)) \\ &= \mathbf{E}[(w_0+w_1x_1+\cdots+w_nx_n)(f(x)-g(x))] \\ &= \mathbf{E}[\mathbf{1}_{\{f(x)\neq g(x)\}}\cdot 2\mathrm{marg}(f,x)].\end{aligned}$$

The first equality is by the assumption that $\widehat{f}(j) = \widehat{g}(j)$ for all $0 \leq j \leq n$, the second equality is linearity of expectation (or Plancherel's identity), and the third equality uses the fact that $f(x) = \mathrm{sgn}(w_0+w_1x_1+\cdots+w_nx_n)$. But since $\mathrm{marg}(f,x)$ is always strictly positive, we must have $\mathbf{Pr}[f(x) \neq g(x)] = 0$ as claimed. □

3. First ingredient: small Chow Distance implies small distance. Our main result in this section is the following.

Theorem 1.6 Restated. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any threshold function and let $g : \{-1,1\}^n \to \{-1,1\}$ be any Boolean function such that $d_{\mathrm{Chow}}(f,g) \leq \epsilon$. Then* $\mathrm{dist}(f,g) \leq \tilde{O}\left(1/\sqrt{\log(1/\epsilon)}\right)$.[4]

Let us compare this with some recent results with a similar qualitative flavor. The main result of Goldberg [20] is a proof that for any threshold function

[3] This notation is slightly informal as it doesn't show the dependence on the *representation* of f.

[4] For a quantity $q < 1$, the notation "$\tilde{O}(q)$" means "$O(q \cdot \log^c(1/q))$ for some absolute constant c."

f and any Boolean function g, if $|\widehat{f}(j) - \widehat{g}(j)| \le (\epsilon/n)^{O(\log(n/\epsilon)\log(1/\epsilon))}$ for all $0 \le j \le n$, then $\mathrm{dist}(f,g) \le \epsilon$. Note that the condition of Goldberg's theorem requires that $d_{\mathrm{Chow}}(f,g) \le n^{-O(\log n)}$. Subsequently Servedio [47] showed that to obtain $\mathrm{dist}(f,g) \le \epsilon$ it suffices to have $|\widehat{f}(j) - \widehat{g}(j)| \le 1/(2^{\tilde{O}(1/\epsilon^2)} \cdot n)$ for all $0 \le j \le n$. This is a worse requirement in terms of ϵ but a better one in terms of n; however it still requires that $d_{\mathrm{Chow}}(f,g) \le 1/\sqrt{n}$. In contrast, Theorem 1.6 allows the Chow Distance between f and g to be an absolute constant *independent* of n. This independence of n will be crucial later on when we use Theorem 1.6 to obtain a computationally efficient algorithm for the Chow Parameters problem.

At a high level, we prove Theorem 1.6 by giving a "robust" version of the proof of Chow's Theorem (Theorem 2.11). A first obvious approach to making the argument robust is to try to show that every threshold function has margin $\Omega(1)$ (independent of n) on every x. However this is well known to be badly false. A next attempt might be to show that every threshold function has a representation with margin $\Omega(1)$ on *almost* every x. This too turns out to be impossible (cf. our discussion after the statement of Lemma 5.1 below). The key to getting an "n-independent" margin lower bound is to also very slightly *alter* the threshold function. Specifically, the next few sections of the paper will be devoted to the proof of the following:

Theorem 3.1. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any threshold function and let $\rho > 0$ be sufficiently small. Then there is a threshold function $f' : \{-1,1\}^n \to \{-1,1\}$ with* $\mathrm{dist}(f,f') \le 2^{-1/\rho}$ *satisfying*

$$\Pr_x[\mathrm{marg}(f',x) \le \rho] \le \tilde{O}\left(1/\sqrt{\log(1/\rho)}\right).$$

In other words, any threshold function f is very close to another threshold function f' satisfying $\mathrm{marg}(f',x) \ge \Omega(1)$ for almost all x. We remark that although the fraction of points failing the margin bound could be as large as inverse-logarithmic in ρ, we only have to change f on a fraction of points which is exponentially small in $1/\rho$ to achieve this.

Theorem 3.1 is the key structural result for threshold functions that allows us to "robustify" the proof of Theorem 2.11. We will now show how Theorem 1.6 follows from Theorem 3.1.

Proof. (Theorem 1.6.) Given f, apply Theorem 3.1 with its parameter ρ set (with foresight) to

$$\rho = \sqrt{\epsilon \log(1/\epsilon)}.$$

This yields a threshold function $f'(x) = \mathrm{sgn}(u_0 + u_1x_1 + \cdots + u_nx_n)$, with $\sum_{j=0}^n u_j^2 = 1$ satisfying

$$\mathrm{dist}(f,f') \le 2^{-1/\rho} \ll \epsilon$$

and

$$\Pr_x[\mathrm{marg}(f',x) \le \rho] \le \tau \stackrel{\mathrm{def}}{=} \tilde{O}\left(1/\sqrt{\log(1/\rho)}\right) = \frac{\mathrm{poly}\log\log(1/\epsilon)}{\sqrt{\log(1/\epsilon)}}. \qquad (3.1)$$

Since $\mathrm{dist}(f,f') \le \epsilon$, by Proposition 1.5 we have $d_{\mathrm{Chow}}(f,f') \le 2\sqrt{\epsilon}$ and thus $d_{\mathrm{Chow}}(f',g) \le 3\sqrt{\epsilon}$ by the triangle inequality. We now follow the proof of Chow's

Theorem 2.11:

$$\begin{aligned} 3\sqrt{\epsilon} \quad \geq \quad d_{\mathrm{Chow}}(f', g) &= \sqrt{\sum_{j=0}^{n} u_j^2} \cdot \sqrt{\sum_{j=0}^{n} (\widehat{f'}(j) - \widehat{g}(j))^2} \\ &\geq \sum_{j=0}^{n} u_j (\widehat{f'}(j) - \widehat{g}(j)) \\ &= \mathbf{E}[\mathbf{1}_{\{f'(x) \neq g(x)\}} \cdot 2\mathrm{marg}(f', x)], \end{aligned} \tag{3.2}$$

where the second inequality is Cauchy-Schwarz.

Now suppose that $\mathbf{Pr}[f'(x) \neq g(x)] \geq 2\tau$. Then by (3.1) we must have that for at least a τ fraction of x's, both $f'(x) \neq g(x)$ and $\mathrm{marg}(f', x) > \rho$. This gives a contribution exceeding $\tau\rho$ to (3.2). But

$$\tau\rho = \sqrt{\epsilon} \cdot \mathrm{poly}\log\log(1/\epsilon) > 3\sqrt{\epsilon},$$

a contradiction. Thus $\mathrm{dist}(f', g) \leq 2\tau$ and so

$$\mathrm{dist}(f, g) \leq \mathrm{dist}(f, f') + \mathrm{dist}(f', g) \leq \epsilon + 2\tau = \tilde{O}\left(1/\sqrt{\log(1/\epsilon)}\right). \qquad \square$$

4. The critical index and anticoncentration. Fix a representation $f(x) = \mathrm{sgn}(w_0 + w_1 x_1 + \cdots + w_n x_n)$ of a threshold function. Throughout this section we adopt the convention that $|w_1| \geq \cdots \geq |w_n| > 0$ (this will be without loss of generality, by permuting indices).

The notion of the "critical index" of the sequence of weights $w_1, \ldots, w_n$ will be useful for us. Roughly speaking, it allows us to approximately decompose any linear form $w_0 + w_1 x_1 + \cdots + w_n x_n$ over random ± 1 x_i's into a short dominant "head", $w_0 + w_1 x_1 + \cdots + w_{\mathrm{small}} x_{\mathrm{small}}$, and a long remaining "tail" which acts like a Gaussian random variable. The "τ-critical index" of $w_1, \ldots, w_n$ is essentially the least index ℓ for which the random variable $w_\ell x_\ell + \cdots + w_n x_n$ behaves like a Gaussian up to error τ. The notion of a critical index was (implicitly) introduced and used in [47].

Towards proving a margin lower bound such as Theorem 3.1 for f, we need to show some kind of "anticoncentration" for the random variable $w_0 + w_1 x_1 + \cdots + w_n x_n$; we want it to rarely be near 0. Let us describe intuitively how analyzing the critical index helps us show this. If the critical index of $w_1, \ldots, w_n$ is large, then it must be the case that the initial weights $w_1, w_2, \ldots$ up to the critical index are rapidly decreasing (roughly speaking, if the weights $w_i, w_{i+1}, \ldots$ stayed about the same for a long stretch this would cause $w_i x_i + \cdots + w_n x_n$ to behave like a Gaussian). This rapid decrease can in turn be shown to imply that the the "head" part $w_0 + w_1 x_1 + \cdots + w_{\mathrm{small}} x_{\mathrm{small}}$ is not too concentrated around any particular value; see Theorem 4.2 below. On the other hand, if the critical index ℓ is small, then the random variable $w_\ell x_\ell + \cdots + w_n x_n$ behaves like a Gaussian. Since Gaussians have good anticoncentration, the overall linear form $w_0 + w_1 x_1 + \cdots + w_n x_n$ will have good anticoncentration, regardless of the head part's value. We need to alter f slightly to make these two cases go through, but having done so, we are able to bound the fraction of inputs x for which $\mathrm{marg}(f, x)$ is very small, leading to Theorem 3.1.

We now give precise definitions. For $1 \leq k \leq n$ we write σ_k to denote the 2-norm of the "tail weights" starting from k; i.e. $\sigma_k \overset{\mathrm{def}}{=} \sqrt{\sum_{i \geq k}^{n} w_i^2}$.

DEFINITION 4.1. *Fix a parameter* $0 < \tau < 1/2$. *We define the* τ-critical index *of the weight vector* w *to be the least index* ℓ *such that* w_ℓ *is "small" relative to* σ_ℓ *in the* *following sense:*

$$\frac{|w_\ell|}{\sigma_\ell} \leq \tau. \tag{4.1}$$

(If no index $1 \leq \ell \leq n$ satisfies (4.1), as is the case for $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, \frac{1}{2^n})$ for example, then we say that the τ-critical index is $+\infty$.) The connection between Equation (4.1) and behaving like a Gaussian up to error τ is given by the Berry-Esseen Theorem, stated in Section 2.2.

The following anticoncentration result shows that if the critical index is large, then the random variable $w_1x_1 + \cdots + w_nx_n$ does not put much probability mass close to any particular value:

THEOREM 4.2. *Let $0 < \tau < 1/2$ and $t \geq 1$ be parameters, and define $k = \lceil O(1)\frac{t}{\tau^2}\ln\left(\frac{t}{\tau}\right)\rceil$. If the τ-critical index ℓ for $w_1, \ldots, w_n$ satisfies $\ell \geq k$, then we have*

$$\mathop{\mathbf{Pr}}_{x}[|w_0 + w_1x_1 + \cdots + w_nx_n| \leq \sqrt{t} \cdot \sigma_k] \leq O(2^{-t}).$$

A similar result was established in [47]. The following subsections §4.1, 4.2, 4.3 are devoted to the proof of Theorem 4.2. Throughout, they assume ℓ denotes the τ-critical index of $w_1, \ldots, w_n$ where $|w_1| \geq \cdots \geq |w_n| > 0$ as in the condition of Theorem 4.2.

4.1. Partitioning weights into blocks. The following simple lemma shows that the tail weight decreases exponentially up to the τ-critical index:

LEMMA 4.3. *For $1 \leq a < b \leq \ell$, we have $\sigma_b^2 < (1-\tau^2)^{b-a}\sigma_a^2 < (1-\tau^2)^{b-a}w_a^2/\tau^2$.*

Proof. Since a is less than the critical index, we have $w_a^2 > \tau^2\sigma_a^2 = \tau^2(w_a^2 + \sigma_{a+1}^2)$, or equivalently $(1-\tau^2)w_a^2 > \tau^2\sigma_{a+1}^2$. Adding $(1-\tau^2)\sigma_{a+1}^2$ to both sides gives $(1-\tau^2)(w_a^2 + \sigma_{a+1}^2) > (1-\tau^2)\sigma_{a+1}^2 + \tau^2\sigma_{a+1}^2$, which is equivalent to $(1-\tau^2)\sigma_a^2 > \sigma_{a+1}^2$. This implies that $\sigma_b^2 < (1-\tau^2)^{b-a}\sigma_a$; the second inequality follows from $w_a^2 > \tau^2\sigma_a^2$. □

Fix a parameter $Z > 1$. We divide the list of weights $w_1, \ldots, w_\ell$ into "Z-blocks" of consecutive weights as follows. The first Z-block B_1 is $w_1, \ldots, w_{k_1}$ where k_1 is defined to be the first index such that w_1 (the largest weight in the block) is "large" relative to σ_{k_1+1} (the total "tail weight" of all weights after the Z-block) in the following sense:

$$|w_1| > Z \cdot \sigma_{k_1+1}.$$

Similarly for $i = 2, 3, \ldots$ the ith Z-block B_i is $w_{k_{i-1}+1}, \ldots, w_{k_i}$ where k_i is the first index such that

$$|w_{k_{i-1}+1}| > Z \cdot \sigma_{k_i+1}.$$

The following lemma says each Z-block must be relatively short prior to the critical index:

LEMMA 4.4. *Suppose that the ith Z-block B_i is such that $k_{i-1} + 1 + m \leq \ell$, where*

$$m \stackrel{def}{=} \frac{1}{\tau^2} \cdot \ln(Z^2/\tau^2). \tag{4.2}$$

Then B_i is of length at most m.

