Nearly optimal solutions for the Chow Parameters Problem and low-weight approximation of halfspaces

[Extended Abstract]

Anindya De[*]
Computer Science Division
University of California
Berkeley, CA 94720
anindya@cs.berkeley.edu

Ilias Diakonikolas[†]
Computer Science Division
University of California
Berkeley, CA 94720
ilias@cs.berkeley.edu

Vitaly Feldman
IBM Almaden
Research Center
San Jose, CA 95120
vitaly@post.harvard.edu

Rocco A. Servedio[‡]
Dept. of Computer Science
Columbia University
New York, NY 10027
rocco@cs.columbia.edu

ABSTRACT

The *Chow parameters* of a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ are its $n+1$ degree-0 and degree-1 Fourier coefficients. It has been known since 1961 [Cho61, Tan61] that the (exact values of the) Chow parameters of any linear threshold function f uniquely specify f within the space of all Boolean functions, but until recently [OS11] nothing was known about efficient algorithms for *reconstructing* f (exactly or approximately) from exact or approximate values of its Chow parameters. We refer to this reconstruction problem as the *Chow Parameters Problem.*

Our main result is a new algorithm for the Chow Parameters Problem which, given (sufficiently accurate approximations to) the Chow parameters of any linear threshold function f, runs in time $\tilde{O}(n^2) \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$ and with high probability outputs a representation of an LTF f' that is ϵ-close to f. The only previous algorithm [OS11] had running time $\mathrm{poly}(n) \cdot 2^{2^{\tilde{O}(1/\epsilon^2)}}$.

As a byproduct of our approach, we show that for any linear threshold function f over $\{-1,1\}^n$, there is a linear threshold function f' which is ϵ-close to f and has all weights that are integers at most $\sqrt{n} \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$. This significantly improves the best previous result of [DS09] which gave a $\mathrm{poly}(n) \cdot 2^{\tilde{O}(1/\epsilon^{2/3})}$ weight bound, and is close to the known lower bound of $\max\{\sqrt{n}, (1/\epsilon)^{\Omega(\log\log(1/\epsilon))}\}$ [Gol06, Ser07]. Our techniques also yield improved algorithms for related problems in learning theory.

In addition to being significantly stronger than previous work, our results are obtained using conceptually simpler proofs. The two main ingredients underlying our results are (1) a new structural result showing that for f any linear threshold function and g any bounded function, if the Chow parameters of f are close to the Chow parameters of g then f is close to g; (2) a new boosting-like algorithm that given approximations to the Chow parameters of a linear threshold function outputs a bounded function whose Chow parameters are close to those of f.

Categories and Subject Descriptors

F.2.2 [**Nonnumerical Algorithms and Problems**]: Computations on Discrete Structures; I.2.6 [**Learning**]: Concept Learning

General Terms

Theory

Keywords

Boolean function; Fourier Analysis; Threshold function; Chow parameters

1. INTRODUCTION

1.1 Background and motivation.

A *linear threshold function*, or LTF, over $\{-1,1\}^n$ is a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ of the form

$$f(x) = \mathrm{sign}\left(\sum_{i=1}^{n} w_i x_i - \theta\right),$$

where $w_1, \ldots, w_n, \theta \in \mathbb{R}$. The function $\mathrm{sign}(z)$ takes value 1 if $z \geq 0$ and takes value -1 if $z < 0$; the w_i's are the *weights* of f and θ is the *threshold*. Linear threshold functions have been intensively studied for decades in many different fields. They are variously known as "halfspaces" or "linear separators" in machine

[*]Research supported by NSF award CCF-0915929, CCF-1017403 and CCF-1118083.
[†]Research supported by a Simons Postdoctoral Fellowship.
[‡]Research supported by NSF grants CNS-0716245, CCF-0915929, and CCF-1115703.

STOC'12, May 19–22, 2012, New York, New York, USA.

learning and computational learning theory, "Boolean threshold functions," "(weighted) threshold gates" and "(Boolean) perceptrons (of order 1)" in computational complexity, and as "weighted majority games" in voting theory and the theory of social choice. Throughout this paper we shall refer to them simply as LTFs.

The *Chow parameters* of a function $f : \{-1,1\}^n \to \mathbb{R}$ are the $n+1$ values

$$\widehat{f}(0) = \mathbf{E}[f(x)], \quad \widehat{f}(i) = \mathbf{E}[f(x)x_i] \text{ for } i = 1, \ldots, n,$$

i.e. the $n+1$ degree-0 and degree-1 Fourier coefficients of f. (Here and throughout the paper, all probabilities and expectations are with respect to the uniform distribution over $\{-1,1\}^n$ unless otherwise indicated.) It is easy to see that in general the Chow parameters of a Boolean function may provide very little information about f; for example, any parity function on at least two variables has all its Chow parameters equal to 0. However, in a surprising result, C.-K. Chow [Cho61] showed that the Chow parameters of an LTF f *uniquely* specify f within the space of all Boolean functions mapping $\{-1,1\}^n \to \{-1,1\}$. Chow's proof (given in Appendix A) is simple and elegant, but is completely non-constructive; it does not give any clues as to how one might use the Chow parameters to find f (or an LTF that is close to f). This naturally gives rise to the following algorithmic question, which we refer to as the "Chow Parameters Problem:"

> **The Chow Parameters Problem** (rough statement): Given (exact or approximate) values for the Chow parameters of an unknown LTF f, output an (exact or approximate) representation of f as $\mathrm{sign}(v_1x_1 + \cdots + v_nx_n - \theta')$.

Motivation and Prior Work. We briefly survey some previous research on the Chow Parameters problem (see Section 1.1 of [OS11] for a more detailed and extensive account). Motivated by applications in electrical engineering, the Chow Parameters Problem was intensively studied in the 1960s and early 1970s; several researchers suggested heuristics of various sorts [Kas63, Win63, KW65, Der65] which were experimentally analyzed in [Win69]. See [Win71] for a survey covering much of this early work and [Bau73, Hur73] for some later work from this period.

Researchers in game theory and voting theory rediscovered Chow's theorem in the 1970s [Lap72], and the theorem and related results have been the subject of study in those communities down to the present [DS79, EL89, TZ92, Fre97, Lee03, Car04, FM04, TT06, APL07]. Since the Fourier coefficient $\widehat{f}(i)$ can be viewed as representing the "influence" of the i-th voter under voting scheme f (under the "Impartial Culture Assumption" in the theory of social choice, corresponding to the uniform distribution over inputs $x \in \{-1,1\}^n$), the Chow Parameters Problem corresponds to designing a set of weights for n voters so that each individual voter has a certain desired level of influence over the final outcome.

In the 1990s and 2000s several researchers in learning theory considered the Chow Parameters Problem. Birkendorf et al. [BDJ$^+$98] showed that the Chow Parameters Problem is equivalent to the problem of efficiently learning LTFs under the uniform distribution in the "1-Restricted Focus of Attention (1-RFA)" model of Ben-David and Dichterman [BDD98] (we give more details on this learning model in Appendix E). Birkendorf et al. showed that if f is an LTF with integer weights of magnitude at most $\mathrm{poly}(n)$, then estimates of the Chow parameters that are accurate to within an additive $\pm\epsilon/\mathrm{poly}(n)$ information-theoretically suffice to specify the halfspace f to within ϵ-accuracy. Other information-theoretic results of this flavor were given by [Gol06, Ser07]. In complexity theory several generalizations of Chow's Theorem were given in [Bru90, RSOK95], and the Chow parameters play an important role in a recent study [CHIS10] of the approximation-resistance of linear threshold predicates in the area of hardness of approximation.

Despite this considerable interest in the Chow Parameters Problem from a range of different communities, the first provably effective and efficient algorithm for the Chow Parameters Problem was only obtained fairly recently. [OS11] gave a $\mathrm{poly}(n) \cdot 2^{2^{\tilde{O}(1/\epsilon^2)}}$-time algorithm which, given sufficiently accurate estimates of the Chow parameters of an unknown n-variable LTF f, outputs an LTF f' that has $\mathbf{Pr}[f(x) \neq f'(x)] \leq \epsilon$.

1.2 Our results.

In this paper we give a significantly improved algorithm for the Chow Parameters Problem, whose running time dependence on ϵ is almost doubly exponentially better than the [OS11] algorithm. Our main result is the following:

THEOREM 1 (MAIN, INFORMAL STATEMENT). *There is an $\tilde{O}(n^2) \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))} \cdot \log(1/\delta)$-time algorithm $\mathcal{A}$ with the following property: Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF and let $0 < \epsilon, \delta < 1/2$. If $\mathcal{A}$ is given as input ϵ, δ and (sufficiently precise estimates of) the Chow parameters of f, then $\mathcal{A}$ outputs integers $v_1, \ldots, v_n, \theta$ such that with probability at least $1-\delta$, the linear threshold function $f^* = \mathrm{sign}(v_1x_1 + \cdots + v_nx_n - \theta)$ satisfies $\mathbf{Pr}_x[f(x) \neq f^*(x)] \leq \epsilon$.*

Thus we obtain an efficient randomized polynomial approximation scheme (ERPAS) with a *quasi-polynomial* dependence on $1/\epsilon$. We note that for the subclass of LTFs with integer weights of magnitude at most $\mathrm{poly}(n)$, our algorithm runs in $\mathrm{poly}(n/\epsilon)$ time, i.e. it is a *fully* polynomial randomized approximation scheme (FPRAS) (see Section 4.1 for a formal statement). Even for this restricted subclass of LTFs, the algorithm of [OS11] runs in time doubly exponential in $1/\epsilon$.