Proof. Suppose that the length $|B_i|$ of the ith Z-block were more than m. Applying Lemma 4.3 with $b - a = m$, we have

$$\sigma^2_{k_{i-1}+1+m} < (1-\tau^2)^m w^2_{k_{i-1}+1}/\tau^2 \leq e^{-\tau^2 m} w^2_{k_{i-1}+1}/\tau^2.$$

But by the assumption that the ith Z-block is longer than m, we also have

$$w^2_{k_{i-1}+1} \leq Z^2 \sigma^2_{k_{i-1}+1+m}.$$

Combining these inequalities and plugging in our expression for m we get a contradiction. □

An easy consequence is that if the critical index is large, then there must be many blocks prior to it:

COROLLARY 4.5. *For $t \geq 1$, suppose that the τ-critical index ℓ is at least tm, where m is defined as in (4.2). Then $k_t \leq tm$, i.e. there are at least t complete Z-blocks by the (tm)-th weight.*

4.2. Block structure and concentration of the random variable $w \cdot x$.

Let $f(x) = \mathrm{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n)$ be a threshold function with $|w_1| \geq \cdots \geq |w_n| > 0$, and let $B_1, B_2, \ldots$ be the Z-blocks for w as defined in the previous subsection. In this subsection we prove the following lemma which is a slight variant of a similar result in [47]. Intuitively the lemma says that if a weight vector v has "many" blocks, then for any $w_0 \in \mathbb{R}$, only an exponentially small fraction of points $x \in \{-1,1\}^n$ will have a "small" margin for the threshold function $\mathrm{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n)$. As we show in the next subsection, Theorem 4.2 will be an easy consequence of this lemma.

LEMMA 4.6. *Fix a value t such that there exist at least t complete Z-blocks $B_1, \ldots, B_t$ in the weight vector w. Then for any $w_0 \in \mathbb{R}$, we have*

$$\mathbf{Pr}[|w_0 + w_1x_1 + \cdots + w_nx_n| \leq \sigma_{k_t+1} \cdot (Z/6)] \leq 2^{-t} + 2te^{-Z^2/72}.$$

Here the probability is taken over a uniform random choice of x from $\{-1,1\}^n$.

We first give some necessary preliminary results and then prove Lemma 4.6. Our approach follows that of [47] with slight modifications.

Let us view the choice of a uniform random assignment x to the variables in Z-blocks $B_1, \ldots, B_t$ as taking place in successive stages, where in the ith stage values are assigned to the variables in the ith Z-block B_i. Immediately after the ith stage, some value—call it ξ_i—has been determined for $w_0 + w_1x_1 + \cdots + w_{k_i}x_{k_i}$. The following simple lemma shows that if ξ_i is too far from 0, then it is unlikely that the remaining variables $x_{k_i+1}, \ldots, x_n$ will come out in such a way as to make the final sum close to 0.

LEMMA 4.7. *For any value $A > 0$ and any $1 \leq i \leq t$, if $|\xi_i| \geq 2\sigma_{k_i+1}\sqrt{2\ln(2/A)}$, then we have*

$$\mathbf{Pr}_{x_{k_i+1},\ldots,x_n}[|w_0 + w_1x_1 + \cdots + w_nx_n| \leq \sigma_{k_i+1}\sqrt{2\ln(2/A)}] \leq A. \qquad (4.3)$$

Proof. By the lower bound on $|\xi_i|$ in the hypothesis of the lemma, it can only be the case that $|w_0 + w_1x_1 + \cdots + w_nx_n| \leq \sigma_{k_i+1}\sqrt{2\ln(2/A)}$ if

$$|w_{k_i+1}x_{k_i+1} + \cdots + w_nx_n| \geq \sigma_{k_i+1}\sqrt{2\ln(2/A)}. \qquad (4.4)$$

We now recall the Hoeffding bound (see e.g. [12]), which says that for any $0 \neq v \in \mathbb{R}^r$ and any $\gamma > 0$, we have $\mathbf{Pr}_{x\in\{-1,1\}^r}[|v_1x_1 + \cdots + v_rx_r| \leq \gamma\sqrt{v_1^2 + \cdots + v_r^2}] \leq 2e^{-\gamma^2/2}$.

Since $w_{k_i+1}^2 + \cdots + w_n^2 = \sigma_{k_i+1}^2$, this Hoeffding bound implies that the probability of (4.4) is at most

$$2e^{-(\sqrt{2\ln(2/A)})^2/2} = A. \qquad \square$$

We henceforth fix A to be $A \overset{\text{def}}{=} 2e^{-Z^2/72}$, so we have $6\sqrt{2\ln(2/A)} = Z$. We now show that regardless of the value of ξ_{i-1}, we have $|\xi_i| \le 2\sigma_{k_i+1}(Z/6)$ with probability at most $1/2$ over the choice of values for variables in block B_i in the ith stage.

LEMMA 4.8. *For any $\xi_{i-1} \in \mathbb{R}$, we have*

$$\mathbf{Pr}_{x_{k_{i-1}+1},\ldots,x_{k_i}}[|\xi_i| \le 2\sigma_{k_i+1}(Z/6) \mid \xi_{i-1}] \le 1/2.$$

Proof. Since ξ_i equals $\xi_{i-1} + (w_{k_{i-1}+1}x_{k_{i-1}+1} + \cdots + w_{k_i}x_{k_i})$, we have $|\xi_i| \le 2\sigma_{k_i+1}(Z/6)$ 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

$$[I_L, I_R] \overset{\text{def}}{=} [-\xi_{i-1} - 2\sigma_{k_i+1}(Z/6), -\xi_{i-1} + 2\sigma_{k_i+1}(Z/6)]$$

of width $\frac{2}{3}\sigma_{k_i+1}Z$.

First suppose that $0 \notin [I_L, I_R]$, i.e. the whole interval has the same sign. If this is the case then $\mathbf{Pr}[w_{k_{i-1}+1}x_{k_{i-1}+1} + \cdots + w_{k_i}x_{k_i} \in [I_L, I_R]] \le \frac{1}{2}$ 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 $\sigma_{k_i+1} \le |w_{k_{i-1}+1}|/Z$, and consequently we have that the width of the interval $[I_L, I_R]$ is at most $\frac{2}{3}|w_{k_{i-1}+1}|$. 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 $[I_L - w_{k_{i-1}+1}x_{k_{i-1}+1}, I_R - w_{k_{i-1}+1}x_{k_{i-1}+1}]$. (Note that in the special case where $k_i = k_{i-1} + 1$, the value $w_{k_{i-1}+2}x_{k_{i-1}+2} + \cdots + w_{k_i}x_{k_i}$ which must hit the target interval is simply 0.) 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 $x_{k_{i-1}+2}, \ldots, x_{k_i}$) 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 $\frac{1}{2}$. $\square$

In order to have $|w_0 + w_1x_1 + \cdots + w_nx_n| \le \sigma_{k_t+1}\sqrt{2\ln(2/A)}$, it must be the case that either

(i) each $|\xi_i| < 2\sigma_{k_i+1}\sqrt{2\ln(2/A)}$ for $i = 1, \ldots, t$; or

(ii) for some $1 \le i \le t$ we have $|\xi_i| \ge 2\sigma_{k_i+1}\sqrt{2\ln(2/A)}$ but nonetheless $|w_0 + w_1x_1 + \cdots + w_nx_n| < \sigma_{k_i+1}\sqrt{2\ln(2/A)}$.

Lemma 4.8 gives us that the probability of (i) is at most $(1/2)^t = 2^{-t}$, and Lemma 4.7 with the union bound gives us that the probability of (ii) is at most $t \cdot A$. This proves Lemma 4.6.

4.3. Proof of Theorem 4.2. Let $Z = 12\sqrt{t}$. We take $m = \frac{1}{\tau^2} \cdot \ln(Z^2/\tau^2)$ as in (4.2), and we have $k = tm + 1$. With these choices the condition $\ell \ge k$ of Theorem 4.2 together with Corollary 4.5 implies that there are at least t complete Z-blocks in the weight vector w. Thus we may apply Lemma 4.6, and we have that

$$\mathbf{Pr}[|w_0 + w_1x_1 + \cdots + w_nx_n| \le \sigma_{k_t+1} \cdot 2\sqrt{t}] \le 2^{-t} + 2te^{-2t} \le O(2^{-t}).$$

Now we further observe that since there are in fact t complete Z-blocks prior to the kth weight, we have $k_t + 1 \le k$ and hence $\sigma_{k_t+1} \ge \sigma_k$, so the above inequality implies

$$\mathbf{Pr}[|w_0 + w_1x_1 + \cdots + w_nx_n| \le \sqrt{t} \cdot \sigma_k] \le O(2^{-t}).$$

This is the desired conclusion of Theorem 4.2.

4.4. Extension of Theorem 4.2. The same proof with a slightly different choice of Z (taking $Z = O(1)t^C$) in fact gives us the following significantly stronger version of Theorem 4.2; however this stronger version is not more useful for our purposes:

THEOREM 4.9. *In the setting of Theorem 4.2, let $C \geq 1/2$ be another parameter, and suppose we instead define*

$$k = \left\lceil O(1)\frac{t}{\tau^2}\ln\left(\frac{t^C}{\tau}\right)\right\rceil.$$

Then if $\ell \geq k$,

$$\mathop{\mathbf{Pr}}_x[|w_0 + w_1x_1 + \cdots + w_nx_n| \leq t^C \cdot \sigma_k] \leq O(2^{-t}).$$

5. Approximating threshold functions using not-too-large head weights.

The main result of this section is a lemma which roughly says that any threshold function f can be approximated by a threshold function f' in which the 2-norm of the tail weights, σ_k, is at least an $\Omega(1)$ fraction of the head weights. This is important so that the Gaussian random variable to which the tail part is close has $\Omega(1)$ variance and thus sufficiently good anticoncentration.

LEMMA 5.1. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any threshold function, $f(x) = \operatorname{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n)$ (recall that we assume $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$). Let $0 < \epsilon < 1/2$ and $1 \leq k \leq n$ be parameters, and write $\sigma_k \overset{def}{=} \sqrt{\sum_{j\geq k} w_j^2}$. Assuming $\sigma_k > 0$, there are numbers $v_0, \ldots, v_{k-1}$ satisfying*

$$|v_i| \leq k^{(k+1)/2} \cdot \sqrt{3\ln(2/\epsilon)} \cdot \sigma_k \tag{5.1}$$

such that the threshold function $f' : \{-1,1\}^n \to \{-1,1\}$ defined by

$$f'(x) = \operatorname{sgn}(v_0 + v_1x_1 + \cdots + v_{k-1}x_{k-1} + w_kx_k + \cdots + w_nx_n)$$

satisfies $\operatorname{dist}(f, f') \leq \epsilon$. *One may further ensure that $|v_1| \geq |v_2| \geq \cdots \geq |v_{k-1}| \geq |w_k|$ and that* $\operatorname{sgn}(v_i) = \operatorname{sgn}(w_i)$ *for all i.*

Before proving this lemma, let us give an illustration. Consider the threshold function

$$f(x) = \operatorname{sgn}(nx_1 + nx_2 + x_3 + \cdots + x_n), \tag{5.2}$$

with $k = 3$. The tail weights here have $\sigma_3 = \sqrt{n-2}$, which of course is not a constant fraction of the two head weights, n. Further, this cannot be fixed just by choosing a different weights-based representation of the same function f. What Lemma 5.1 shows here is that we can shrink the head weights from n all the way down to $\Theta(\sqrt{\ln(1/\epsilon)})\sqrt{n}$ without changing the function on more than an ϵ fraction of points (this heavily uses the fact that the tail acts like a Gaussian with standard deviation $\sqrt{n-2}$). Then indeed σ_3 is an $\Omega(f(\epsilon))$ fraction of the head weights for a function $f(\epsilon)$ that is independent of n, as desired.

We now give the proof of Lemma 5.1, a modification of the classic argument of [40] which bounds the weights required for exact representation of any threshold function.

Proof. We will first prove the theorem without the extra constraints $|v_1| \geq |v_2| \geq \cdots \geq |v_{k-1}| \geq |w_k|$ and $\operatorname{sgn}(v_i) = \operatorname{sgn}(w_i)$. At the end of the proof we will show how these constraints can also be ensured.

Let $h : \{-1,1\}^{k-1} \to \mathbb{R}$ denote the head of f,

$$h(x) = w_0 + w_1x_1 + \cdots + w_{k-1}x_{k-1}.$$

Consider the system $\mathcal{S}$ of 2^{k-1} linear equations in k unknowns named $u_0, \dots, u_{k-1}$: for each $x \in \{-1,1\}^{k-1}$ we include the equation

$$u_0 + u_1x_1 + \cdots + u_{k-1}x_{k-1} = h(x).$$

Of course, the linear system $\mathcal{S}$ is satisfiable, since $(u_0, \dots, u_{k-1}) = (w_0, \dots, w_{k-1})$ is a solution.