Our main result has a range of interesting implications in learning theory. First, it directly gives an efficient algorithm for learning LTFs in the uniform distribution 1-RFA model. Second, it yields a very fast agnostic-type algorithm for learning LTFs in the standard uniform distribution PAC model. Both these algorithms run in time quasi-polynomial in $1/\epsilon$. We elaborate on these learning applications in Appendix E.

An interesting feature of our algorithm is that it outputs an LTF with integer weights of magnitude at most $\sqrt{n} \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$. Hence, as a corollary of our approach, we obtain essentially optimal bounds on approximating arbitrary LTFs using LTFs with small integer weights. It has been known since the 1960s that every n-variable LTF f has an exact representation $\mathrm{sign}(w \cdot x - \theta)$ in which all the weights w_i are integers satisfying $|w_i| \leq 2^{O(n \log n)}$, and Håstad [Hås94] has shown that there is an n-variable LTF f for which *any* integer-weight representation must have each $|w_i| \geq 2^{\Omega(n \log n)}$. However, by settling for an approximate representation (i.e. a representation $f' = \mathrm{sign}(w \cdot x - \theta)$ such that $\mathbf{Pr}_x[f(x) \neq f'(x)] \leq \epsilon$), it is possible to get away with much smaller integer weights. Servedio [Ser07] showed that every LTF f can be ϵ-approximated using integer weights each at most $\sqrt{n} \cdot 2^{\tilde{O}(1/\epsilon^2)}$, and this bound was subsequently improved (as a function of ϵ) to $n^{3/2} \cdot 2^{\tilde{O}(1/\epsilon^{2/3})}$ in [DS09]. (We note that ideas and tools that were developed in work on low-weight approximators for LTFs have proved useful in a range of other contexts, including hardness of approximation [FGRW09], property testing [MORS10], and explicit constructions of pseudorandom objects [DGJ$^+$10].)

Formally, our approach to proving Theorem 1 yields the following nearly-optimal weight bound on ϵ-approximators for LTFs:

THEOREM 2 (LOW-WEIGHT APPROXIMATORS FOR LTFS). *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any LTF. There is an LTF $f^* = \mathrm{sign}(v_1x_1 + \cdots + v_nx_n - \theta)$ such that $\mathbf{Pr}_x[f(x) \neq f^*(x)] \leq \epsilon$ and the weights v_i are integers that satisfy*

$$\sum_{i=1}^{n} v_i^2 = n \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}.$$

The bound on the magnitude of the weights in the above theorem is optimal as a function of n and nearly optimal as a function of ϵ. Indeed, as shown in [Hås94, Gol06], in general any ϵ-approximating LTF f^* for an arbitrary n-variable LTF f may need to have integer weights at least $\max\{\Omega(\sqrt{n}), (1/\epsilon)^{\Omega(\log\log(1/\epsilon))}\}$. Thus, Theorem 2 nearly closes what was previously an almost exponential gap between the known upper and lower bounds for this problem. Moreover, the proof of Theorem 2 is constructive (as opposed e.g. to the one in [DS09]), i.e. there is a randomized $\mathrm{poly}(n) \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$-time algorithm that constructs an ϵ-approximating LTF.

Techniques. We stress that not only are the quantitative results of Theorems 1 and 2 dramatically stronger than previous work, but the proofs are significantly more self-contained and elementary as well. The [OS11] algorithm relied heavily on several rather sophisticated results on spectral properties of linear threshold functions; moreover, its proof of correctness required a careful re-tracing of the (rather involved) analysis of a fairly complex property testing algorithm for linear threshold functions given in [MORS10]. In contrast, our proof of Theorem 1 entirely bypasses these spectral results and does not rely on [MORS10] in any way. Turning to low-weight approximators, the improvement from $2^{\tilde{O}(1/\epsilon^2)}$ in [Ser07] to $2^{\tilde{O}(1/\epsilon^{2/3})}$ in [DS09] required a combination of rather delicate linear programming arguments and powerful results on the anti-concentration of sums of independent random variables due to Halász [Hal77]. In contrast, our proof of Theorem 2 bypasses anti-concentration entirely and does not require any sophisticated linear programming arguments.

Two main ingredients underlie the proof of Theorem 1. The first is a new structural result relating the "Chow distance" and the ordinary (Hamming) distance between two functions f and g, where f is an LTF and g is an arbitrary bounded function. The second is a new and simple algorithm which, given (approximations to) the Chow parameters of an arbitrary Boolean function f, efficiently construct a "linear bounded function" (LBF) g – a certain type of bounded function – whose "Chow distance" from f is small. We describe each of these contributions in more detail below.

1.3 The main structural result.

In this subsection we first give the necessary definitions regarding Chow parameters and Chow distance, and then state Theorem 7, our main structural result.

1.3.1 Chow parameters and distance measures.

We formally define the Chow parameters of a function on $\{-1,1\}^n$:

DEFINITION 3. *Given any function $f : \{-1,1\}^n \to \mathbb{R}$, its* Chow Parameters *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 say that the* Chow vector *of f is $\vec{\chi}_f = (\widehat{f}(0), \widehat{f}(1), \ldots, \widehat{f}(n))$.*

The Chow parameters naturally induce a distance measure between functions f, g:

DEFINITION 4. *Let $f, g : \{-1,1\}^n \to \mathbb{R}$. We define the* Chow distance *between f and g to be $d_{\mathrm{Chow}}(f,g) \stackrel{def}{=} \|\vec{\chi}_f - \vec{\chi}_g\|_2$, i.e. the Euclidean distance between the Chow vectors.*

This is in contrast with the familiar L_1-distance between functions:

DEFINITION 5. *The* distance *between $f, g : \{-1,1\}^n \to \mathbb{R}$ is defined as $\mathrm{dist}(f,g) \stackrel{def}{=} \mathbf{E}[|f(x) - g(x)|]$. If $\mathrm{dist}(f,g) \leq \epsilon$, we say that f and g are ϵ-*close.

We note that if f, g are Boolean functions with range $\{-1,1\}$ then $\mathrm{dist}(f,g) = 2\,\mathbf{Pr}[f(x) \neq g(x)]$ and thus dist is equivalent (up to a factor of 2) to the familiar Hamming distance.

1.3.2 The main structural result: small Chow-distance implies small distance.

The following fact can be proved easily using basic Fourier analysis (see Proposition 1.5 in [OS11]):

FACT 6. *Let $f, g : \{-1,1\}^n \to \mathbb{R}$. We have that $d_{\mathrm{Chow}}(f,g) \leq 2\sqrt{\mathrm{dist}(f,g)}$.*

Our main structural result, Theorem 7, is essentially a converse which bounds $\mathrm{dist}(f,g)$ in terms of d_{Chow} when f is an LTF and g is any bounded function:

THEOREM 7 (MAIN STRUCTURAL RESULT). *Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF and $g : \{-1,1\}^n \to [-1,1]$ be any bounded function. If $d_{\mathrm{Chow}}(f,g) \leq \epsilon$ then*

$$\mathrm{dist}(f,g) \leq 2^{-\Omega\left(\sqrt[3]{\log(1/\epsilon)}\right)}.$$

Since Chow's theorem says that if f is an LTF and g is any bounded function then $d_{\mathrm{Chow}}(f,g) = 0$ implies that $\mathrm{dist}(f,g) = 0$, Theorem 7 may be viewed as a "robust" version of Chow's Theorem. Note that the assumption that g is bounded is necessary for the above statement, since the function $g(x) = \sum_{i=0}^{n} \hat{f}(i)x_i$ (where $x_0 \equiv 1$) has $d_{\mathrm{Chow}}(f,g) = 0$, but may have $\mathrm{dist}(f,g) = \Omega(1)$. Results of this sort but with weaker quantitative bounds were given earlier in [BDJ+98, Gol06, Ser07, OS11]; we discuss the relationship between Theorem 7 and some of this prior work below.

Discussion. Theorem 7 should be contrasted with Theorem 1.6 of [OS11], the main structural result of that paper. That theorem says that for $f : \{-1,1\}^n \to \{-1,1\}$ any LTF and $g : \{-1,1\}^n \to [-1,1]$ any bounded function[1], if $d_{\mathrm{Chow}}(f,g) \leq \epsilon$ then $\mathrm{dist}(f,g) \leq \tilde{O}(1/\sqrt{\log(1/\epsilon)})$. Our new Theorem 7 provides a bound on $\mathrm{dist}(f,g)$ which is almost exponentially stronger than the [OS11] bound.

Theorem 7 should also be contrasted with Theorem 4 (the main result) of [Gol06], which says that for f an n-variable LTF and g any Boolean function, if $d_{\mathrm{Chow}}(f,g) \leq (\epsilon/n)^{O(\log(n/\epsilon)\log(1/\epsilon))}$ then $\mathrm{dist}(f,g) \leq \epsilon$. Phrased in this way, Theorem 7 says that for f an LTF and g any bounded function, if $d_{\mathrm{Chow}}(f,g) \leq \epsilon^{O(\log^2(1/\epsilon))}$ then $\mathrm{dist}(f,g) \leq \epsilon$. So our main structural result may be viewed as an improvement of Goldberg's result that removes its dependence on n. Indeed, this is not a coincidence; Theorem 7 is proved by carefully extending and strengthening Goldberg's arguments using the "critical index" machinery developed in recent studies of structural properties of LTFs [Ser07, OS11, DGJ+10].

[1] The theorem statement in [OS11] actually requires that g have range $\{-1,1\}$, but the proof is easily seen to extend to $g : \{-1,1\}^n \to [-1,1]$ as well.

It is natural to wonder whether the conclusion of Theorem 7 can be strengthened to "$\mathrm{dist}(f,g) \leq \epsilon^c$" where $c > 0$ is some absolute constant. We show that no such strengthening is possible, and in fact, no conclusion of the form "$\mathrm{dist}(f,g) \leq 2^{-\gamma(\epsilon)}$" is possible for any function $\gamma(\epsilon) = \omega(\log(1/\epsilon)/\log\log(1/\epsilon))$; we prove this in Section 4.2.

1.4 The algorithmic component.

A straightforward inspection of the arguments in [OS11] shows that by using our new Theorem 7 in place of Theorem 1.6 of that paper throughout, the running time of the [OS11] algorithm can be improved to $\mathrm{poly}(n)\cdot 2^{(1/\epsilon)^{O(\log^2(1/\epsilon))}}$. This is already a significant improvement over the $\mathrm{poly}(n)\cdot 2^{2^{\tilde{O}(1/\epsilon^2)}}$ running time of [OS11], but is significantly worse than the $\mathrm{poly}(n)\cdot(1/\epsilon)^{O(\log^2(1/\epsilon))}$ running time which is our ultimate goal.

The second key ingredient of our results is a new algorithm for constructing an LTF from the (approximate) Chow parameters of an LTF f. The previous approach to this problem [OS11] constructed an LTF with Chow parameters close to $\vec{\chi}_f$ directly and applied the structural result to the constructed LTF. Instead, our approach is based on the insight that it is substantially easier to find a bounded real-valued function g that is close to f in Chow distance. The structural result can then be applied to g to conclude that g is close to f in L_1-distance. The problem with this idea is, of course, that we need an LTF that is close to f and not a general bounded function. However, we show that it is possible to find g which is a "linear bounded function" (LBF), a type of bounded function closely related to LTFs. An LBF can then be easily converted to an LTF with only a small increase in distance from f. We now proceed to define the notion of an LBF and state our main algorithmic result formally. We first need to define the notion of a projection:

DEFINITION 8. *For a real value a, we denote its projection to $[-1,1]$ by $P_1(a)$. That is, $P_1(a) = a$ if $|a| \leq 1$ and $P_1(a) = \mathrm{sign}(a)$, otherwise.*

DEFINITION 9. *A function $g : \{-1,1\}^n \to [-1,1]$ is referred to as a* linear bounded function *(LBF) if there exists a vector of real values $w = (w_0, w_1, \ldots, w_n)$ such that $g(x) = P_1(w_0 + \sum_{i=1}^n w_i x_i)$. The vector w is said to represent g.*

We are now ready to state our main algorithmic result:

THEOREM 10 (MAIN ALGORITHMIC RESULT). *There exists a randomized algorithm* `ChowReconstruct` *that for every Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, given $\epsilon > 0, \delta > 0$ and a vector $\vec{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_n)$ such that $\|\vec{\chi}_f - \vec{\alpha}\| \leq \epsilon$, with probability at least $1-\delta$, outputs an LBF g such that $\|\vec{\chi}_f - \vec{\chi}_g\| \leq 6\epsilon$. The algorithm runs in time $\tilde{O}(n^2\epsilon^{-4}\log(1/\delta))$. Further, g is represented by a weight vector $\kappa v \in \mathbb{R}^{n+1}$, where $\kappa \in \mathbb{R}$ and v is an integer vector of length $\|v\| = O(\sqrt{n}/\epsilon^3)$.*

We remark that the condition on the weight vector v given by Theorem 10 is the key for the proof of Theorem 2.

Note that the running time of `ChowReconstruct` is polynomial in the relation between Chow distance and L_1-distance. By the structural result of [BDJ+98], this implies that for the subclass of LTFs with integer weights of magnitude bounded by $\mathrm{poly}(n)$, we obtain a $\mathrm{poly}(n/\epsilon)$ time algorithm, i.e. an FPRAS.

Discussion. It is interesting to note that the approach underlying Theorem 10 is much more efficient and significantly simpler than the algorithmic approach of [OS11]. The algorithm in [OS11] roughly works as follows: In the first step, it constructs a "small" set of candidate LTFs such that at least one of them is close to f, and in the second step it identifies such an LTF by searching over all such candidates. The first step proceeds by enumerating over "all" possible weights assigned to the "high influence" variables. This brute force search makes the [OS11] algorithm very inefficient. Moreover, its proof of correctness requires some sophisticated spectral results from [MORS10], which make the approach rather complicated.

In this work, our algorithm is based on a boosting-based approach, which is novel in this context. Our approach is much more efficient than the brute force search of [OS11] and its analysis is much simpler, since it completely bypasses the spectral results of [MORS10]. We also note that the algorithm of [OS11] crucially depends on the fact that the relation between Chow distance and distance has no dependence on n. (If this was not the case, the approach would not lead to a polynomial time algorithm.) Our boosting-based approach is quite robust, as it has no such limitation. This fact is crucial for us to obtain the aforementioned FPRAS for small-weight LTFs.

While we are not aware of any prior results similar to Theorem 10 being stated explicitly, we note that weaker forms of our theorem can be obtained from known results. In particular, Trevisan *et al.* [TTV09] describe an algorithm that given oracle access to a Boolean function f, $\epsilon' > 0$, and a set of functions $H = \{h_1, h_2, \ldots h_k\}$, efficiently finds a bounded function g that for every $i \leq n$ satisfies $|\mathbf{E}[f \cdot h_i] - \mathbf{E}[g \cdot h_i]| \leq \epsilon'$. One can observe that if $H = \{1, x_1, \ldots, x_n\}$, then the function g returned by their algorithm is in fact an LBF and that the oracle access to f can be replaced with approximate values of $\mathbf{E}[f \cdot h_i]$ for every i. Hence, the algorithm in [TTV09], applied to the set of functions $H = \{1, x_1, x_2, \ldots, x_n\}$, would find an LBF g which is close in Chow distance to f. A limitation of this algorithm is that, in order to obtain an LBF which is Δ-close in Chow distance to f, it requires that every Chow parameter of f be given to it with accuracy of $O(\Delta/\sqrt{n})$. In contrast, our algorithm only requires that the total distance of the given vector to $\vec{\chi}_f$ is at most $\Delta/6$. In addition, the bound on the integer weight approximation of LTFs that can be obtained from the algorithm in [TTV09] is linear in $n^{3/2}$, whereas we obtain the optimal dependence of $\sqrt{n}$.

The algorithm in [TTV09] is a simple adaptation of the hardcore set construction technique of Impagliazzo [Imp95]. Our algorithm is also based on the ideas from [Imp95] and, in addition, uses ideas from the distribution-specific boosting technique in [Fel10].

Learning (or approximating) a function by constructing a bounded function with approximately the same low-degree Fourier coefficients is also used in a very recent algorithm for learning a certain of class of polynomial threshold functions (which includes polynomial-size DNF formulae) [Fel12]. The algorithm in [Fel12] uses the same boosting-based approach but, like the algorithm in [TTV09], requires that every low-degree Fourier coefficient be given to it with high accuracy. As a result it would be similarly less efficient in our application. The application of this approach in [Fel12] is based on a substantially simpler structural result (a generalization of the one we use to obtain our FPRAS for small-weight LTFs).

Organization. In Section 2 we prove the main structural result, Theorem 7. In Section 3 we give our main algorithmic ingredient, Theorem 10. Section 4 puts the pieces together and proves our main theorem, Theorem 15, and our other main results.

2. PROOF OF MAIN STRUCTURAL RESULT: THEOREM 7

In this section we prove Theorem 7, restated here for convenience:

THEOREM 7 (MAIN STRUCTURAL RESULT). *Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF and $g : \{-1,1\}^n \to [-1,1]$ be any bounded function. If $d_{\mathrm{Chow}}(f,g) \le \epsilon$ then $\mathrm{dist}(f,g) \le 2^{-\Omega\left(\sqrt[3]{\log(1/\epsilon)}\right)}$.*

We give an informal overview of the main ideas of the proof of Theorem 7 in Section 2.1, and then give the actual proof outline of Theorem 7 in Section 2.2.

2.1 Proof overview of Theorem 7.

We first note that throughout the informal explanation given in this subsection, for the sake of clarity we restrict our attention to the case in which $g : \{-1,1\}^n \to \{-1,1\}$ is a Boolean rather than a bounded function. In the actual proof we deal with bounded functions using a suitable weighting scheme for points of $\{-1,1\}^n$ (see the discussion before Fact 31 near the start of the proof of Theorem 7).

To better explain our approach, we begin with a few words about how Theorem 1.6 of [OS11] (the only previously known statement of this type that is "independent of n") is proved. The key to that theorem is a result on approximating LTFs using LTFs with "good anti-concentration"; more precisely, [OS11] shows that for any LTF f there is an LTF $f'(x) = \mathrm{sign}(v \cdot x - \nu), \|v\| = 1$, that is extremely close to f (Hamming distance roughly $2^{-1/\epsilon}$) and which has "moderately good anticoncentration at radius ϵ," in the sense that $\mathbf{Pr}[|v \cdot x - \nu| \le \epsilon] \le \tilde{O}(1/\sqrt{\log(1/\epsilon)})$. Given this, Theorem 1.6 of [OS11] is proved using a modification of the proof of the original Chow's Theorem. However, for this approach based on the original Chow proof to work, it is crucial that the Hamming distance between f and f' (namely $2^{-1/\epsilon}$) be very small compared to the anti-concentration radius (which is ϵ). Subject to this constraint it seems very difficult to give a significant quantitative improvement of the approximation result in a way that would improve the bound of Theorem 1.6 of [OS11].