Let C be defined by

$$C = \sqrt{3\ln(2/\epsilon)} \cdot \sigma_k,$$

and consider the system $\mathcal{LP}$ of 2^{k-1} linear *inequalities* over unknowns $u_0, \dots, u_{k-1}$: for each $x \in \{-1,1\}^{k-1}$ we include the (in)equality

$$u_0 + u_1x_1 + \cdots + u_{k-1}x_{k-1} \begin{cases} \geq C & \text{if } h(x) \geq C, \\ = h(x) & \text{if } |h(x)| < C, \\ \leq -C & \text{if } h(x) \leq -C. \end{cases} \tag{5.3}$$

We have that $\mathcal{LP}$ is feasible, since it is a relaxation of the satisfiable system $\mathcal{S}$.

Now we use the following standard result from the theory of linear inequalities, which is a straightforward consequence of Cramer's rule and is implicit in several works (see e.g. the proof at the start of Section 3 of [24]):

LEMMA 5.2. *Let $\mathcal{LP}$ denote a feasible linear program over k variables $u_0, \dots, u_{k-1}$ in which the constraint matrix has all entries from $\{-1, 0, 1\}$ and the right-hand side has all entries at most C in absolute value. Then there is a feasible solution $(v_0, \dots, v_{k-1})$ in which*

$$|v_i| \leq k^{(k+1)/2} \cdot C$$

for each i.

This implies that there is a feasible solution $(u_0, \dots, u_{k-1}) = (v_0, \dots, v_{k-1})$ to $\mathcal{LP}$ in which the numbers v_i are not too large in magnitude: specifically, using Lemma 5.2 we may obtain

$$|v_i| \leq k^{(k+1)/2} \cdot C. \tag{5.4}$$

We now show that the threshold function

$$f'(x) = \operatorname{sgn}(v_0 + v_1x_1 + \cdots + v_{k-1}x_{k-1} + w_kx_k + \cdots w_nx_n)$$

satisfies $\operatorname{dist}(f, f') \leq \epsilon$.

Given $x \in \{-1,1\}^n$, let us abuse notation by writing

$$h(x) = h(x_1, \dots, x_{k-1}) = w_0 + w_1x_1 + \cdots + w_{k-1}x_{k-1};$$

let us also write

$$h'(x) = v_0 + v_1x_1 + \cdots + v_{k-1}x_{k-1}$$

for the head of f' and

$$t(x) = \sum_{j \geq k} w_j x_j$$

for the tail, which is common to both f and f'. Now if x is any input for which $|h(x)| < C$ then we have $h(x) = h'(x)$ by construction, and hence $f(x) = f'(x)$. Thus in order for $f(x)$ to disagree with $f(x')$ it must at least be the case that $|h(x)| \geq C$. Moreover, it must also be the case that $|t(x)| \geq C$, for otherwise $\mathrm{sgn}(h(x) + t(x))$ will equal $\mathrm{sgn}(h'(x) + t(x))$, because $h(x)$ and $h'(x)$ have the same sign by construction. But the Hoeffding bound implies that

$$\mathop{\mathbf{Pr}}_x[|t(x)| \geq C] \leq \mathop{\mathbf{Pr}}_x[|t(x)| \geq \sqrt{2\ln(2/\epsilon)} \cdot \sigma_k] \leq 2e^{-\ln(2/\epsilon)} = \epsilon.$$

Hence indeed $\mathbf{Pr}[f(x) \neq f'(x)] \leq \epsilon$, as desired.

Finally, we complete the proof by showing how to ensure the extra constraints $|v_1| \geq |v_2| \geq \cdots \geq |v_{k-1}| \geq |w_k|$ and $\mathrm{sgn}(v_i) = \mathrm{sgn}(w_i)$. First, the constraints $\mathrm{sgn}(u_i) = \mathrm{sgn}(w_i)$ can be added into $\mathcal{LP}$—by this we mean adding constraints like $u_1 \geq 0$, $u_2 \leq 0$, etc. Next, the constraints

$$\begin{aligned} \mathrm{sgn}(w_1)u_1 &\geq \mathrm{sgn}(w_2)u_2 \\ \mathrm{sgn}(w_2)u_2 &\geq \mathrm{sgn}(w_3)u_3 \\ &\cdots \\ \mathrm{sgn}(w_{k-2})u_{k-2} &\geq \mathrm{sgn}(w_{k-1})u_{k-1} \end{aligned}$$

can be added into $\mathcal{LP}$; again, these are constraints like $-u_i \geq u_{i+1}$. Finally, we can add the constraint $\mathrm{sgn}(w_{k-1})u_{k-1} \geq |w_k|$. Of course, $\mathcal{LP}$ remains feasible after the addition of all of these constraints, since $(u_0, \ldots, u_{k-1}) = (w_0, \ldots, w_{k-1})$ is still a solution. It remains to show that there is still a solution satisfying the bounds in (5.4). But this still follows from Lemma 5.2: the added constraints only have coefficients in $\{-1, 0, 1\}$, and the added right-hand side entries are all 0, except for the last, which is $|w_k| \leq \sigma_k \leq C$. □

6. Every threshold function is close to a threshold function for which few points have small margin. In this subsection we show how to combine Theorem 4.2 and Lemma 5.1 to establish the following:

THEOREM 6.1. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any threshold function and let $0 < \tau < 1/2$. Then there is a threshold function $f' : \{-1,1\}^n \to \{-1,1\}$ with* $\mathrm{dist}(f, f') \leq \epsilon$ *satisfying* $\mathbf{Pr}_x[\mathrm{marg}(f', x) \leq \rho] \leq O(\tau)$,

$$\textit{where} \qquad \epsilon = \epsilon(\tau) = 2^{-2^{O(\log^3(1/\tau)/\tau^2)}} \qquad \textit{and} \qquad \rho = \rho(\tau) = 2^{-O(\log^3(1/\tau)/\tau^2)}.$$

Our main structural results about margins, Theorem 3.1, is simply a rephrasing of the above theorem. Hence proving Theorem 6.1 completes the proof of Theorem 1.6, the "first ingredient" in our solution to the Chow Parameters Problem.

The plan for the proof of Theorem 6.1 follows the intuition described in the beginning of Section 4. We consider the location of the τ-critical index of f. Case 1 is that it occurs quite early. In that case, the resulting tail acts like a Gaussian (up to error τ), and hence we can get a good anticoncentration bound so long as the tail's variance is large enough. To ensure this, we alter f at the beginning of the argument using Lemma 5.1, which yields tail weights with total variance lower bounded by a function that depends only on τ. Case 2 is that the critical index occurs late. In this case we get anticoncentration by appealing to Theorem 4.2. We again use Lemma 5.1 so that the σ_k parameter is not too small.

We now give the formal proof.

Proof. (Theorem 6.1) We intend to apply Theorem 4.2 in Case 2 with its t parameter set to $\log(1/\tau)$, so that the anticoncentration is $O(\tau)$. Thus we will need to ensure the τ-critical index parameter ℓ is at least

$$k \overset{\text{def}}{=} \left\lceil O(1)\frac{\log(1/\tau)}{\tau^2}\ln\left(\frac{\log(1/\tau)}{\tau}\right)\right\rceil. \tag{6.1}$$

To that end, fix a weights-based representation of f,

$$f(x) = \operatorname{sgn}(w_0 + w_1x_1 + \cdots + w_nx_n),$$

where we may assume that $|w_1| \geq |w_2| \geq \cdots \geq |w_n| > 0$. Write $\sigma_k = \sqrt{\sum_{j\geq k} w_j^2}$, and observe that $\sigma_k > 0$ since each $w_i \neq 0$. Now apply Lemma 5.1, with its parameter ϵ set to $2^{-k^{O(k)}}$. This yields a new threshold function

$$f'(x) = \operatorname{sgn}(v_0 + v_1x_1 + \cdots + v_{k-1}x_{k-1} + w_kx_k + \cdots w_nx_n), \tag{6.2}$$

where each v_i satisfies

$$|v_i| \leq k^{O(k)} \cdot \sigma_k, \tag{6.3}$$

and also $|v_1| \geq |v_2| \geq \cdots \geq |v_{k-1}| \geq |w_k|$. This f' has $\operatorname{dist}(f, f') \leq \epsilon = 2^{-k^{O(k)}}$.

To analyze $\operatorname{marg}(f', x)$, let us normalize the weights of f' by dividing each weight by $\sqrt{v_0^2 + \cdots + v_{k-1}^2 + w_k^2 + \cdots + w_n^2}$. We thus may write

$$f'(x) = \operatorname{sgn}(u_0 + u_1x_1 + \cdots + u_{k-1}x_{k-1} + u_kx_k + \cdots u_nx_n),$$

where $\sum_{j\geq 0} u_j^2 = 1$. Equation (6.2) implies that for each of the k values $i = 0, \ldots, k-1$ we have that v_i^2 is at most $k^{O(k)}$ times as large as $w_k^2 + \cdots + w_n^2$. Letting σ_i' denote $\sqrt{\sum_{j\geq i} u_j^2}$ and recalling that $\sum_{j\geq 0} u_j^2 = 1$, this is easily seen to imply that

$$\sigma_k' \geq k^{-O(k)}. \tag{6.4}$$

Recalling that we still have $|u_1| \geq |u_2| \geq \cdots \geq |u_n| > 0$, let ℓ be the τ-critical index for $u_1, \ldots, u_n$, and consider two cases:

Case 1: $\ell < k$. In this case, consider any fixed choice for $x_1, \ldots, x_{\ell-1}$ and write $h = u_0 + u_1x_1 + \cdots + u_{\ell-1}x_{\ell-1}$. Using the definition of τ-critical index and applying the Berry-Esseen Corollary 2.8 to $u_\ell x_\ell + \cdots + u_nx_n$ we get

$$\mathop{\mathbf{Pr}}_{x_\ell,\ldots,x_n}[-h - \gamma \leq u_\ell x_\ell + \cdots + u_nx_n \leq -h + \gamma] \leq \frac{2\gamma}{\sigma_\ell'} + 2\tau,$$

for any choice of $\gamma \geq 0$. Taking $\gamma = \tau\sigma_\ell' \geq \tau\sigma_k'$ we conclude

$$\mathop{\mathbf{Pr}}_{x}[\operatorname{marg}(f', x) \leq \tau\sigma_k'] \leq 4\tau.$$

Case 2: $\ell \geq k$. In this case we apply Theorem 4.2, with its parameter t set to $\log(1/\tau)$, as described at the beginning of the proof. With k defined as in (6.1), we conclude

$$\mathop{\mathbf{Pr}}_{x}[\mathrm{marg}(f', x) \leq \sqrt{\log(1/\tau)} \cdot \sigma'_k] \leq O(\tau).$$

Combining the results of the two cases and using $\sigma'_k \geq k^{-O(k)}$ from (6.4), we conclude that we always have

$$\mathop{\mathbf{Pr}}_{x}[\mathrm{marg}(f', x) \leq \tau k^{-O(k)}] \leq O(\tau).$$

Now it only remains to observe that by definition (6.1) of k,

$$k^{-O(k)} = 2^{-O(\log^3(1/\tau)/\tau^2)}.$$

Hence we have that

$$\mathrm{dist}(f, f') \leq 2^{-k^{O(k)}} \leq \epsilon(\tau)$$

and

$$\tau k^{-O(k)} \geq \tau 2^{-O(\log^3(1/\tau)/\tau^2)} \geq \rho(\tau). \qquad \square$$

7. Second ingredient: using Chow Parameters as weights for tail variables. We begin this section with some informal motivation for and description of our "second ingredient".

We first recall that every threshold function f is unate; this means that for every i, f is either monotone increasing or monotone decreasing as a function of its i-th coordiante. A well-known consequence of unateness is that the magnitude of the Fourier coefficient $|\hat{f}(i)|$ is equal to the *influence* of the variable x_i on f; i.e. $\mathbf{Pr}[f(x) \neq f(y)]$ where x is drawn uniformly from $\{-1, 1\}^n$ and y is x with the ith bit flipped. As done in the "first ingredient", it is natural to group together the high-influence variables, forming the "head" indices of f. We refer to the remaining indices as the "tail" indices. Note that an algorithm for the Chow Parameters problem can do this grouping, since it is given the $\hat{f}(i)$'s.

The following theorem states that any threshold function f is either already close to a junta over the head indices, or is close to a threshold function obtained by replacing the tail weights with (suitably scaled versions of) the tail Chow Parameters. (We have made no effort to optimize the precise polynomial dependence of $\tau(\epsilon)$ on ϵ.)

THEOREM 7.1. *There is a polynomial function $\tau(\epsilon) = \mathrm{poly}(\epsilon)$ such that the following holds: Let f be a Boolean threshold function over head indices H and tail indices T,*

$$f(x) = \mathrm{sgn}\left(v_0 + \sum_{i \in H} v_i x_i + \sum_{i \in T} w_i x_i\right),$$

and let $0 < \epsilon < 1/2$. Assume that H contains all indices i such that $|\widehat{f}(i)| \geq \tau(\epsilon)^2$. Then one of the following holds:

(i) *f is $O(\epsilon)$-close to a junta over H; or,*

(ii) *we can normalize the weights so that $\sum_{i \in T} w_i^2 = 1$, in which case f is $O(\epsilon)$-close to the Boolean threshold function*

$$f'(x) = \operatorname{sgn}\left(v_0 + \sum_{i \in H} v_i x_i + \sum_{i \in T} \frac{\widehat{f}(i)}{\sigma} x_i \right),$$

where σ denotes $\sqrt{\sum_{i \in T} \widehat{f}(i)^2}$.