Instead, we hew more closely to the approach used to prove Theorem 4 of [Gol06]. This approach also involves a perturbation of the LTF f, but instead of measuring closeness in terms of Hamming distance, a more direct geometric view is taken. In the rest of this subsection we give a high-level explanation of Goldberg's proof and of how we modify it to obtain our improved bound.

The key to Goldberg's approach is a (perhaps surprising) statement about the geometry of hyperplanes as they relate to the Boolean hypercube. He establishes the following key geometric result (see Theorem 13 for a precise statement):

> If $\mathbf{H}$ is any n-dimensional hyperplane such that an α fraction of points in $\{-1,1\}^n$ lie "very close" in Euclidean distance (essentially $1/\mathrm{quasipoly}(n/\alpha)$) to $\mathbf{H}$, then there is a hyperplane $\mathbf{H}'$ which actually *contains* all those $\alpha 2^n$ points of the hypercube.

With this geometric statement in hand, an iterative argument is used to show that if the Hamming distance between LTF f and Boolean function g is large, then the Euclidean distance between the centers of mass of (the positive examples for f on which f and g differ) and (the negative examples for f on which f and g differ) must be large; finally, this Euclidean distance between centers of mass corresponds closely to the Chow distance between f and g.

However, the $1/\mathrm{quasipoly}(n)$ closeness requirement in the key geometric statement means that Goldberg's Theorem 4 not only depends on n, but this dependence is superpolynomial. The heart of our improvement is to combine Goldberg's key geometric statement with ideas based on the "critical index" of LTFs to get a version of the statement which is completely independent of n. Roughly speaking, our analogue of Goldberg's key geometric statement is the following (a precise version is given as Lemma 14 below):

> If $\mathbf{H}$ is any n-dimensional hyperplane such that an α fraction of points in $\{-1,1\}^n$ lie within Euclidean distance $\alpha^{O(\log(1/\alpha))}$ of $\mathbf{H}$, then there is a hyperplane $\mathbf{H}'$ which contains *all but a tiny fraction* of those $\alpha 2^n$ points of the hypercube.

Our statement is much stronger than Goldberg's in that there is no dependence on n in the distance bound from $\mathbf{H}$, but weaker in that we do not guarantee $\mathbf{H}'$ passes through every point; it may miss a tiny fraction of points, but we are able to handle this in the subsequent analysis. Armed with this improvement, a careful sharpening of Goldberg's iterative argument (to get rid of another dependence on n, unrelated to the tiny fraction of points missed by $\mathbf{H}'$) lets us prove Theorem 7.

2.2 Chow vs. Hamming distance for bounded functions: Proof of Theorem 7.

As discussed in Section 2.1, the key to proving Theorem 7 is an improvement of Theorem 3 in [Gol06].

DEFINITION 11. *Given a hyperplane $\mathbf{H}$ in $\mathbb{R}^n$ and $\beta > 0$, the β-*neighborhood *of $\mathbf{H}$ is defined as the set of points in $\mathbb{R}^n$ at Euclidean distance at most β from $\mathbf{H}$.*

We recall the following fact which shows how to express the Euclidean distance of a point from a hyperplane using the standard representation of the hyperplane:

FACT 12. *Let $\mathbf{H} = \{x : w \cdot x - \theta = 0\}$ be a hyperplane in $\mathbb{R}^n$ where $\|w\| = 1$. Then for any $x \in \mathbb{R}^n$, the Euclidean distance $d(x,\mathbf{H})$ of x from $\mathbf{H}$ is $|w \cdot x - \theta|$.*

THEOREM 13 (THEOREM 3 IN [GOL06]). *Given any hyperplane in $\mathbb{R}^n$ whose β-neighborhood contains a subset S of vertices of $\{-1,1\}^n$, where $|S| = \alpha \cdot 2^n$, there exists a hyperplane which contains all elements of S provided that*

$$0 \le \beta \le \left((2/\alpha) \cdot n^{5+\lfloor \log(n/\alpha) \rfloor} \cdot (2 + \lfloor \log(n/\alpha)! \rfloor)\right)^{-1}.$$

Before stating our improved version of the above theorem, we define the set $U = \cup_{i=1}^n \mathbf{e}_i \cup \mathbf{0}$ where $\mathbf{0} \in \mathbb{R}^n$ is the all zeros vector and $\mathbf{e}_i \in \mathbb{R}^n$ is the unit vector in the i^{th} direction.

Our improved version of Theorem 13 is the following:

LEMMA 14. *Let $\mathbf{H}$ be a hyperplane in $\mathbb{R}^n$ whose β-neighborhood contains a subset S of vertices of $\{-1,1\}^n$, where $|S| = \alpha \cdot 2^n$. Fix $0 < \kappa < \alpha/2$. Then there exists a hyperplane $\mathbf{H}'$ in $\mathbb{R}^n$ that contains a subset $S^* \subseteq S$ of cardinality at least $(\alpha - \kappa) \cdot 2^n$ provided that $0 \le \beta \le \beta_0$, where*

$$\beta_0 \stackrel{def}{=} (\log(1/\kappa))^{-1/2} \cdot (\log\log(1/\kappa))^{-O(\log\log\log(1/\kappa))} \cdot \alpha^{O(\log(1/\alpha))}.$$

Moreover, the coefficient vector defining $\mathbf{H}'$ has at most

$$O\left((1/\alpha^2) \cdot \left(\log\log(1/\kappa) + \log^2(1/\alpha)\right)\right)$$

nonzero coordinates. Further, for any $x \in U$, if x lies on $\mathbf{H}$ then x lies on $\mathbf{H}'$ as well.

Discussion. We note that while Lemma 14 may appear to be incomparable to Theorem 13 because it "loses" $\kappa 2^n$ points from the

set S, in fact by taking $\kappa = 1/2^{n+1}$ it must be the case that our S^* is the same as S, and with this choice of κ, Lemma 14 gives a strict quantitative improvement of Theorem 13. (We stress that for our application, though, it will be crucial for us to use Lemma 14 by setting the κ parameter to depend only on α independent of n.) We further note that in any statement like Lemma 14 that does not "lose" any points from S, the bound on β must necessarily depend on n; we show this in Appendix F. Finally, the condition at the end of Lemma 14 (that if $x \in U$ lies on $\mathbf{H}$, then it lies on $\mathbf{H}'$ as well) is something we will require later for technical reasons.

We give the detailed proof of Lemma 14 in Appendix C.2. We now briefly sketch the main idea underlying the proof of the lemma. At a high level, the proof proceeds by reducing the number of variables from n down to

$$m \stackrel{\text{def}}{=} O\left((1/\alpha^2) \cdot (\log(1/\beta) + \log\log(1/\kappa))\right)$$

followed by an application of Theorem 44, a technical generalization of Theorem 13 proved in Appendix G, in $\mathbb{R}^m$. (As we will see later, we use Theorem 44 instead of Theorem 13 because we need to ensure that points of U which lie on $\mathbf{H}$ continue to lie on $\mathbf{H}'$.) The reduction uses the notion of the τ-critical index applied to the vector w defining $\mathbf{H}$. (See Appendix C.1 for the relevant definitions.)

The idea of the proof is that for coordinates i in the "tail" of w (intuitively, where $|w_i|$ is small) the value of x_i does not have much effect on $d(x, \mathbf{H})$, and consequently the condition of the lemma must hold true in a space of much lower dimension than n. To show that tail coordinates of x do not have much effect on $d(x, \mathbf{H})$, we do a case analysis based on the τ-critical index $c(w, \tau)$ of w to show that (in both cases) the 2-norm of the entire "tail" of w must be small. If $c(w, \tau)$ is large, then this fact follows easily by properties of the τ-critical index. On the other hand, if $c(w, \tau)$ is small we argue by contradiction as follows: By the definition of the τ-critical index and the Berry-Esséen theorem, the "tail" of w (approximately) behaves like a normal random variable with standard deviation equal to its 2-norm. Hence, if the 2-norm was large, the entire linear form $w \cdot x$ would have good anti-concentration, which would contradict the assumption of the lemma. Thus in both cases, we can essentially ignore the tail and make the effective number of variables be m which is independent of n.

As described earlier, we view the geometric Lemma 14 as the key to the proof of Theorem 7; however, to obtain Theorem 7 from Lemma 14 requires a delicate iterative argument, which we give in full in Appendix C.3. This argument is essentially a refined version of Theorem 4 of [Gol06] with two main modifications: one is that we generalize the argument to allow g to be a bounded function rather than a Boolean function, and the other is that we get rid of various factors of $\sqrt{n}$ which arise in the [Gol06] argument (and which would be prohibitively "expensive" for us). We give the detailed proof in Appendix C.3.

3. THE ALGORITHM

In this section we give a proof overview of Theorem 10, restated below for convenience. We give the formal details of the proof in Appendix D.

THEOREM 10 (MAIN ALGORITHMIC RESULT). *There exists a randomized algorithm* `ChowReconstruct` *that for every Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, given $\epsilon > 0, \delta > 0$ and a vector $\vec{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_n)$ such that $\|\vec{\chi}_f - \vec{\alpha}\| \leq \epsilon$, with probability at least $1 - \delta$, outputs an LBF g such that $\|\vec{\chi}_f - \vec{\chi}_g\| \leq 6\epsilon$. The algorithm runs in time $\tilde{O}(n^2\epsilon^{-4}\log(1/\delta))$. Further, g is represented by a weight vector $\kappa v \in \mathbb{R}^{n+1}$, where $\kappa \in \mathbb{R}$ and v is an integer vector of length $\|v\| = O(\sqrt{n}/\epsilon^3)$.*

PROOF OUTLINE. Our algorithm is motivated by the following intuitive reasoning: since the function $\alpha_0 + \sum_{i\in[n]} \alpha_i \cdot x_i$ has the desired Chow parameters, why not just use it to define an LBF g_1 as $P_1(\alpha_0 + \sum_{i\in[n]} \alpha_i \cdot x_i)$? The answer, of course, is that as a result of applying the projection operator, the Chow parameters of g_1 can become quite different from the desired vector $\vec{\alpha}$. Nevertheless, it seems quite plausible to expect that g_1 will be better than a random guess.