We remark that it can be shown that statement (ii) in this theorem in fact *always* holds (assuming $\sigma \neq 0$), even when f is close to a junta. We omit the proof as our overall results do not need this strengthening. We also remark that by Parseval's identity, one can take the set $H \subset [n]$ in the theorem to be the $1/\tau(\epsilon)^4 = \operatorname{poly}(1/\epsilon)$ indices for which the weights are the largest.

Theorem 7.1 has the following immediate corollary:

Corollary 7.2. *Under the hypotheses of Theorem 7.1, there exists a threshold function $f'(x) = \operatorname{sgn}(v_0 + v_1 x_1 + \cdots + v_n x_n)$ which is $O(\epsilon)$-close to f in which $v_i = \widehat{f}(i)$ for all $i \notin H$.*

Proof. In case (i), Lemma 2.3 implies that f is $O(\epsilon)$-close to the junta $\operatorname{sgn}(v_0 + \sum_{i \in H} v_i x_i)$. We can put this junta over H into the desired format by scaling the weights $\{v_i\}_{i \in H}$ so large that the weights $\{v_i = \widehat{f}(i)\}_{i \notin H}$ are collectively irrelevant. Otherwise, we are in case (ii) and we can scale all weights by σ. □

Theorem 7.1 suggests an approach to constructing a "small" list of candidate threshold functions for the Chow Parameters problem. We take H to be all indices with Chow Parameter of magnitude at least $\tau(\epsilon)^2$; as mentioned, there are at most $1/\tau(\epsilon)^4$ such indices. If f is close to a junta over H (case (i)), we can construct a list of candidates that will contain such a close-to-f junta by simply enumerating all junta threshold functions over H; intuitively this is a "small" number of candidates since $|H|$ is "small." On the other hand, if we are in case (ii) then simply using the Chow Parameters as the tail weights almost gives us a threshold function which is ϵ-close to f—it remains only to fill in the $|H|$ unknown head weights.

We deal with the unknown head weights via the following extension of Theorem 7.1, which shows that it is enough to consider head weights with bounded precision within a bounded range:

Theorem 7.3. *Statement (ii) in Theorem 7.1 can be replaced by the following:*

(ii) f is $O(\epsilon)$-close to a Boolean threshold function f' of the form

$$f'(x) = \operatorname{sgn}\left(u_0 + \sum_{i \in H} u_i x_i + \sum_{i \in T} \frac{\widehat{f}(i)}{\sigma} x_i \right),$$

where the weights u_i are integer multiples of $\sqrt{\tau(\epsilon)}/|H|$ with magnitude at most $2^{O(|H| \log |H|)} \sqrt{\ln(1/\tau(\epsilon))}$.

Theorem 7.3 is sufficient if we are given the exact values of the Chow Parameters, but as described in Section 1.2 we consider the more difficult scenario in which we are only given approximations to the Chow Parameters (this is the scenario required for 1-RFA learning). Thus we want an extension of Theorem 7.3 which requires only that the input vector be close to the Chow Parameters of f. We prove the following:

Theorem 7.4. *Theorem 7.3 continues to hold if, instead of using the vector $\vec{\gamma} = [\widehat{f}(i)]_{i \in T}$ for the (pre-scaled) tail weights, we used a vector $\vec{\alpha}$ satisfying*

$$\|\vec{\alpha} - \vec{\gamma}\| \leq \Omega(\epsilon^4). \tag{7.1}$$

Since Theorem 7.4 is our ultimate goal we prove it directly; this is the object of Section 8 The proof builds on ideas developed in the proof of correctness of the poly$(1/\epsilon)$-query testing algorithm for the class of threshold functions given by Matulef et al. [37]. In the remainder of this section we give a sketch of the proof of Theorem 7.4, and also develop some technical geometric lemmas needed in the proof.

Proof sketch. The "completeness" analysis of [37] (together with geometric lemmas) helps us show that if f is far from a junta over H, then all restrictions of the head indices give rise to Chow vectors (of the different restrictions of f) that are mutually "approximately parallel":

DEFINITION 7.5. *We say two vectors $\vec{\beta}$ and $\vec{\gamma}$ are η-*approximately parallel *if*

$$\|\vec{\beta}\| \cdot \|\vec{\gamma}\| - \vec{\beta} \cdot \vec{\gamma} \leq \eta. \tag{7.2}$$

The completeness argument of [37] also gives us that there is a set of weights for the head indices lying in the required range and with the required precision, that are compatible in a certain technical sense with all the restrictions of the head. Additional geometric arguments show that the *average* of the Chow Vectors of the restrictions—which equals the tail of the Chow Vector of f itself—is a "long" vector which is itself approximately parallel to the Chow vectors of the restrictions. Next, these properties, along with the "soundness" analysis of [37], are used to show that replacing the tail weights with the tail Chows of f causes very little error for each restriction to the head indices. Finally, the "compatible" head weights from above are used to obtain an overall high-accuracy approximator for f whose head weights have the stated bounded magnitude and granularity and whose tail weights are the tail Chow parameters of f.

We now state and prove the required geometric lemmas. It is straightforward to characterize the vectors that are η-approximately parallel to a fixed vector $\vec{\beta}$:

PROPOSITION 7.6. *Fix a nonzero vector $\vec{\beta}$ and a value $\eta > 0$. The set $S = \{$all vectors $\vec{\gamma}$ that are η-approximately parallel to $\vec{\beta}\}$ is a closed solid paraboloid of revolution along an axis in the direction $\vec{\beta}$, with vertex located at*

$$\frac{-\eta\vec{\beta}}{2\|\vec{\beta}\|^2}.$$

Proof. We have $\vec{\gamma} \in S$ if and only if $\|\vec{\gamma}\| \leq \frac{\eta}{\|\vec{\beta}\|} + \gamma \cdot \frac{\vec{\beta}}{\|\vec{\beta}\|}$, i.e. γ is closer to the origin than to the hyperplane $\frac{\eta}{\|\vec{\beta}\|} + \frac{\vec{\beta}}{\|\vec{\beta}\|} \cdot x = 0$. Thus S is the paraboloid of revolution whose focus is the origin and whose directrix plane is $\frac{\eta}{\|\vec{\beta}\|} + \frac{\vec{\beta}}{\|\vec{\beta}\|} \cdot x = 0$; this paraboloid has axis vector in the direction $\vec{\beta}$ and has vertex $\frac{-\eta\vec{\beta}}{2\|\vec{\beta}\|^2}$ as claimed. □

The next lemma gives sufficient conditions for two vectors $\vec{\beta}, \vec{\beta}'$ to be approximately parallel:

LEMMA 7.7. *Let $\vec{\beta}$ and $\vec{\beta}'$ be vectors with $\|\vec{\beta}\|, \|\vec{\beta}'\| \leq 1$ and let W, W' be numbers satisfying*

$$\big|\|\vec{\beta}\|^2 - W\big| \leq \tau^{1/12}, \quad \big|\|\vec{\beta}'\|^2 - W'\big| \leq \tau^{1/12}, \tag{7.3}$$

$$\big|(\vec{\beta} \cdot \vec{\beta}')^2 - W \cdot W'\big| \leq \tau^{1/12}, \tag{7.4}$$

where $0 < \tau < 1/2$. *Assume also that* $\vec{\beta} \cdot \vec{\beta}' \geq 0$. *Then* $\vec{\beta}$ *and* $\vec{\beta}'$ *are* $O(\tau^{1/36})$-*approximately parallel.*

Proof. If either $\|\vec{\beta}\| \leq \tau^{1/36}$ or $\|\vec{\beta}'\| \leq \tau^{1/36}$ then the claim holds easily:

$$\|\vec{\beta}\| \cdot \|\vec{\beta}'\| - \vec{\beta} \cdot \vec{\beta}' \leq 1 \cdot \tau^{1/36} - 0 = \tau^{1/36}.$$

Otherwise, substituting (7.3) into (7.4) (and using $0 \leq W, W' \leq 1 + \tau^{1/12} = O(1)$) we get

$$\big|(\vec{\beta} \cdot \vec{\beta}')^2 - \|\vec{\beta}\|^2 \cdot \|\vec{\beta}'\|^2\big| \leq O(\tau^{1/12}).$$

Now we apply Fact 2.5 and use $\|\vec{\beta}\| \cdot \|\vec{\beta}'\| \geq \tau^{2/36}$ to conclude

$$\big||\vec{\beta} \cdot \vec{\beta}'| - \|\vec{\beta}\| \cdot \|\vec{\beta}'\|\big| \leq O\Big(\frac{\tau^{1/12}}{\tau^{1/18}}\Big) = O(\tau^{1/36}).$$

Since $|\vec{\beta} \cdot \vec{\beta}'| = \vec{\beta} \cdot \vec{\beta}'$ by assumption, the proof is complete. □

Roughly speaking, the next lemma says that in a group of vectors that are all mutually approximately parallel, if a "large" fraction of the vectors are "long" then the average of all the vectors in the group must also be fairly long.

LEMMA 7.8. *Suppose* $\{\beta_\pi\}_{\pi \in \Pi}$ *is a collection of vectors which are mutually* η-*approximately parallel. Write* Π_ϵ *for those* π *such that* $\|\beta_\pi\| \geq \epsilon$ *and write* $\lambda = |\Pi_\epsilon|/|\Pi|$. *Assume that* $\eta \leq (3/4)\epsilon^2$. *Then* $\vec{\gamma} = \operatorname{avg}_{\pi \in \Pi}[\beta_\pi]$ *satisfies*

$$\|\vec{\gamma}\| \geq \lambda\epsilon/2 - \eta/\epsilon.$$

Proof. Write $\gamma = \lambda\vec{g} + (1-\lambda)\vec{e}$, where $\vec{g} = \operatorname{avg}_{\pi \in \Pi_\epsilon}[\beta_\pi]$ and $\vec{e}$ is the average of the remaining β_π's. We have

$$\|\vec{g}\|^2 = \frac{1}{|\Pi_\epsilon|^2} \sum_{\pi,\pi' \in \Pi_\epsilon} \beta_\pi \cdot \beta_{\pi'} \geq \frac{1}{|\Pi_\epsilon|^2} \sum_{\pi,\pi' \in \Pi_\epsilon} (\|\beta_\pi\| \cdot \|\beta_{\pi'}\| - \eta) \geq \epsilon^2 - \eta \geq \epsilon^2/4,$$

where we used the fact that each pair $\beta_\pi, \beta_{\pi'}$ is η-approximately parallel and also $\eta \leq (3/4)\epsilon^2$. On the other hand, by the convexity of the paraboloid from Proposition 7.6, $\vec{g}$ is η-approximately parallel to all β_π for $\pi \notin \Pi_\epsilon$; hence, $\vec{e}$ is η-approximately parallel to $\vec{g}$. Given this and Proposition 7.6, the least value that $\|\lambda\vec{g} + (1-\lambda)\vec{e}\|$ can take is if $\vec{e}$ were $-\frac{\eta}{2\|\vec{g}\|^2}\vec{g}$. Thus we have a lower bound

$$\|\gamma\| \geq \lambda\|\vec{g}\| - (1-\lambda)\frac{\eta}{2\|\vec{g}\|} \geq \lambda\|\vec{g}\| - \frac{\eta}{2\|\vec{g}\|}.$$

The above quantity is an increasing function of $\|\vec{g}\|$, so using $\|\vec{g}\| \geq \sqrt{\epsilon^2/4} = \epsilon/2$ we get

$$\|\gamma\| \geq \lambda\epsilon/2 - \eta/\epsilon,$$

as claimed. □

8. Proof of Theorem 7.4 via Property Testing. As mentioned, our proof of Theorem 7.4 builds on the efficient property testing algorithm for threshold functions developed in [37]. We are able to use in a mostly black-box fashion their "soundness" theorem:

THEOREM 8.1. *Let $g : \{-1,1\}^n \to \{-1,1\}$ be a θ-regular Boolean threshold function and write $\vec{\beta} = (\widehat{g}(1), \ldots, \widehat{g}(n))$, so $|\vec{\beta}(i)| \le \theta$ for each i. Let $\vec{\gamma}$ be a vector with $\|\vec{\gamma}\| = 1$ which is $\theta^{1/9}$-approximately parallel to $\vec{\beta}$. Assume also $|\vec{\gamma}(i)| \le \theta^{1/9}$ for all i. Then g is $O(\theta^{1/18})$-close to the threshold function* $\mathrm{sgn}(h)$*, where*

$$h(x) = t + \vec{\gamma}(1)x_1 + \cdots + \vec{\gamma}(n)x_n, \qquad t = \mu^{-1}(\mathbf{E}[g]).$$

(Here μ is the function $\mu(\theta) = \Phi[-\theta, \theta]$.[5])

Roughly, this theorem says that if g is a threshold function with all its degree-1 Chow Parameters small and $\vec{\gamma}$ is approximately parallel to the Chow Vector of g, then g is well approximated by a threshold function whose weights are the coordinates of $\vec{\gamma}$. Theorem 8.1 is essentially proved in [37] as their "Theorem 49". But since it does not appear there exactly as we need it, we prove Theorem 8.1 in Section 8.1.

The remainder of this section is devoted to the proof of Theorem 7.4. This will require extending the "completeness" results in [37]'s main property testing algorithm for Boolean threshold functions, the "Test-LTF" algorithm given in Section 6.5 of [37]. In order to have as self-contained a presentation as possible in this paper, we present a streamlined version of that test in Appendix A. (We can use a streamlined version because in the original testing scenario it was necessary to estimate certain parameters and implicitly identify certain sets; here we can work directly with the desired parameters and sets.) We consider this testing algorithm as being applied to a Boolean threshold function f. We begin by reviewing the steps of the algorithm and recalling a few facts that are established in the proof of the testing algorithm's completeness, Theorem 62 of [37].