Given the Chow parameters of g_1 we can try to correct them by adding the difference between $\vec{\alpha}$ and $\vec{\chi}_{g_1}$ to the vector that represents g_1. Again, intuitively we are adding a real-valued function $h_1 = \alpha_0 - \hat{g}_1(0) + \sum_{i\in[n]}(\alpha_i - \hat{g}_1(i)) \cdot x_i$ with the Chow parameters that we would like to add to the Chow parameters of g_1. And, again, the projection operation is likely to ruin our intention but we could still hope that we got closer to f and that by doing this operation for a while we will converge to an LBF with Chow parameters close to $\vec{\alpha}$.

While this idea might appear too naive, this is almost exactly what we do in `ChowReconstruct`. The main difference between this naive proposal and our actual algorithm is that at step t we actually add only half the difference between $\vec{\alpha}$ and the Chow vector of the current hypothesis $\vec{\chi}_{g_t}$. This is necessary in our proof to offset the fact that $\vec{\alpha}$ is only an approximation to $\vec{\chi}_f$ and we can only approximate the Chow parameters of g_t. An additional minor modification is required to ensure that the resulting weight vector is a multiple of an integer weight vector of length $O(\sqrt{n}/\epsilon^3)$.

Proving the correctness of this algorithm is also relatively easy. If the difference vector is sufficiently large (namely, more than a small multiple of the difference between $\|\vec{\chi}_f - \vec{\alpha}\|$) then the linear function h_t defined by this vector can be easily seen as being correlated with $f - g_t$, namely $\mathbf{E}[(f - g_t)h_t] \geq c\|\vec{\chi}_{g_t} - \vec{\alpha}\|^2$ for a constant $c > 0$. As was shown in [TTV09] and [Fel10] this condition for a Boolean h_t can be used to decrease a simple potential function measuring $\mathbf{E}[(f - g_t)^2]$, the l_2^2 distance of the current hypothesis to f. One issue that arises is this: while the l_2^2 distance is only reduced if h_t is added to g_t, in order to ensure that g_{t+1} is an LBF, we need to add the vector of difference (used to define h_t) to the weight vector representing g_t. To overcome this problem the proof in [TTV09] uses an additional point-wise counting argument from [Imp95]. This counting argument can be adapted to the real valued h_t, but the resulting argument becomes quite cumbersome. Instead, we augment the potential function in a way that captures the additional counting argument from [Imp95] and easily generalizes to the real-valued case.

4. THE MAIN RESULTS

4.1 Proofs of Theorems 1 and 2.

In this subsection we put the pieces together and prove our main results. We start by giving a formal statement of Theorem 1:

THEOREM 15 (MAIN). *There is a function $\kappa(\epsilon) \stackrel{def}{=} 2^{-O(\log^3(1/\epsilon))}$ such that the following holds: Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF and let $0 < \epsilon, \delta < 1/2$. Write $\vec{\chi}_f$ for the Chow vector of f and assume that $\vec{\alpha} \in \mathbb{R}^{n+1}$ is a vector satisfying $\|\vec{\alpha} - \vec{\chi}_f\| \leq \kappa(\epsilon)$. Then, there is an algorithm $\mathcal{A}$ with the following property: Given as input $\vec{\alpha}$, ϵ and δ, $\mathcal{A}$ performs $\tilde{O}(n^2) \cdot \mathrm{poly}(1/\kappa(\epsilon)) \cdot \log(1/\delta)$ bit operations and outputs the (weights-based) representation of an LTF f^* which with probability at least $1-\delta$ satisfies $\mathrm{dist}(f, f^*) \leq \epsilon$.*

PROOF OF THEOREM 15. Suppose that we are given a vector $\vec{\alpha} \in \mathbb{R}^{n+1}$ that satisfies $\Delta := \|\vec{\alpha} - \vec{\chi}_f\| \leq \kappa(\epsilon)$, where f is the unknown LTF to be learned. To construct the desired f^*, we run algorithm `ChowReconstruct` (from Theorem 10) on input $\vec{\alpha}$. The algorithm runs in time $\mathrm{poly}(1/\Delta) \cdot \tilde{O}(n^2) \cdot \log(1/\delta)$ and outputs an LBF g such that with probability at least $1-\delta$ we have $d_{\mathrm{Chow}}(f,g) \leq 6\Delta \leq 6\kappa(\epsilon)$. (We can set the constants appropriately in the definition of the function $\kappa(\epsilon)$ above, so that the quantity on the RHS of the latter relation is smaller than the "quasi-polynomial" quantity we need in the main structural theorem, so that the conclusion is "$\mathrm{dist}(f,g) \leq \epsilon/2$".) By Theorem 7 we get that with probability at least $1-\delta$ we have $\mathrm{dist}(f,g) \leq \epsilon/2$. Writing the LBF g as $g(x) = P_1(v_0 + \sum_{i=1}^n v_i x_i)$, we now claim that $f^*(x) = \mathrm{sign}(v_0 + \sum_{i=1}^n v_i x_i)$ has $\mathrm{dist}(f,f^*) \leq \epsilon$. This is simply because for each input $x \in \{-1,1\}^n$, the contribution that x makes to to $\mathrm{dist}(f,f^*)$ is at most twice the contribution x makes to $\mathrm{dist}(f,g)$. This completes the proof of Theorem 15. □

As a simple corollary, we obtain Theorem 2.

PROOF OF THEOREM 2. Let $f : \{-1,1\}^n \to \{-1,1\}$ be an arbitrary LTF. We apply Theorem 15 above, for $\delta = 1/3$, and consider the LTF f^* produced by the above proof. Note that the weights v_i defining f^* are identical to the weights of the LBF g output by the algorithm `ChowReconstruct`. It follows from Theorem 10 that these weights are integers that satisfy $\sum_{i=1}^n v_i^2 = O(n \cdot \Delta^{-6})$, where $\Delta = \Omega(\kappa(\epsilon))$, and the proof is complete. □

As pointed out in Section 1.2 our algorithm runs in $\mathrm{poly}(n/\epsilon)$ time for LTFs whose integer weight is at most $\mathrm{poly}(n)$. Formally, we have:

THEOREM 16. *Let $f = \mathrm{sign}(\sum_{i=1}^n w_i x_i - \theta)$ be an LTF with integer weights w_i such that $W \stackrel{def}{=} \sum_{i=1}^n |w_i| = \mathrm{poly}(n)$. Fix $0 < \epsilon, \delta < 1/2$. Write $\vec{\chi}_f$ for the Chow vector of f and assume that $\vec{\alpha} \in \mathbb{R}^{n+1}$ is a vector satisfying $\|\vec{\alpha} - \vec{\chi}_f\| \leq \epsilon/(12W)$. Then, there is an algorithm $\mathcal{A}'$ with the following property: Given as input $\vec{\alpha}$, ϵ and δ, $\mathcal{A}'$ performs $\mathrm{poly}(n/\epsilon) \cdot \log(1/\delta)$ bit operations and outputs the (weights-based) representation of an LTF f^* which with probability at least $1-\delta$ satisfies $\mathrm{dist}(f,f^*) \leq \epsilon$.*

The algorithm and proof of the above theorem are identical to the ones in Theorem 15 above, except that we apply a simpler structural result from [BDJ+98] relating $d_{\mathrm{Chow}}(f,g)$ and $\mathrm{dist}(f,g)$ when f is a LTF with small integer weights rather than applying our Theorem 7.

4.2 Near-optimality of Theorem 7.

Theorem 7 says that if f is an LTF and $g : \{-1,1\}^n \to [-1,1]$ satisfy $d_{\mathrm{Chow}}(f,g) \leq \epsilon$ then $\mathrm{dist}(f,g) \leq 2^{-\Omega(\sqrt[3]{\log(1/\epsilon)})}$. It is natural to wonder whether the conclusion can be strengthened to "$\mathrm{dist}(f,g) \leq \epsilon^c$" where $c > 0$ is some absolute constant. Here we observe that no conclusion of the form "$\mathrm{dist}(f,g) \leq 2^{-\gamma(\epsilon)}$" is possible for any function $\gamma(\epsilon) = \omega(\log(1/\epsilon)/\log\log(1/\epsilon))$.

To see this, fix γ to be any function such that

$$\gamma(\epsilon) = \omega(\log(1/\epsilon)/\log\log(1/\epsilon)).$$

If there were a stronger version of Theorem 7 in which the conclusion is "then $\mathrm{dist}(f,g) \leq 2^{-\gamma(\epsilon)}$," the arguments of Section 4.1 would give that for any LTF f, there is an LTF $f' = \mathrm{sign}(v \cdot x - \nu)$ such that $\mathbf{Pr}[f(x) \neq f'(x)] \leq \epsilon$, where each $v_i \in \mathbb{Z}$ satisfies $|v_i| \leq \mathrm{poly}(n) \cdot (1/\epsilon)^{o(\log\log(1/\epsilon))}$. Taking $\epsilon = 1/2^{n+1}$, this tells us that f' must agree with f on every point in $\{-1,1\}^n$, and each integer weight in the representation $\mathrm{sign}(v \cdot x - \nu)$ is at most $2^{o(n \log n)}$. But choosing f to be Håstad's function from [Hås94], this is a contradiction, since any integer representation of that function must have every $|v_i| \geq 2^{\Omega(n \log n)}$.

4.3 Applications to computational learning theory.

In this section we state some consequences of our approach in algorithmic learning theory. For a more detailed discussion (with the proofs), see Appendix E. The first application is to the "Restricted Focus of Attention" (RFA) learning framework introduced by Ben-David and Dichterman [BDD98].

THEOREM 17. *There is an algorithm which performs $\tilde{O}(n^2) \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))} \cdot \log(\frac{1}{\delta})$ bit-operations and properly learns LTFs to accuracy ϵ and confidence $1-\delta$ in the uniform distribution 1-RFA model.*

A variant of our technique also results in a very fast "agnostic-type" algorithm for learning LTFs under the uniform distribution.

THEOREM 18. *There is an algorithm $\mathcal{B}$ with the following performance guarantee: Let f be any Boolean function and let $\mathsf{opt} = \mathrm{dist}(f,\mathcal{H})$. Given $0 < \epsilon, \delta < 1/2$ and access to independent uniform examples $(x, f(x))$, algorithm $\mathcal{B}$ outputs the (weights-based) representation of an LTF f^* which with probability $1-\delta$ satisfies $\mathrm{dist}(f^*,f) \leq 2^{-\Omega(\sqrt[3]{\log(1/\mathsf{opt})})} + \epsilon$. The algorithm performs $\tilde{O}(n^2) \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))} \cdot \log(1/\delta)$ bit operations.*

The case $\mathsf{opt} = 0$ corresponds to learning LTFs under the uniform distribution with no noise (non-agnostic learning). In this case it is easy to see that the above proof gives a $\tilde{O}(n^2) \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))}$ time algorithm, which is a significant improvement over the $\tilde{O}(n^2) \cdot 2^{\mathrm{poly}(1/\epsilon)}$ running time of the previous fastest algorithm [OS11].

5. REFERENCES

[APL07] H. Aziz, M. Paterson, and D. Leech. Efficient algorithm for designing weighted voting games. In *IEEE Intl. Multitopic Conf.*, pages 1–6, 2007.

[Bau73] C. R. Baugh. Chow parameters in pseudothreshold logic. In *SWAT (FOCS)*, pages 49–55, 1973.

[BDD98] S. Ben-David and E. Dichterman. Learning with restricted focus of attention. *Journal of Computer and System Sciences*, 56(3):277–298, 1998.

[BDJ+98] A. Birkendorf, E. Dichterman, J. Jackson, N. Klasner, and H.U. Simon. On restricted-focus-of-attention learnability of Boolean functions. *Machine Learning*, 30:89–123, 1998.

[Bru90] J. Bruck. Harmonic analysis of polynomial threshold functions. *SIAM Journal on Discrete Mathematics*, 3(2):168–177, 1990.

[Car04] F. Carreras. On the design of voting games. *Mathematical Methods of Operations Research*, 59(3):503–515, 2004.

[CHIS10] M. Cheraghchi, J. Håstad, M. Isaksson, and O. Svensson. Approximating Linear Threshold Predicates. In *13th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems — APPROX 2010*, pages 110–123, 2010.

[Cho61] C.K. Chow. On the characterization of threshold functions. In *Proceedings of the Symposium on Switching Circuit Theory and Logical Design (FOCS)*, pages 34–38, 1961.

[Der65] M. Dertouzos. *Threshold Logic: A Synthesis Approach*. MIT Press, Cambridge, MA, 1965.

[DGJ+10] I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. *SIAM J. on Comput.*, 39(8):3441–3462, 2010.

[DS79] P. Dubey and L.S. Shapley. Mathematical properties of the Banzhaf power index. *Mathematics of Operations Research*, 4:99–131, 1979.

[DS09] I. Diakonikolas and R. Servedio. Improved approximation of linear threshold functions. In *Proc. 24th Annual IEEE Conference on Computational Complexity (CCC)*, pages 161–172, 2009.

[EL89] E. Einy and E. Lehrer. Regular simple games. *International Journal of Game Theory*, 18:195–207, 1989.

[Fel68] W. Feller. *An introduction to probability theory and its applications*. John Wiley & Sons, 1968.

[Fel10] V. Feldman. Distribution-specific agnostic boosting. In *Proceedings of Innovations in Computer Science*, pages 241–250, 2010.

[Fel12] V. Feldman. Learning DNF expressions from Fourier spectrum. *CoRR*, abs/1203.0594, 2012.

[FGRW09] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. In *FOCS*, pages 385–394, 2009.

[FM04] D. Felsenthal and M. Machover. A priori voting power: what is it all about? *Political Studies Review*, 2(1):1–23, 2004.

[Fre97] J. Freixas. Different ways to represent weighted majority games. *Top (Journal of the Spanish Society of Statistics and Operations Research)*, 5(2):201–212, 1997.

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

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

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

[Hur73] S.L. Hurst. The application of Chow Parameters and Rademacher-Walsh matrices in the synthesis of binary functions. *The Computer Journal*, 16:165–173, 1973.

[Imp95] Russell Impagliazzo. Hard-core distributions for somewhat hard problems. In *Proc. 36th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 538–545. IEEE Computer Society Press, 1995.

[Kas63] P. Kaszerman. A geometric test-synthesis procedure for a threshold device. *Information and Control*, 6(4):381–398, 1963.

[KKMS05] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2005.

[KSS94] M. Kearns, R. Schapire, and L. Sellie. Toward Efficient Agnostic Learning. *Machine Learning*, 17(2/3):115–141, 1994.

[KW65] K.R. Kaplan and R.O. Winder. Chebyshev approximation and threshold functions. *IEEE Trans. Electronic Computers*, EC-14:315–325, 1965.

[Lap72] E. Lapidot. The counting vector of a simple game. *Proceedings of the AMS*, 31:228–231, 1972.

[Lee03] D. Leech. Power indices as an aid to institutional design: the generalised apportionment problem. In M. Holler, H.Kliemt, D. Schmidtchen, and M. Streit, editors, *Yearbook on New Political Economy*, 2003.

[MORS10] K. Matulef, R. O'Donnell, R. Rubinfeld, and R. Servedio. Testing halfspaces. *SIAM J. on Comput.*, 39(5):2004–2047, 2010.

[Odl88] A. M. Odlyzko. On subspaces spanned by random selections of ± 1 vectors. *J. Comb. Theory, Ser. A*, 47(1):124–133, 1988.

[OS11] R. O'Donnell and R. Servedio. The Chow Parameters Problem. *SIAM J. on Comput.*, 40(1):165–199, 2011.

[RSOK95] V.P. Roychowdhury, K.-Y. Siu, A. Orlitsky, and T. Kailath. Vector analysis of threshold functions. *Information and Computation*, 120(1):22–31, 1995.

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

[Tan61] M. Tannenbaum. The establishment of a unique representation for a linearly separable function. Technical report, Lockheed Missiles and Space Co., 1961. Threshold Switching Techniques Note 20, pp. 1-5.

[TT06] K. Takamiya and A. Tanaka. Computational complexity in the design of voting games. Technical Report 653, The Institute of Social and Economic Research, Osaka University, 2006.

[TTV09] Luca Trevisan, Madhur Tulsiani, and Salil P. Vadhan. Regularity, boosting, and efficiently simulating every high-entropy distribution. In *IEEE Conference on Computational Complexity*, pages 126–136, 2009.

[TV09] T.Tao and V. H. Vu. Inverse Littlewood-Offord theorems and the condition number of random discrete matrices. *Annals of Mathematics*, 169:595–632, 2009.

[TZ92] A. Taylor and W. Zwicker. A Characterization of Weighted Voting. *Proceedings of the AMS*, 115(4):1089–1094, 1992.

[Win63] R.O. Winder. Threshold logic in artificial intelligence. *Artificial Intelligence*, IEEE Publication S-142:107–128, 1963.

[Win69] R.O. Winder. Threshold gate approximations based on chow parameters. *IEEE Transactions on Computers*, pages 372–375, 1969.

[Win71] R.O. Winder. Chow parameters in threshold logic. *Journal of the ACM*, 18(2):265–289, 1971.

APPENDIX

A. PROOF OF CHOW'S THEOREM

For completeness we state and prove Chow's Theorem here:

THEOREM 19 ([CHO61]). *Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF and let $g : \{-1,1\}^n \to [-1,1]$ be a bounded function such that $\widehat{g}(j) = \widehat{f}(j)$ for all $0 \le j \le n$. Then $g = f$.*

PROOF. Write $f(x) = \text{sign}(w_0 + w_1 x_1 + \cdots + w_n x_n)$, where the weights are scaled so that $\sum_{j=0}^{n} w_j^2 = 1$. We may assume

without loss of generality that $|w_0 + w_1x_1 + \cdots + w_nx_n| \neq 0$ for all x. (If this is not the case, first translate the separating hyperplane by slightly perturbing w_0 to make it hold; this can be done without changing f's value on any point of $\{-1,1\}^n$.) Now we have

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

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

$$\mathrm{sign}(f(x) - g(x)) = f(x) = \mathrm{sign}(w_0 + w_1x_1 + \cdots + w_nx_n)$$

for any bounded function g with range $[-1,1]$. But since $|w_0 + w_1x_1 + \cdots + w_nx_n|$ is always strictly positive, we must have $\mathbf{Pr}[f(x) \neq g(x)] = 0$ as claimed. □

B. USEFUL BACKGROUND

B.1 Probabilistic Facts.