Step 1 of **Test-LTF** (see Appendix A) defines the set $H = \{i \in [n] : |\hat{f}(i)| \ge \tau^2\}$. (Note that this definition of H is consistent with our assumption on H from Section 7, see Theorem 7.1 in particular.)

Step 2 defines the set Π of all restrictions π that fix the variables in H.

We introduce some notation before discussing Step 3. Given a restriction π which assigns values to the variables in H, we write f_π to denote the function obtained by applying the restriction π to f. We introduce the notation

$$\vec{\beta}_\pi = [\widehat{f_\pi}(i)]_{i \in T},$$

a vector whose coordinates are indexed by the tail indices T. (We remind the reader that the "tail" T is the set $[n] \setminus H$.)

We next come to the testing algorithm's Step 3. If at least a $1-\epsilon$ fraction of the restrictions π to H satisfy $|\mathbf{E}[f_\pi]| \ge 1-\epsilon$ (this corresponds to Step 3a), then it is easy to see that f is $O(\epsilon)$-close to being a junta over H; in this case condition (i) of Theorem 7.4 holds, and we are done.

Otherwise, it must be the case that less than a $1-\epsilon$ fraction of the restrictions π to H satisfy $|\mathbf{E}[f_\pi]| \ge 1-\epsilon$ (i.e. **Test-LTF** executes Step 3(b)). In this case we observe from Lemma 2.4 that

$$\|\vec{\beta}_\pi\| \ge \Omega(\epsilon) \qquad \text{for at least an } \epsilon \text{ fraction of restrictions } \pi, \tag{8.1}$$

[5] In [37] it was $\Phi[\theta, -\theta]$, as their definition of μ used a different sign convention.

and we proceed to show that condition (ii) of Theorem 7.4 must hold. From now on we shall assume the weights of f are normalized so that $\sum_{i \in T} w_i^2 = 1$.

We now consider the analysis of Step 3b in the testing algorithm's proof of completeness. In particular, we recall Claim 64 of [37] (a component of the completeness proof) which is as follows:

CLAIM 8.2. *(Claim 64 of [37]) Under the conditions of Step 3b, there is a vector $\ell \in \mathbb{R}^T$ with $\|\ell\| = 1$ and all coefficients of magnitude at most $\Omega(\sqrt{\tau})$, such that the following two statements hold: 1. For every restriction $\pi \in \Pi$ fixing the variables in H, the LTF f_π is expressed as $f_\pi(x) = \operatorname{sgn}(\ell \cdot x - (\theta' - w' \cdot \pi))$. 2. For every restriction $\pi \in \Pi$ fixing the variables in H, f_π is $\sqrt{\tau}$-regular.*

We now would like to recall "Theorem 48" from [37], the first part of which intuitively states that every regular threshold function h with mean μ has essentially the same value of $\sum_{i=1}^n \widehat{h}(i)^2$; namely, a certain number $W(\mu)$.

DEFINITION 8.3. *The function $W : [-1,1] \to [0, 2/\pi]$ is defined to be $W(\nu) = (2\varphi(\mu^{-1}(-\nu)))^2$.*[6]

We now restate Theorem 48 from [37]:

THEOREM 8.4. *Let f_1 be a τ-regular threshold function (where τ is assumed less than a sufficiently small constant). Then*

$$\left| \sum_{i=1}^n \widehat{f_1}(i)^2 - W(\mathbf{E}[f_1]) \right| \le \tau^{1/6}. \tag{8.2}$$

Further, suppose $f_2 : \{-1,1\}^n \to \{-1,1\}$ is another τ-regular threshold function that can be expressed using the same linear form as f_1; i.e., $f_1(x) = \operatorname{sgn}(w \cdot x - \theta_1)$ and $f_2(x) = \operatorname{sgn}(w \cdot x - \theta_2)$ for some w, θ_1, θ_2. Then

$$\left| \left(\sum_{i=1}^n \widehat{f_1}(i) \widehat{f_2}(i) \right)^2 - W(\mathbf{E}[f_1]) W(\mathbf{E}[f_2]) \right| \le \tau^{1/6}. \tag{8.3}$$

Since f_π is $\sqrt{\tau}$-regular, we may apply this result and we conclude that

$$\left| \|\vec{\beta}_\pi\|^2 - W(\mathbf{E}[f_\pi]) \right| \le \tau^{1/12}$$

and

$$\left| (\vec{\beta}_\pi \cdot \vec{\beta}_{\pi'})^2 - W(\mathbf{E}[f_\pi]) W(\mathbf{E}[f_{\pi'}]) \right| \le \tau^{1/12}$$

hold for all restrictions π, π'. We also observe here that $\vec{\beta}_\pi \cdot \vec{\beta}_{\pi'} \ge 0$ always holds; this is because f_π and $f_{\pi'}$ are Boolean threshold functions over the same linear form (modulo the constant term) and thus $\operatorname{sgn}(\widehat{f_\pi}(i)) = \operatorname{sgn}(\widehat{f_{\pi'}}(i))$ for all $i \in T$. Using these facts along with our Lemma 7.7 lets us deduce the following:

$$\vec{\beta}_\pi \text{ and } \vec{\beta}_{\pi'} \text{ are } O(\tau^{1/36})\text{-approximately parallel for all restrictions } \pi, \pi'. \tag{8.4}$$

We next recall Lemma 65 from [37]:

LEMMA 8.5. *(Lemma 65 of [37]) Suppose that $|\mathbf{E}[f_\pi] - \mu(\theta' - w' \cdot \pi)| \le \sqrt{\tau}$ holds for every restriction $\pi \in \Pi$ fixing the variables in H. Then there is a vector w^* whose*

[6] The exact formula for $W(\cdot)$ is not important in this paper but we provide it for completeness.

entries are integer multiples of $\sqrt{\tau}/|H|$ at most $2^{O(|H|\log|H|)}\sqrt{\ln(1/\tau)}$ in absolute value, and an integer multiple θ^* of $\sqrt{\tau}/|H|$, also at most $2^{O(|H|\log|H|)}\sqrt{\ln(1/\eta)}$ in absolute value, such that $|\mathbf{E}[f_\pi] - \mu(\theta^* - w^* \cdot \pi)| \le 4\eta^{1/6}$ also holds for all $\pi \in \Pi$.

Rephrasing, this lemma establishes the existence of head weights u_i, each of which is an integer multiple of $\sqrt{\tau}/|H|$ with magnitude at most $2^{O(|H|\log|H|)}\sqrt{\ln(1/\tau)}$, such that

$$\left|\mu(u_0 + \sum_{i\in H} u_i\pi_i) - \mathbf{E}[f_\pi]\right| \le O(\tau^{1/6}) \qquad \text{for all restrictions } \pi. \tag{8.5}$$

(Recall again that μ is the function $\mu(\theta) = \Phi[-\theta,\theta]$.)

We now come to the key part of the present proof: analyzing

$$\vec{\gamma} \stackrel{\text{def}}{=} \text{avg}_\pi[\vec{\beta}_\pi] = [\widehat{f}(i)]_{i\in T}.$$

We first observe that

$$|\vec{\gamma}(i)| \le \sqrt{\tau} \qquad \text{for all } i \in T. \tag{8.6}$$

This is because each f_π is $\sqrt{\tau}$-regular and thus $|\vec{\beta}_\pi(i)| \le \sqrt{\tau}$ for all $i \in T$. Using Proposition 7.6 (in particular, the convexity of the set of vectors that are $O(\tau^{1/36})$-approximately parallel to $\vec{\beta}_\pi$), we deduce from (8.4) that

$$\vec{\beta}_\pi \text{ and } \vec{\gamma} \text{ are } O(\tau^{1/36})\text{-approximately parallel for all restrictions } \pi. \tag{8.7}$$

At this point we fix $\tau = \Omega(\epsilon^{144})$. This gives us that $O(\tau^{1/36}) \le (3/4)\epsilon^2$, so combining (8.1), (8.4), and Lemma 7.8, we conclude that

$$\sigma \stackrel{\text{def}}{=} \|\vec{\gamma}\| \ge \Omega(\epsilon^2) - O(\tau^{1/36})/\epsilon \ge \Omega(\epsilon^2) = \Omega(\tau^{1/72}). \tag{8.8}$$

With conditions (8.6), (8.7), and (8.8) in hand, we are in a position to apply Theorem 8.1 (i.e., essentially Theorem 49 from [37]) which analyzes the soundness of the testing algorithm. Applying Theorem 8.1 to the vector $\vec{\gamma}$ would be sufficient to prove Theorem 7.3; to prove the present theorem, we further observe that condition (7.1)—i.e., $\|\vec{\alpha} - \vec{\gamma}\| \le \Omega(\epsilon^4) = \Omega(\tau^{1/36})$—easily implies that the vector $\vec{\alpha}$ satisfies the following conditions:

$$\sigma' \stackrel{\text{def}}{=} \|\vec{\alpha}\| \ge \Omega(\tau^{1/72}); \tag{8.9}$$

$$|\vec{\alpha}(i)| \le O(\tau^{1/36}) \text{ for all } i \in T;$$

$$\vec{\beta}_\pi \text{ and } \vec{\alpha} \text{ are } O(\tau^{1/36})\text{-approximately parallel for all restrictions } \pi.$$

These allow us to apply Theorem 8.1 where each f_π plays the role of "g" and $\vec{\alpha'} = \vec{\alpha}/\sigma'$ plays the role of "$\vec{\gamma}$" in Theorem 8.1. Note that we have

$$|\vec{\alpha'}(i)| \le O(\tau^{1/72}) \qquad \text{for all } i \in T$$

and

$$\vec{\alpha'} \text{ is } O(\tau^{1/72})\text{-approximately parallel to each } \vec{\beta}_\pi.$$

We may take θ in Theorem 8.1 to be $O(\tau^{1/8})$ and conclude that each f_π is $O(\tau^{1/144})$-close to the threshold function

$$h_\pi \stackrel{\text{def}}{=} \operatorname{sgn}\left(\mu^{-1}(\mathbf{E}[f_\pi]) + \sum_{i\in T} \vec{\alpha}'(i)x_i\right).$$

Let

$$f'(x) \stackrel{\text{def}}{=} \operatorname{sgn}\left(u_0 + \sum_{i\in H} u_i x_i + \sum_{i\in T} \vec{\alpha}'(i)x_i\right).$$

We shall show that for each π we have $\operatorname{dist}(f_\pi, f'_\pi) \le O(\tau^{1/144})$; by our choice of $\tau = \Omega(\epsilon^{144})$ this means that $\operatorname{dist}(f_\pi, f'_\pi) \le O(\epsilon)$ for all π, which implies the theorem.

We have

$$\operatorname{dist}(f_\pi, f'_\pi) \le \operatorname{dist}(f_\pi, h_\pi) + \operatorname{dist}(h_\pi, f'_\pi) \le O(\tau^{1/144}) + \operatorname{dist}(h_\pi, f'_\pi)$$

so it remains only to show that $\operatorname{dist}(h_\pi, f'_\pi) \le O(\tau^{1/144})$. Since h_π and f'_π are threshold functions with the same linear part $\sum_{i\in T} \vec{\alpha}(i)x_i$ but different thresholds, we have that $\operatorname{dist}(f_\pi, h_\pi)$ equals

$$\mathbf{Pr}\left[-\sum_{i\in T} \vec{\alpha}'(i)x_i \text{ lies between } \mu^{-1}(\mathbf{E}[f_\pi]) \text{ and } u_0 + \sum_{i\in H} u_i\pi_i\right].$$

We may apply the Berry-Esseen theorem to deduce that the above probability is at most $O(\tau^{1/72})$ plus

$$\left|\Phi\left[\mu^{-1}(\mathbf{E}[f_\pi]), u_0 + \sum_{i\in H} u_i\pi_i\right]\right|. \tag{8.10}$$

It is easy to check that by definition of the μ function this is at most

$$\frac{1}{2}\left|\mu(\mu^{-1}(\mathbf{E}[f_\pi])) - \mu\left(u_0 + \sum_{i\in H} u_i\pi_i\right)\right| = \frac{1}{2}\left|\mathbf{E}[f_\pi] - \mu\left(u_0 + \sum_{i\in H} u_i\pi_i\right)\right|$$

which is at most $O(\tau^{1/6})$ by (8.5). This concludes the proof of Theorem 7.4. □

8.1. Proof of Theorem 8.1. Here we give the proof of Theorem 8.1, which as mentioned essentially appears already in [37]:

Proof. We follow closely the proof of Theorem 49 from [37]. We first handle the case that $|\mathbf{E}[g]| \ge 1 - \theta^{1/18}$; without loss of generality, assume $\mathbf{E}[g] \ge 1 - \theta^{1/18}$. Now by definition,

$$1 - \theta^{1/18} \le \mathbf{E}[g] = \mu(t) = \Phi[-t, t] = 1 - 2\,\mathbf{Pr}[N(0,1) \le -t]$$

and hence $2\,\mathbf{Pr}[N(0,1) \le -t] \le \theta^{1/18}$. But Pinelis's subgaussian inequality [44] implies that

$$\mathbf{Pr}[\vec{\gamma}'(1)x_1 + \cdots + \vec{\gamma}'(n)x_n \le -t] \le O(\mathbf{Pr}[N(0,1) \le -t])$$

and hence $\mathbf{Pr}[h(x) \le 0] \le O(\theta^{1/18})$. Thus $\mathbf{E}[\operatorname{sgn}(h)] \ge 1 - O(\theta^{1/18})$ and we have that g and $\operatorname{sgn}(h)$ are $O(\theta^{1/18})$-close.

We henceforth assume $|\mathbf{E}[g]| \le 1 - \theta^{1/18}$. Exactly as in the proof of Theorem 49 of [37], this implies

$$\sqrt{W(\mathbf{E}[g])} \ge \Omega(\theta^{1/18}). \tag{8.11}$$

(Here W still denotes the function mentioned in Definition 8.3.) Again, as in Theorem 49 we consider $\mathbf{E}[gh]$ and get (analogous to equation (29) in [37])

$$\mathbf{E}[|h|] - \mathbf{E}[gh] \le \left(\sqrt{W(\mathbf{E}[g])} - \vec{\beta} \cdot \vec{\gamma}\right) + O(\theta^{1/9}).$$

Since $\vec{\beta}$ and $\vec{\gamma}$ are $\theta^{1/9}$-approximately parallel and $\|\vec{\gamma}\| = 1$, we get

$$\mathbf{E}[|h|]-\mathbf{E}[gh] \le \left(\sqrt{W(\mathbf{E}[g])} - \|\vec{\beta}\|\right)+\theta^{1/9}+O(\theta^{1/9}) = \left(\sqrt{W(\mathbf{E}[g])} - \|\vec{\beta}\|\right)+O(\theta^{1/9}). \tag{8.12}$$

Since g is θ-regular, Theorem 48 in [37] (restated as Theorem 8.4 of this paper) implies

$$\left|\|\vec{\beta}\|^2 - W(\mathbf{E}[g])\right| \le \theta^{1/6};$$

together with Fact 2.5 and (8.11) this gives

$$\left|\|\vec{\beta}\| - \sqrt{W(\mathbf{E}[g])}\right| \le \theta^{1/6}/\sqrt{W(\mathbf{E}[g])} = O(\theta^{1/9}).$$

Substituting this into (8.12) yields

$$\mathbf{E}[|h|] - \mathbf{E}[gh] \le C\theta^{1/9} \tag{8.13}$$

for some universal constant C. Now similarly to the proof of Theorem 49 we note that (using Corollary 2.8) we have

$$\mathbf{Pr}[|h(x)| \le C\theta^{1/18}] \le 2C\theta^{1/18} + 2\theta^{1/9} \le (2C+2)\theta^{1/18}.$$

If $\mathbf{Pr}[h(x) \ne g(x)] \ge (2C+3)\theta^{1/18}$ then we have that for at least a $\theta^{1/18}$ fraction of points x both $h(x) \ne g(x)$ and $|h| > C\theta^{1/18}$. This implies that

$$\mathbf{E}[|h|] - \mathbf{E}[gh] > \theta^{1/18} \cdot C\theta^{1/18} = C\theta^{1/9}$$

in contradiction to (8.13). Thus $\mathbf{Pr}[h \ne g] \le (2C+3)\theta^{1/18}$, completing the proof. □

9. Proof of the main theorem.

Having established the two ingredients, we are able to prove are main theorem, restated here for convenience:

MAIN THEOREM. *There is a randomized algorithm $\mathcal{A}$ and a function $\kappa(\epsilon) = 2^{-\tilde{O}(1/\epsilon^2)}$ such that the following holds. Let $f : \{-1,1\}^n \to \{-1,1\}$ be a threshold function and let $0 < \epsilon < 1/2$. Write $\vec{\chi}$ for the Chow Vector of f and assume that $\vec{\alpha}$ is a vector satisfying*

$$\|\vec{\alpha} - \vec{\chi}\| \le \kappa(\epsilon). \tag{9.1}$$

Then given as input $\vec{\alpha}$ and ϵ the algorithm $\mathcal{A}$ performs $2^{\mathrm{poly}(1/\kappa(\epsilon))} \cdot n^2 \cdot \log n \cdot \log(\frac{n}{\delta})$ bit operations and outputs the (weights-based) representation of a threshold function f^* which with probability at least $1-\delta$ satisfies $\mathrm{dist}(f, f^*) \le \epsilon$.

Proof. We first present a high-level description of the entire algorithm. We then give a more detailed explanation of how the algorithm performs its main step, Step 1, and prove correctness of the algorithm. Finally we analyze the running time.

High-level description of $\mathcal{A}$. Algorithm $\mathcal{A}$ is given $\epsilon > 0$ and the vector $\vec{\alpha}$ as input. The algorithm executes the following steps:

Step 0. Truncate each $\vec{\alpha}(i)$ to additive accuracy $\pm\sqrt{\kappa(\epsilon)/(n+1)}$. (Note that this changes the location of $\vec{\alpha}$ by distance at most $\kappa(\epsilon)$, so absorbing the factor of 2 into the definition of $\kappa(\epsilon)$ we have that (9.1) still holds for the new $\vec{\alpha}$.)

Step 1. Generate a list of $2^{\mathrm{poly}(1/\kappa(\epsilon))}$ "candidate" threshold functions f'.

Step 2. Let $\epsilon_0 = 2^{-\tilde{O}(1/\epsilon^2)}$ be such that in an application of Theorem 1.6, having $d_{\mathrm{Chow}}(f, f^*) \leq 6\sqrt{\epsilon_0}$ implies $\mathrm{dist}(f, f^*) \leq \epsilon$. Estimate each of the candidates' Chow Vectors to within distance $\sqrt{\epsilon_0}$ (see Fact 9.2 for how this is done), and output any f^* whose Chow Vector estimate has distance at most $4\sqrt{\epsilon_0}$ from $\vec{\alpha}$.

Having outlined the algorithm, we now give a detailed explanation of Step 1 and prove correctness. Running time analysis is deferred to the end.

Step 1 details and correctness. The way that $\mathcal{A}$ generates the $2^{\mathrm{poly}(1/\kappa(\epsilon))}$ "candidate" threshold functions in Step 1 is based on Theorem 7.4. Let τ_0 denote $\tau(\epsilon_0)$. The set H in Theorem 7.4 is taken to be the set of all indices $1 \leq i \leq n$ for which $|\vec{\alpha}(i)| \geq \tau_0^2/2$. If we now fix $\kappa(\epsilon) = \tau_0^2/2$ (which is indeed $2^{-\tilde{O}(1/\epsilon^2)}$), we are assured that H contains all indices i for which $|\vec{\chi}(i)| = |\widehat{f}(i)| \geq \tau_0^2$, since if H were missing even one such index this would cause $\|\vec{\alpha} - \vec{\chi}\| > \kappa(\epsilon)$ contrary to (9.1). Note also that $|H| \leq O(1/\tau_0^4) = \mathrm{poly}(1/\kappa(\epsilon))$, since $\sum \vec{\alpha}(i)^2 \approx \sum \widehat{f}(i)^2 \leq 1$.

Algorithm $\mathcal{A}$ performs Step 1 by generating two sets of candidate threshold functions, corresponding to the two cases in Theorem 7.4. The first set simply consists of all threshold functions which are juntas over H. Recalling the classic fact [40] that every threshold function over $|H|$ Boolean variables can be represented using integer weights each of magnitude $2^{O(|H|\log|H|)}$, algorithm $\mathcal{A}$ can construct all candidate threshold functions in the first set in time $2^{O(|H|^2\log|H|)} = 2^{\mathrm{poly}(1/\kappa(\epsilon))}$ by simply creating a candidate from each possible vector of integer weights in this range. The second set of candidates consists of all threshold functions whose "head weights" (for indices in H) are integer multiples of $\sqrt{\tau_0}/|H|$ with magnitude at most $2^{O(|H|\log|H|)}\sqrt{\ln(1/\tau_0)}$ and whose "tail weights" (for indices in $T = [n] \setminus H$) are given by $\vec{\alpha}/\|\vec{\alpha}\|$. It is not difficult to see that there are again at most $2^{\mathrm{poly}(1/\kappa(\epsilon))}$ such candidates.

By Theorem 7.4, at least one of the two sets of candidates contains a threshold function f' which has $\mathrm{dist}(f, f') \leq \epsilon_0$. (This uses the fact that as required by statement (ii) of Theorem 7.4, we indeed have $\|\vec{\alpha} - \vec{\chi}\| \leq \kappa(\epsilon) \leq \Omega(\epsilon_0^4)$.) By Proposition 1.5 this f' also satisfies $d_{\mathrm{Chow}}(f, f') \leq 2\sqrt{\epsilon_0}$; writing $\vec{\chi'}$ for the Chow vector of f', the triangle inequality implies

$$\|\vec{\alpha} - \vec{\chi'}\| \leq \|\vec{\alpha} - \vec{\chi}\| + \|\vec{\chi} - \vec{\chi'}\| \leq 3\sqrt{\epsilon_0}$$

(this uses the fact that $\kappa(\epsilon)$ is smaller than $\sqrt{\epsilon_0}$).

To conclude the proof of correctness, we now observe that since Step 2 estimates the Chow Vector of each candidate to within distance $\sqrt{\epsilon_0}$, there must indeed be at least one candidate f^* whose Chow Vector estimate has distance at most $4\sqrt{\epsilon_0}$ from $\vec{\alpha}$. So f^*'s true Chow Vector has distance at most $5\sqrt{\epsilon_0}$ from $\vec{\alpha}$, and the triangle inequality implies $d_{\mathrm{Chow}}(f, f^*) \leq 6\sqrt{\epsilon_0}$ (again using $\kappa(\epsilon) \leq \sqrt{\epsilon_0}$). Now Theorem 1.6 implies $\mathrm{dist}(f, f^*) \leq \epsilon$, as desired. This concludes the proof of correctness.

Running time analysis. We first observe that as a result of the truncation performed in Step 0, each number $\vec{\alpha}(i)$ is represented using only $O(\log n + \log(1/\kappa(\epsilon)))$ bits. (Without this truncation, if the input vector $\vec{\alpha}$ were the exact Chow Vector of a threshold function f then each $\vec{\alpha}(i)$ could require an n-bit representation; working with these numbers would slow down the algorithm by almost a factor of n.)

As described above, algorithm $\mathcal{A}$ generates two sets with a total of $2^{\mathrm{poly}(1/\kappa(\epsilon))}$ candidate threshold functions f'. Each candidate in either set has $\mathrm{poly}(1/\kappa(\epsilon))$ many "head weights," each of which is represented with $\mathrm{poly}(1/\kappa(\epsilon))$ bits. The tail weights in the second set of candidates are each represented using $O(\log n + \log(1/\kappa(\epsilon)))$ bits before normalization (i.e. before dividing by $\|\vec{\alpha}\|$), and also $O(\log n + \log(1/\kappa(\epsilon)))$ bits after normalization (this is a consequence of the lower bound $\|\vec{\alpha}\| \geq \Omega(\epsilon_0^2) \geq \mathrm{poly}(1/\kappa(\epsilon))$ which follows from (8.9) in the proof of Theorem 7.4). This straightforwardly yields the following facts:

FACT 9.1. *Given an input $x \in \{-1,1\}^n$ and a candidate threshold function f' as described above, one can evaluate $f'(x)$ using* $\mathrm{poly}(1/\kappa(\epsilon)) + O(1) \cdot n(\log n + \log(1/\kappa(\epsilon)))$ *bit operations.*

FACT 9.2. *Given a candidate threshold function f' as described above and an accuracy parameter $0 < \eta < 1/2$, one can estimate the Chow vector $\vec{\chi}_{f'}$ to within L_2-distance η with confidence $1-\delta$ using a total of at most* $\mathrm{poly}(1/\kappa(\epsilon)) \cdot n^2 \log(\frac{n}{\delta}) \cdot (\log n + \log(1/\kappa(\epsilon)))/\eta^2$ *bit operations.*

Proof sketch: Generate a sample of $O(n \log(\frac{n}{\delta})/\eta^2)$ uniform random examples and use it to empirically estimate each of the $n+1$ Chow Parameters ($\mathbf{E}[f(x)x_i]$ or $\mathbf{E}[f(x)]$) to accuracy $\pm\eta/\sqrt{n+1}$ with confidence $1 - \frac{\delta}{n+1}$. The bit complexity is linear in the number of examples times the time required to evaluate each example, which is upper bounded by the claimed bound. □

By taking δ/M in place of δ in the above, one can estimate the Chow Vector for each of a set of M candidate threshold functions to within distance η with confidence $1-\delta$ using

$$M \cdot \mathrm{poly}(1/\kappa(\epsilon)) \cdot n^2 \log\left(\frac{Mn}{\delta}\right) \cdot (\log n + \log(1/\kappa(\epsilon)))/\eta^2$$

bit operations. In our context we have $\eta = \sqrt{\epsilon_0} = 2^{-\tilde{O}(1/\epsilon^2)}$ and $M = 2^{\mathrm{poly}(1/\kappa(\epsilon))}$, so the overall running time for constructing all the estimates in Step 2, which dominates the overall running time of Algorithm $\mathcal{A}$, is at most

$$2^{\mathrm{poly}(1/\kappa(\epsilon))} \cdot n^2 \cdot \log n \cdot \log(\tfrac{n}{\delta})$$

bit operations as claimed. This concludes the running time analysis of Algorithm $\mathcal{A}$, and with it the proof of the Main Theorem. □

10. Applications to learning theory.

As we now explain, our main theorem has a range of interesting consequences in learning theory.