We require some basic probability results including the standard additive Hoeffding bound:

THEOREM 20. *Let $X_1, \ldots, X_n$ be independent random variables such that for each $j \in [n]$, X_j is supported on $[a_j, b_j]$ for some $a_j, b_j \in \mathbb{R}$, $a_j \leq b_j$. Let $X = \sum_{j=1}^n X_j$. Then, for any $t > 0$, $\mathbf{Pr}\left[|X - \mathbf{E}[X]| \geq t\right] \leq 2\exp\left(-2t^2/\sum_{j=1}^n (b_j - a_j)^2\right)$.*

The Berry-Esséen theorem (see e.g. [Fel68]) gives explicit error bounds for the Central Limit Theorem:

THEOREM 21. *(Berry-Esséen) Let $X_1, \ldots, X_n$ be independent random variables satisfying $\mathbf{E}[X_i] = 0$ for all $i \in [n]$, $\sqrt{\sum_i \mathbf{E}[X_i^2]} = \sigma$, and $\sum_i \mathbf{E}[|X_i|^3] = \rho_3$. Let $S = (X_1 + \cdots + X_n)/\sigma$ and let F denote the cumulative distribution function (cdf) of S. Then $\sup_x |F(x) - \Phi(x)| \leq \rho_3/\sigma^3$ where Φ denotes the cdf of the standard gaussian random variable.*

An easy consequence of the Berry-Esséen theorem is the following fact, which says that a regular linear form has good anti-concentration (i.e. it assigns small probability mass to any small interval):

FACT 22. *Let $w = (w_1, \ldots, w_n)$ be a τ-regular vector in $\mathbb{R}^n$ and write σ to denote $\|w\|_2$. Then for any interval $[a,b] \subseteq \mathbb{R}$, we have $\left|\mathbf{Pr}[\sum_{i=1}^n w_ix_i \in (a,b]] - \Phi([a/\sigma, b/\sigma])\right| \leq 2\tau$, where $\Phi([c,d]) \stackrel{def}{=} \Phi(d) - \Phi(c)$. In particular, it follows that*

$$\mathbf{Pr}\left[\sum_{i=1}^{n} w_ix_i \in (a,b]\right] \leq |b-a|/\sigma + 2\tau.$$

B.2 Useful inequalities.

We will use the following basic inequalities.

FACT 23. *For $a, b \in (0,1)$,*

$$(ab)^{\log(1/a)+\log(1/b)} \geq a^{2\log(1/a)} \cdot b^{2\log(1/b)}.$$

PROOF.

$$\begin{aligned} (ab)^{\log(1/a)+\log(1/b)} &= 2^{-\log^2(1/a) - \log^2(1/b) - 2\log(1/a)\cdot\log(1/b)} \\ &\geq 2^{-2\log^2(1/a) - 2\log^2(1/b)} \\ &= a^{2\log(1/a)} \cdot b^{2\log(1/b)}, \end{aligned}$$

where the inequality is the Arithmetic-Geometric mean inequality. □

Similarly, we obtain:

FACT 24. *For $x, y \geq 1$,*

$$(x+y)^{-\log(x+y)} \geq (2x)^{-\log(2x)} \cdot (2y)^{-\log(2y)}.$$

B.3 Useful facts about affine spaces.

A subset $V \subseteq \mathbb{R}^n$ is said to be an *affine subspace* if it is closed under affine combinations of vectors in V. Equivalently, V is an affine subspace of $\mathbb{R}^n$ if $V = X + b$ where $b \in \mathbb{R}^n$ and X is a linear subspace of $\mathbb{R}^n$. The affine dimension of V is the same as the dimension of the linear subspace X. A hyperplane in $\mathbb{R}^n$ is an affine space of dimension $n-1$. Throughout the paper we use bold capital letters such as $\mathbf{H}$ to denote hyperplanes.

In this paper whenever we refer to a "subspace" we mean an affine subspace unless explicitly otherwise indicated. The dimension of an affine subspace V is denoted by $\dim(V)$. Similarly, for a set $S \subseteq \mathbb{R}^n$, we write $\mathrm{span}(S)$ to denote the affine span of S, i.e.

$$\mathrm{span}(S) = \{s + \sum_{i=1}^{m} w_i(x^i - y^i) \mid s, x^i, y^i \in S, w_i \in \mathbb{R}, m \in \mathbb{N}\}.$$

The following very useful fact about affine spaces was proved by Odlyzko[Odl88].

FACT 25. *[Odl88] Any affine subspace of $\mathbb{R}^n$ of dimension d contains at most 2^d elements of $\{-1,1\}^n$.*

C. MISSING PROOFS FROM SECTION 2

In this section we provide a detailed proof of our main structural result (Theorem 7).

C.1 Tools for Theorem 7.

As described above, a key ingredient in the proof of Theorem 7 is the notion of the "critical index" of an LTF f. The critical index was implicitly introduced and used in [Ser07] and was explicitly used in [DS09, DGJ+10, OS11] and other works. To define the critical index we need to first define "regularity":

DEFINITION 26 (REGULARITY). *Fix $\tau > 0$. We say that a vector $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ is τ-*regular *if $\max_{i\in[n]} |w_i| \leq \tau\|w\| = \tau\sqrt{w_1^2 + \cdots + w_n^2}$. A linear form $w \cdot x$ is said to be τ-regular if w is τ-regular, and similarly an LTF is said to be τ-regular if it is of the form $\mathrm{sign}(w \cdot x - \theta)$ where w is τ-regular.*

Regularity is a helpful notion because if w is τ-regular then the Berry-Esséen theorem (stated below) tells us that for uniform $x \in \{-1,1\}^n$, the linear form $w \cdot x$ is "distributed like a Gaussian up to error τ." This can be useful for many reasons; in particular, it will let us exploit the strong anti-concentration properties of the Gaussian distribution.

Intuitively, the critical index of w is the first index i such that from that point on, the vector $(w_i, w_{i+1}, \ldots, w_n)$ is regular. A precise definition follows:

DEFINITION 27 (CRITICAL INDEX). *Given a vector $w \in \mathbb{R}^n$ such that $|w_1| \geq \cdots \geq |w_n| > 0$, for $k \in [n]$ we denote by σ_k the quantity $\sqrt{\sum_{i=k}^{n} w_i^2}$. We define the* τ-critical index $c(w,\tau)$ of w *as the smallest index $i \in [n]$ for which $|w_i| \leq \tau \cdot \sigma_i$. If this inequality does not hold for any $i \in [n]$, we define $c(w,\tau) = \infty$.*

The following simple fact states that the "tail weight" of the vector w decreases exponentially prior to the critical index:

FACT 28. *For any vector $w = (w_1, \ldots, w_n)$ such that $|w_1| \geq \cdots \geq |w_n| > 0$ and $1 \leq a \leq c(w,\tau)$, we have $\sigma_a < (1-\tau^2)^{(a-1)/2} \cdot \sigma_1$.*

PROOF. If $a < c(w,\tau)$, then by definition $|w_a| > \tau \cdot \sigma_a$. This implies that $\sigma_{a+1} < \sqrt{1-\tau^2} \cdot \sigma_a$. Applying this inequality repeatedly, we get that $\sigma_a < (1-\tau^2)^{(a-1)/2} \cdot \sigma_1$ for any $1 \leq a \leq c(w,\tau)$. $\square$

C.2 Proof of Lemma 14.

At a high level, the proof proceeds by reducing the number of variables from n down to

$$m \stackrel{\text{def}}{=} O\left((1/\alpha^2) \cdot (\log(1/\beta) + \log\log(1/\kappa))\right)$$

followed by an application of Theorem 44, a technical generalization of Theorem 13 proved in Appendix G, in $\mathbb{R}^m$. (As we will see later, we use Theorem 44 instead of Theorem 13 because we need to ensure that points of U which lie on $\mathbf{H}$ continue to lie on $\mathbf{H}'$.) The reduction uses the notion of the τ-critical index applied to the vector w defining $\mathbf{H}$.

The idea of the proof is that for coordinates i in the "tail" of w (intuitively, where $|w_i|$ is small) the value of x_i does not have much effect on $d(x,\mathbf{H})$, and consequently the condition of the lemma must hold true in a space of much lower dimension than n. To show that tail coordinates of x do not have much effect on $d(x,\mathbf{H})$, we do a case analysis based on the τ-critical index $c(w,\tau)$ of w. If $c(w,\tau)$ is large then the 2-norm of the entire "tail" of w must be small, and if $c(w,\tau)$ is small then we use the regularity of the tail of w to show again that its 2-norm must be small. Thus in both cases, we can essentially ignore the tail and make the effective number of variables be m which is independent of n.

Let $0 < \tau < \alpha$. Let $\mathbf{H} = \{x \in \mathbb{R}^n \mid w \cdot x = \theta\}$ where we can assume (by rescaling) that $\|w\|_2 = 1$ and (by reordering the coordinates) that $|w_1| \geq |w_2| \geq \ldots \geq |w_n|$. Note that the Euclidean distance of any point $x \in \mathbb{R}^n$ from $\mathbf{H}$ is $|w \cdot x - \theta|$. Let us also define $V \stackrel{\text{def}}{=} \mathbf{H} \cap U$. Set $\tau \stackrel{\text{def}}{=} \alpha/4$ (for conceptual clarity we will continue to use "τ" for as long as possible in the arguments below). We consider the τ-critical index $c(w,\tau)$ of the vector $w \in \mathbb{R}^n$ and proceed by case analysis based on its value. Fix the parameter $K_0 \stackrel{\text{def}}{=} \Theta\left((1/\tau^2) \cdot (\log\log(1/\kappa) + \log(1/\beta))\right)$.