10.1. Learning threshold functions in the 1-RFA model.

We briefly recall the 1-RFA model that was introduced by Ben-David and Dichterman [3] to model the phenomenon of a learner having incomplete access to examples. In this model there is a target function f and a distribution $\mathcal{D}$ over n-bit examples. Each time the learner is about to receive a labeled example it specifies an index $1 \leq i \leq n$, then an n-bit string x is drawn from the distribution $\mathcal{D}$ and the learner is given $(x_i, f(x))$, i.e. she is only shown the i-th bit of the example along with the label. It is not difficult to show (see Birkendorf et al. [5]) that it is information-theoretically impossible to learn threshold functions in the 1-RFA model if the distribution $\mathcal{D}$ is allowed to be arbitrary. Thus, we restrict our attention to the uniform distribution setting in which $\mathcal{D}$ is uniform over $\{-1,1\}^n$.

Birkendorf et al. [5] showed that a sample of $O(nW^2 \log(\frac{n}{\delta})/\epsilon^2)$ many examples is information-theoretically sufficient for learning an unknown threshold function with integer weights w_i that satisfy $\sum_i |w_i| \le W$. For constant ϵ, the results of Goldberg [20] and Servedio [47] mentioned in Section 3 respectively yield $n^{O(\log n)}$ and poly(n) sample complexity bounds for learning arbitrary threshold functions. However, no efficient algorithms were proposed to accompany any of these information-theoretic bounds.

Birkendorf et al. [5] asked whether there is an efficient uniform-distribution 1-RFA learning algorithm for threshold functions.[7] For constant ϵ, our Main Theorem gives an affirmative answer: each of the $n+1$ Chow Parameters ($\mathbf{E}[f(x)x_i]$ or $\mathbf{E}[f(x)]$) can be empirically estimated in the 1-RFA model, so it is straightforward to construct an approximation $\vec{\alpha}$ to the Chow Vector $\vec{\chi}_f$ of f as required by our Main Theorem. Since the running time of the algorithm $\mathcal{A}$ dominates the time required to construct $\vec{\alpha}$, we have:

Theorem 1.7 Restated. *There is an algorithm which properly learns threshold functions to accuracy ϵ and confidence $1-\delta$ in the uniform distribution 1-RFA model. The algorithm performs $2^{2^{\tilde{O}(1/\epsilon^2)}} \cdot n^2 \cdot \log n \cdot \log(\frac{n}{\delta})$ bit operations.*

10.2. A fast agnostic-type learning algorithm for halfspaces under the uniform distribution. The agnostic learning model was introduced by Kearns et al. in 1994 [31], but quite recently there has been considerable progress in both positive and negative results on agnostically learning threshold functions. Let $\mathcal{D}$ be a distribution over $\{-1,1\}^n$ and let $g : \{-1,1\}^n \to \{-1,1\}$ be an arbitrary Boolean function. We write opt to denote the optimal error rate of any threshold function for approximating g with respect to $\mathcal{D}$, i.e.

$$\mathsf{opt} \stackrel{\text{def}}{=} \min_f \mathbf{Pr}_{x\sim D}[f(x) \ne g(x)]$$

where the min is taken over all threshold functions f. An algorithm which, for any g and any $\mathcal{D}$, constructs a hypothesis h satisfying

$$\mathbf{Pr}_{x\sim D}[h(x) \ne g(x)] \le \mathsf{opt} + \epsilon \tag{10.1}$$

is said to be an *agnostic* learning algorithm for threshold functions.

Positive results. Kalai et al. [28] gave a uniform distribution agnostic learning algorithm for threshold functions: if $\mathcal{D}$ is the uniform distribution over $\{-1,1\}^n$, their algorithm outputs a hypothesis h which satisfies (10.1) as desired. However, the hypothesis that the algorithm constructs is of the form sgn$(p(x))$ where $p(x)$ is a polynomial of degree $O(1/\epsilon^4)$, so the algorithm is not proper since it does not output a threshold function. Perhaps more significantly, the running time of their algorithm is $n^{O(1/\epsilon^4)}$.

Negative results. Results of Klivans and Sherstov [32] and Feldman et al. [15] show that under plausible cryptographic hardness assumptions, there is no polynomial-time algorithm that can agnostically learn threshold functions under arbitrary distributions. Feldman et al. [15] also showed that complexity-theoretic assumptions rule out even a very weak form of *proper* agnostic learning for threshold functions. More precisely, they showed that for any constant $\epsilon > 0$, if P $\ne$ NP then there is no algorithm which, given a data set of labeled examples (x,y) (where each $x \in \mathbb{Q}^n$) that has $\mathsf{opt} = 1-\epsilon$, outputs a threshold function hypothesis that agrees with $\frac{1}{2}+\epsilon$ fraction of the labeled examples. Guruswami and Raghavendra [23] proved that this result holds even if the data points x belong to the Boolean cube $\{-1,1\}^n$.

[7] More precisely, they explicitly asked whether there is a *proper* learning algorithm, i.e. one which constructs a threshold function as its hypothesis; our algorithm is of course proper.

Our results. As we now show, the tools we have developed quite directly yield a very fast agnostic-type uniform distribution learning algorithm for threshold functions. We call our algorithm "agnostic-type" instead of agnostic because the hypothesis it constructs is guaranteed to have error at most $O(\mathsf{opt}^{\Omega(1)}) + \epsilon$ instead of $\mathsf{opt} + \epsilon$.[8] However, our algorithm has some significant advantages to offset this drawback: chief among these is its running time, which is $\tilde{O}(n^2)$ for any constant ϵ. So for example, if $\mathsf{opt} > 0$ is a sufficiently small constant then our algorithm can construct a hypothesis with error rate 0.01 in time $\tilde{O}(n^2)$, while to construct a similarly accurate hypothesis the [28] algorithm would need running time something like n^{10^8}. We also note that our algorithm constructs a threshold function hypothesis and hence is *proper*; this is in contrast with the [28] algorithm. Indeed, it is interesting to observe that the result of [23] shows that (assuming P $\neq$ NP) no analogue of our algorithm with a similar performance guarantee can exist for learning under arbitrary distributions $\mathcal{D}$ over $\{-1,1\}^n$.

THEOREM 10.1. *There is an algorithm $\mathcal{B}$ with the following performance guarantee: Let g be any Boolean function and let $\mathsf{opt} = \min_f \mathbf{Pr}[f(x) \neq g(x)]$ where the min is over all threshold functions and the probability is uniform over $\{-1,1\}^n$. Given an input parameter $\epsilon > 0$ and access to independent uniform examples $(x, g(x))$, algorithm $\mathcal{B}$ outputs the (weights-based) representation of a threshold function f^* which with probability at least $1-\delta$ satisfies $\mathbf{Pr}[h(x) \neq g(x)] \leq O(\mathsf{opt}^{\Omega(1)}) + \epsilon$. The algorithm performs*

$$\mathrm{poly}(1/\epsilon) \cdot n^2 \cdot \log(\tfrac{n}{\delta}) \;+\; 2^{\mathrm{poly}(1/\epsilon)} \cdot n \cdot \log n \cdot \log(\tfrac{1}{\delta})$$

bit operations.

The algorithm and analysis are similar to Algorithm $\mathcal{A}$ from Section 9, but slightly simpler since we do not need to estimate Chow Parameters and use Theorem 1.6 to gauge the accuracy of each candidate – instead we can just directly estimate the empirical accuracy of each candidate using random examples. This is what enables the algorithm to save an exponential in the dependence on ϵ compared with the running time of Algorithm $\mathcal{A}$, and also a $\log n$ factor since we do not have to take a union bound over all $n+1$ estimated Chow Parameters of each candidate. We now give a detailed proof.

Proof. (Theorem 10.1.) We will initially assume that the algorithm $\mathcal{B}$ is given the value of opt as input. At the end of the proof we briefly discuss how this assumption can be removed.

Let $\gamma \overset{\text{def}}{=} \Theta(\epsilon^{O(1)})$. Algorithm $\mathcal{B}$ works as follows:

Step 0. Using uniform examples $(x, g(x))$, empirically estimate each of the $n+1$ Chow Parameters $\widehat{g}(i) \overset{\text{def}}{=} \mathbf{E}[g(x)x_i]$ and $\widehat{g}(0) \overset{\text{def}}{=} \mathbf{E}[g(x)]$ to additive accuracy $\pm\gamma/\sqrt{n+1}$. Let $\vec{\alpha}$ denote the estimated Chow vector of g obtained this way.

Step 1. Use $\vec{\alpha}$ to generate a list of candidate threshold functions f'.

Step 2. For each candidate threshold function f', empirically estimate the error rate $\mathbf{Pr}[f'(x) \neq g(x)]$ to within additive accuracy $\pm\epsilon/4$, using the source of uniform examples $(x, g(x))$. Output the candidate f^* whose estimated error rate is lowest.

[8] We remark here that [28] in fact show that achieving $\mathsf{opt} + \epsilon$ accuracy in time $n^{1/\epsilon^{2-\kappa}}$ for any constant $\kappa > 0$ would imply a very substantial improvement in the fastest known algorithms for the challenging problem of learning parity with noise: in particular, this would give an algorithm running in time $2^{n^{1-\kappa'}}$, improving on the current $2^{n/\log n}$ running time of [6]. We feel that this motivates research into algorithms which, like the one we present, have higher error rates but faster running times.

Let f be the threshold function such that $\mathbf{Pr}[f(x) \neq g(x)] = \mathsf{opt}$. From Proposition 1.5 we have $d_{\mathrm{Chow}}(f, g) \leq 2\sqrt{\mathsf{opt}}$. Since the estimated Chow Vector $\vec{\alpha}$ has distance at most γ from the true Chow Vector $\vec{\chi}_g$, we have

$$\|\vec{\alpha} - \vec{\chi}_f\| \leq \mu, \qquad \text{where } \mu \stackrel{\text{def}}{=} 2\sqrt{\mathsf{opt}} + \gamma. \tag{10.2}$$

(Here μ essentially plays the role of "$\kappa(\epsilon)$" in the right-hand side of (9.1).

Let H denote the set of all indices $1 \leq i \leq n$ for which $|\vec{\alpha}(i)| \geq \mu$. This set H must contain all indices i for which $|\vec{\chi}_f(i)| \geq 2\mu$, for if H were missing even one such index this would cause $\|\vec{\alpha} - \vec{\chi}_f\| > \mu$ in violation of (10.2). We have $|H| \leq O(1)/\mu^2$ since $\sum \vec{\alpha}(i)^2 \approx \sum \widehat{f}(i)^2 \leq 1$.

Similar to Algorithm $\mathcal{A}$, Algorithm $\mathcal{B}$ performs Step 1 by generating two sets of candidates according to Theorem 7.4 (where now the role of "$\tau(\epsilon)^2$" of that theorem is played by 2μ). The first set of candidates are all threshold function juntas over H[9] and the second set are all threshold functions whose tail weights (for variables in $[n] \setminus H$) are given by $\vec{\alpha}/\|\vec{\alpha}\|$ and whose head weights (for variables in H) are integer multiples of $(2\mu)^{1/4}/|H|$ with magnitude at most $2^{O(|H| \log |H|)}\sqrt{\ln(1/\mu)}$. Our bound on $|H|$ implies that there are at most $2^{\mathrm{poly}(1/\mu)}$ candidates in total.

Now let $\tau(\cdot)$ denote the polynomial function from Theorem 7.4. There is a value $\kappa = O(\mu^{\Omega(1)})$ such that both of the following hold: (i) $\|\vec{\alpha} - \vec{\chi}_f\| \leq \; C\kappa^4$, where C is the constant implicit in the RHS of (7.1); and (ii) $\tau(\kappa) \leq \mu$. (Note that if $\tau(\kappa)$ is greater than μ, we may take a larger polynomial for τ in Theorem 7.4 without affecting the correctness of that theorem.) We may thus apply Theorem 7.4 (with κ playing the role of "ϵ" in the theorem's statement) and conclude that at least one candidate is $O(\kappa)$-close to f. Since f is opt-close to g, this means that some candidate is $(O(\kappa) + \mathsf{opt})$-close to g. Since each candidate's error rate with respect to g is empirically estimated to within additive accuracy $\pm\epsilon/4$, this means that a candidate will be found with estimated error w.r.t. g at most $O(\kappa) + \mathsf{opt} + \epsilon/4$, and that the true error rate of this candidate w.r.t. g is at most $O(\kappa) + \mathsf{opt} + \epsilon/2$. Tracing back through the definitions of κ, μ and γ, we have that the error rate of this candidate w.r.t. g is at most

$$O(\kappa)+\mathsf{opt}+\epsilon/2 \leq O(\mu^{\Omega(1)})+\mathsf{opt}+\epsilon/2 \leq O((2\sqrt{\mathsf{opt}}+\gamma)^{\Omega(1)})+\mathsf{opt}+\epsilon/2 \leq O(\mathsf{opt}^{\Omega(1)})+\epsilon$$

as claimed.

Now we analyze the running time of the algorithm. In Step 0, the algorithm needs $O(n \log(n/\delta)/\gamma^2)$ examples to obtain all $n+1$ estimates with total failure probability $\delta/2$; this takes $O(n^2 \log(n/\delta)/\gamma^2)$ bit operations. For Steps 1 and 2, as described above there are at most $2^{\mathrm{poly}(1/\mu)}$ candidates that are generated. In each candidate representation, the first $\mathrm{poly}(1/\mu)$ many weights have $\mathrm{poly}(1/\mu)$-bit representations and the remaining weights have $O(\log n + \log(1/\epsilon))$-bit representations (even after normalization). Thus the time required to evaluate each candidate representation on an example x is $\mathrm{poly}(1/\mu)+O(n(\log n+\log(1/\epsilon)))$ many bit operations. For each candidate, the empirical error rate estimate can be obtained using $O(\log(2^{\mathrm{poly}(1/\mu)}/\delta)/\epsilon^2)$ many examples, with total failure probability $\delta/2$ over all candidates. Thus the overall running time required for Steps 1 and 2 is at most

$$2^{\mathrm{poly}(1/\mu)} \cdot n \cdot (\log n + \log(1/\epsilon)) \cdot \log(1/\delta)/\epsilon^2.$$

[9] Note that to construct the set H the algorithm must know the value of opt.

Recalling that $\mu \geq \Omega(\epsilon^{\Omega(1)})$, we can bound the runtime of Steps 1 and 2 by

$$2^{\mathrm{poly}(1/\epsilon)} \cdot n \cdot \log n \cdot \log(1/\delta).$$

Thus the overall running time (including Step 0) is at most

$$2^{\mathrm{poly}(1/\epsilon)} \cdot n \cdot \log n \cdot \log(1/\delta) \ + \ \mathrm{poly}(1/\epsilon) \cdot n^2 \cdot (\log n + \log(1/\delta)) \tag{10.3}$$

as claimed.

We now briefly explain how the assumption that $\mathcal{B}$ is given the value of opt can be removed. We first observe that in the proof given above, if the value of opt used by the algorithm is off by an additive $\pm O(\epsilon^{\Omega(1)})$ the algorithm will still generate a hypothesis whose final error rate is at most $O(\mathsf{opt}^{\Omega(1)}) + 2\epsilon$. So if the algorithm is not given the value of opt, it can successively try $\mathsf{opt}' = 1/2, 1/2 - \Delta, 1/2 - 2\Delta, \dots$ where $\Delta = O(\epsilon^{\Omega(1)})$, perform hypothesis testing on the result of each attempt, and ultimately output the best-performing one. Since there are at most $\mathrm{poly}(1/\epsilon)$ runs of the algorithm, (10.3) is still an upper bound on the running time. This concludes the proof of Theorem 1.8. □

10.3. A fast uniform-distribution PAC learning algorithm for halfspaces. The usual (noise-free) uniform distribution PAC learning model corresponds to the special case of the agnostic model in which the target function g is required to actually be a threshold function, i.e. $\mathsf{opt} = 0$. Theorem 1.8 thus immediately gives us an algorithm that can PAC learn threshold functions in the usual (noise-free) uniform distribution model in the stated time bound.

We observe that for constant ϵ the running time of this algorithm is close to optimal even in this noise-free scenario. Known information-theoretic lower bounds [4, 33] imply that *any* algorithm that learns threshold functions to fixed constant accuracy (say $\epsilon = 0.01$) under the uniform distribution must use $\Omega(n)$ labeled examples; this is true even if the algorithm is allowed to make membership queries. Thus the information-theoretic minimum input length that is required for this problem is $\Omega(n^2)$ bits—this is very close to the $O(n^2 \log n)$ bit operations our algorithm performs. As far as we are aware, the previous fastest known approach to learning threshold functions to constant accuracy under the uniform distribution on $\{-1,1\}^n$ would require using linear programming and require [55] at least $\tilde{O}(n^{4.5})$ bit operations (more precisely, $\tilde{O}(n^{3.5})$ arithmetic operations on $\tilde{O}(n)$-bit operands).

REFERENCES

[1] H. Aziz, M. Paterson, and D. Leech, *Efficient algorithm for designing weighted voting games*, in IEEE Intl. Multitopic Conf., 2007, pp. 1–6.

[2] J. Banzhaf, *Weighted voting doesn't work: A mathematical analysis*, Rutgers Law Review, 19 (1965), pp. 317–343.

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

[4] G. Benedek and A. Itai, *Learnability with respect to fixed distributions*, Theor. Comput. Sci., 86 (1991), pp. 377–390.

[5] A. Birkendorf, E. Dichterman, J. Jackson, N. Klasner, and H.U. Simon, *On restricted-focus-of-attention learnability of Boolean functions*, Machine Learning, 30 (1998), pp. 89–123.

[6] A. Blum, A. Kalai, and H. Wasserman, *Noise-tolerant learning, the parity problem, and the statistical query model*, J. ACM, 50 (2003), pp. 506–519.

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

[8] F. Carreras, *On the design of voting games*, Mathematical Methods of Operations Research, 59 (2004), pp. 503–515.

[9] C.K. Chow, *On the characterization of threshold functions*, in Proceedings of the Symposium on Switching Circuit Theory and Logical Design (FOCS), 1961, pp. 34–38.

[10] M. Dertouzos, *Threshold Logic: A Synthesis Approach*, MIT Press, Cambridge, MA, 1965.

[11] P. Dubey and L.S. Shapley, *Mathematical properties of the banzhaf power index*, Mathematics of Operations Research, 4 (1979), pp. 99–131.

[12] D. Dubhashi and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms*, Cambridge University Press, Cambridge, 2009.

[13] E. Einy and E. Lehrer, *Regular simple games*, International Journal of Game Theory, 18 (1989), pp. 195–207.

[14] C.C. Elgot, *Truth functions realizable by single threshold organs*, in Proceedings of the Symposium on Switching Circuit Theory and Logical Design (FOCS), 1960, pp. 225–245.

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

[16] D. Felsenthal and M. Machover, *A priori voting power: what is it all about?*, Political Studies Review, 2 (2004), pp. 1–23.

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

[18] J. Freixas, *Different ways to represent weighted majority games*, Top (Journal of the Spanish Society of Statistics and Operations Research), 5 (1997), pp. 201–212.

[19] M. Garman and M. Kamien, *The paradox of voting: probability calculations*, Behavioral Sci., 13 (1968), pp. 306–16.

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

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

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

[23] V. Guruswami and P. Raghavendra, *Hardness of learning halfspaces with noise*, in Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS), IEEE Computer Society, 2006, pp. 543–552.

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

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

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

[27] J.R. Isbell, *A Counterexample in Weighted Majority Games*, Proceedings of the AMS, 20 (1969), pp. 590–592.

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

[29] K.R. Kaplan and R.O. Winder, *Chebyshev approximation and threshold functions*, IEEE Trans. Electronic Computers, EC-14 (1965), pp. 315–325.

[30] P. Kaszerman, *A geometric test-synthesis procedure for a threshold device*, Information and Control, 6 (1963), pp. 381–398.

[31] M. Kearns, R. Schapire, and L. Sellie, *Toward Efficient Agnostic Learning*, Machine Learning, 17 (1994), pp. 115–141.

[32] A. Klivans and A. Sherstov, *Cryptographic hardness for learning intersections of halfspaces*, in Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS), 2006, pp. 553–562.

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

[34] E. Lapidot, *The counting vector of a simple game*, Proceedings of the AMS, 31 (1972), pp. 228–231.

[35] D. Leech, *Power indices as an aid to institutional design: the generalised apportionment problem*, in Yearbook on New Political Economy, M. Holler, H.Kliemt, D. Schmidtchen, and M. Streit, eds., 2003.

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

[37] K. Matulef, R. O'Donnell, R. Rubinfeld, and R. Servedio, *Testing halfspaces*, SIAM J.

Comp., 39 (2010), pp. 2004–2047.

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

[39] S. Muroga, I. Toda, and M. Kondo, *Majority decision functions of up to six variables*, Math. Comput., 16 (1962), pp. 459–472.

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

[41] S. Muroga, T. Tsuboi, and C.R. Baugh, *Enumeration of threshold functions of eight variables*, Tech. Report 245, Univ. of Illinois, Urbana, 1967.

[42] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, 1944.

[43] L.S. Penrose, *The elementary statistics of majority voting*, Journal of the Royal Statistical Society, 109 (1946), pp. 53–57.

[44] I. Pinelis, *Extremal probabilistic problems and Hotelling's t^2 test under a symmetry condition*, Ann. Statist., 22 (1994), pp. 357–368.

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

[46] V.P. Roychowdhury, K.-Y. Siu, A. Orlitsky, and T. Kailath, *Vector analysis of threshold functions*, Information and Computation, 120 (1995), pp. 22–31.

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

[48] Q. Sheng, *Threshold Logic*, London, New York, Academic Press, 1969.

[49] I.S. Shiganov, *Refinement of the upper bound of the constant in the central limit theorem*, Journal of Soviet Mathematics, (1986), pp. 2545–2550.

[50] S.J. Szarek, *On the best constants in the Khinchine inequality*, Studia Math., 58 (1976), pp. 197–208.

[51] K. Takamiya and A. Tanaka, *Computational complexity in the design of voting games*, Tech. Report 653, The Institute of Social and Economic Research, Osaka University, 2006.

[52] M. Talagrand, *How much are increasing sets positively correlated?*, Combinatorica, 16 (1996), pp. 243–258.

[53] M. Tannenbaum, *The establishment of a unique representation for a linearly separable function*, tech. report, Lockheed Missiles and Space Co., 1961. Threshold Switching Techniques Note 20, pp. 1-5.

[54] A. Taylor and W. Zwicker, *A Characterization of Weighted Voting*, Proceedings of the AMS, 115 (1992), pp. 1089–1094.

[55] P. M. Vaidya, *A new algorithm for minimizing convex functions over convex sets*, Math. Prog., 73 (1996), pp. 291–341.

[56] R.O. Winder, *Threshold logic in artificial intelligence*, Artificial Intelligence, IEEE Publication S-142 (1963), pp. 107–128.

[57] ———, *Threshold functions through $n = 7$*, Tech. Report 7, Air Force Cambridge Research Laboratories, 1964.

[58] ———, *Threshold gate approximations based on chow parameters*, IEEE Transactions on Computers, (1969), pp. 372–375.

[59] ———, *Chow parameters in threshold logic*, Journal of the ACM, 18 (1971), pp. 265–289.

Appendix A. The algorithm of [37] for testing Boolean threshold functions.

```
**Streamlined-Test-LTF** (inputs are $\epsilon > 0$ and black-box access to $f : \{-1,1\}^n \to \{-1,1\}$)
    0. Let $\tau = \epsilon^{108}$.
    1. Let $H = \{i \in [n] : |\hat{f}(i)| \geq \tau^2\}$.
    2. Let $\Pi$ denote the set of all $2^{|H|}$ restrictions $\pi$ that assign $\pm 1$ values to the variables in $H$.
    3. At this point there are two cases depending on whether or not the set $\Pi' := \{\pi \in \Pi : \ |\mathbf{E}[f_\pi]| \geq 1 - \epsilon\}$ is at least a $1 - \epsilon$ fraction of $\Pi$:
        (a) (The case that $|\Pi'|/|\Pi| \geq 1 - \epsilon$.)
            In this case, enumerate all possible length-$|H|$ integer vectors $w$ with entries up to $2^{O(|H| \log |H|)}$ in absolute value, and also all possible integer thresholds $\theta$ in the same range. For each pair $(w, \theta)$, check whether $\mathrm{sgn}(w \cdot \pi - \theta) = \mathrm{sgn}(\mathbf{E}[f_\pi])$ holds for at least a $1 - 20\epsilon$ fraction of all $\pi \in \Pi$. If this is the case for any $(w, \theta)$ pair then ACCEPT. If it fails for all $(w, \theta)$ then REJECT.
        (b) (The case that $|\Pi'|/|\Pi| < 1 - \epsilon$, i.e. at least an $\epsilon$ fraction of restrictions $\pi$ have $|\mathbf{E}[f_\pi]| < 1 - \epsilon$.)
            In this case, pick any $\pi^*$ such that $|\mathbf{E}[f_{\pi^*}]| < 1 - \epsilon$. Then:
            i. Check that $\sum_{|S|=1} \widehat{f_{\pi^*}}(S)^4 \leq 2\tau$. If this fails, REJECT.
            ii. Check that $|\sum_{|S|=1} \widehat{f_{\pi^*}}(S)^2 - W(\mathbf{E}[f_{\pi^*}])| \leq 2\tau^{1/12}$. If this fails, REJECT.
            iii. Check that both
                $$\left| \left( \sum_{|S|=1} \widehat{f_{\pi^*}}(S) \widehat{f_\pi}(S) \right)^2 - W(\mathbf{E}[f_{\pi^*}]) W(\mathbf{E}[f_\pi]) \right| \leq 2\tau^{1/12}$$
                and $\sum_{|S|=1} \widehat{f_{\pi^*}}(S) \widehat{f_\pi}(S) \geq -\eta$ hold for all $\pi \in \Pi$. If this fails, REJECT.
            iv. Enumerate all possible length-$|H|$ vectors $w$ whose entries are integer multiples of $\sqrt{\tau}/|H|$, up to $2^{O(|H| \log |H|)} \sqrt{\ln(1/\tau)}$ in absolute value, and also all possible thresholds $\theta$ with the same properties. For each pair $(w, \theta)$, check that $|\mathbf{E}[f_\pi] - \mu(w \cdot \pi^i - \theta)| \leq 5\sqrt{\tau}$ holds for all $\pi \in \Pi$. If this ever happens, ACCEPT. If it fails for all $(w, \theta)$, REJECT.
```