Case I: $c(w,\tau) > K_0$. In this case, we partition $[n]$ into a set of "head" coordinates $H = [K_0]$ and a complementary set of "tail" coordinates $T = [n] \setminus H$. Writing w as (w_H, w_T) and likewise for x, it follows from Fact 28 that $\|w_T\| \leq O(\beta/\sqrt{\log(1/\kappa)})$. By the Hoeffding bound, for $(1-\kappa)$ fraction of $x \in \{-1,1\}^n$ we have that $|w_T \cdot x_T| \leq \beta$. Therefore, for $(1-\kappa)$ fraction of $x \in \{-1,1\}^n$ we have

$$|w_H \cdot x_H - \theta| \leq |w \cdot x - \theta| + |w_T \cdot x_T| \leq |w \cdot x - \theta| + \beta.$$

By the assumption of the lemma, there exists a set $S \subseteq \{-1,1\}^n$ of cardinality at least $\alpha \cdot 2^n$ such that for all $x \in S$ we have $|w \cdot x - \theta| \leq \beta$. A union bound and the above inequality imply that there exists a set $S^* \subseteq S$ of cardinality at least $(\alpha - \kappa) \cdot 2^n$ with the property that for all $x \in S^*$, we have

$$|w_H \cdot x_H - \theta| \leq 2\beta.$$

Also, any $x \in U$ satisfies $\|x_T\| \leq 1$. Hence for any $x \in V$, we have that

$$\begin{aligned} |w_H \cdot x_H - \theta| &\leq |w \cdot x - \theta| + |w_T \cdot x_T| = |w_T \cdot x_T| \\ &\leq \|w_T\| \cdot \|x_T\| \leq O(\beta/\sqrt{\log(1/\kappa)}) \leq \beta. \end{aligned}$$

Define the projection mapping $\phi_H : \mathbb{R}^n \to \mathbb{R}^{|H|}$ by $\phi_H : x \mapsto x_H$ and consider the image of S^*, i.e. $S' \stackrel{\text{def}}{=} \phi_H(S^*)$. It is clear that $|S'| \geq (\alpha - \kappa) \cdot 2^{|H|}$ and that for all $x_H \in S'$, we have

$$|w_H \cdot x_H - \theta| \leq 2\beta.$$

Similarly, if V' is the image of V under ϕ_H, then for every $x_H \in V'$ we have $|w_H \cdot x_H - \theta| \leq \beta$. It is also clear that $\|w_T\| < 1/2$ and hence $\|w_H\| > 1/2$. Thus for every $x_H \in (S' \cup V')$ we have

$$\left| \frac{w_H \cdot x_H}{\|w_H\|} - \frac{\theta}{\|w_H\|} \right| \leq 4\beta.$$

We now define the K_0-dimensional hyperplane $\mathbf{H}_H$ as $\mathbf{H}_H \stackrel{\text{def}}{=} \{x_H \in \mathbb{R}^{|H|} \mid w_H \cdot x_H = \theta\}$. As all points in $S' \cup V'$ are in the 4β-neighborhood of $\mathbf{H}_H$, we may now apply Theorem 44 for the hyperplane $\mathbf{H}_H$ over $\mathbb{R}^{|H|}$ to deduce the existence of an alternate hyperplane $\mathbf{H}'_H = \{x_H \in \mathbb{R}^{|H|} \mid v_H \cdot x_H = \nu\}$ that contains all points in $S' \cup V'$. The only condition we need to verify in order that Theorem 44 may be applied is that 4β is upper bounded by

$$\left(\frac{2}{\alpha-\kappa} \cdot K_0^{5+\lfloor \log(K_0/(\alpha-\kappa)) \rfloor} \cdot \left(2 + \lfloor \log\left(\frac{K_0}{\alpha-\kappa}\right) \rfloor !\right)\right)^{-1}.$$

In the following C_1, C_2, etc. denote unspecified absolute positive constants. Using $\kappa < \alpha/2$, it suffices to ensure

$$\beta < (\alpha/K_0)^{C_1(\log(K_0/\alpha))}.$$

Recalling that $\tau = \alpha/4$ and plugging in the value of K_0 in terms of α, κ and β, we need to verify that

$$\beta < \left(\frac{\alpha^3}{\log\log(1/\kappa) + \log(1/\beta)}\right)^{C_2(\log(1/\alpha^3) + \log(\log\log(1/\kappa) + \log(1/\beta)))}.$$

Using Fact 23, we get that the right hand side is lower bounded by

$$\alpha^{C_3 \log(1/\alpha)} \cdot (\log\log(1/\kappa) + \log(1/\beta))^{-C_3 \log(\log\log(1/\kappa) + \log(1/\beta))}.$$

Using Fact 24, we get that the above expression is lower bounded by

$$\alpha^{C_4 \log(1/\alpha)} \cdot \log\log(1/\kappa)^{-C_4 \log\log\log(1/\kappa)} \cdot \log(1/\beta)^{-C_4 \log\log(1/\beta)}.$$

Thus it suffices to verify that

$$\beta \leq \alpha^{C_4 \log(1/\alpha)} \cdot \log\log(1/\kappa)^{-C_4 \log\log\log(1/\kappa)} \cdot \log(1/\beta)^{-C_4 \log\log(1/\beta)}.$$

It is easy to see that for

$$\beta \leq \alpha^{O(\log(1/\alpha))} \cdot \log\log(1/\kappa)^{-O(\log\log\log(1/\kappa))}$$

(with sufficiently large constants inside the $O(\cdot)$ notation), the above inequality is indeed true and hence it is true for $\beta \leq \beta_0$.

Thus, we get a new hyperplane $\mathbf{H}'_{K_0} = \{x_H \in \mathbb{R}^{|H|} \mid v_H \cdot x_H = \nu\}$ that contains all points in $S' \cup V'$. It is then clear that the n-dimensional hyperplane $\mathbf{H}' = \{x \in \mathbb{R}^n \mid v_H \cdot x_H = \nu\}$ contains all the points in $S^* = (\phi_H)^{-1}(S')$ and the points in V, and

that the vector v_H defining $\mathbf{H}'$ has the claimed number of nonzero coordinates. So the theorem is proved in Case I.

Case II: $c(w,\tau) \leq K_0$. In this case, we partition $[n]$ into "head" and "tail" based on the value of $c(w,\tau)$ by taking $H = [c(w,\tau)]$ and $T = [n] \setminus H$. We use the fact that w_T is τ-regular to deduce that the norm of the tail must be small.

CLAIM 29. *We have* $\|w_T\|_2 \leq 2\beta/(\alpha - 3\tau) = 8\beta/\alpha$.

PROOF. Suppose for the sake of contradiction that

$$\|w_T\|_2 > 2\beta/(\alpha - 3\tau).$$

By the Berry-Esséen theorem (Theorem 21, or more precisely Fact 22), for all $\delta > 0$ we have

$$\sup_{t\in\mathbb{R}} \mathbf{Pr}_{x_T}\left[|w_T \cdot x_T - t| \leq \delta\right] \leq \frac{2\delta}{\|w_T\|} + 2\tau.$$

By setting $\delta \stackrel{\text{def}}{=} (\alpha - 3\tau)\|w_T\|/2 > \beta$ we get that

$$\sup_{t\in\mathbb{R}} \mathbf{Pr}_{x_T}\left[|w_T \cdot x_T - t| \leq \delta\right] < \alpha,$$

and consequently

$$\begin{aligned}\mathbf{Pr}_x[|w\cdot x - \theta| \leq \beta] &\leq \sup_{t\in\mathbb{R}} \mathbf{Pr}_{x_T}[|w_T \cdot x_T - t| \leq \beta] \\ &\leq \sup_{t\in\mathbb{R}} \mathbf{Pr}_{x_T}\left[|w_T \cdot x_T - t| \leq \delta\right] \\ &< \alpha\end{aligned}$$

which contradicts the existence of the set S in the statement of the lemma. □

The rest of the proof proceeds similarly to Case I. By the Hoeffding bound, for $1-\kappa$ fraction of $x \in \{-1,1\}^n$ we have

$$|w_H \cdot x_H - \theta| \leq |w \cdot x - \theta| + \beta'$$

where $\beta' = O\left((\beta/\alpha)\cdot\sqrt{\log(1/\kappa)}\right)$. By the assumption of the lemma and a union bound, there exists a set $S^* \subseteq S$ of cardinality at least $(\alpha-\kappa)\cdot 2^n$ with the property that for all $x \in S^*$ we have

$$|w_H \cdot x_H - \theta| \leq \beta' + \beta.$$

Turning to V, for every point $x \in V$ we have that $|w_H \cdot x_H - \theta| \leq |w\cdot x - \theta| + |w_T \cdot x_T| = |w_T \cdot x_T|$. For $x \in V$ the value $w_T \cdot x_T$ is either 0 (if $x = \mathbf{0}$) or is $(w_T)_i$ (if $x = \mathbf{e}_i$) for some $i \in T$. Since w_T is τ-regular we have $|(w_T)_i| \leq \tau \cdot \|w_T\| \leq (\alpha/4)\cdot(8\beta/\alpha) = 2\beta$, so for every $x \in V$ we have $|w_H \cdot x_H - \theta| \leq 2\beta \leq \beta + \beta'$.

As before, we define the projection mapping $\phi_H : \mathbb{R}^n \to \mathbb{R}^{|H|}$ by $\phi_H : x \mapsto x_H$. We let $S' \stackrel{\text{def}}{=} \phi_H(S^*)$ and $V' \stackrel{\text{def}}{=} \phi_H(V)$. It is clear that $|S'| \geq (\alpha-\kappa)\cdot 2^{|H|}$ and that for all $x_H \in (S' \cup V')$ we have

$$|w_H \cdot x_H - \theta| \leq \beta' + \beta.$$

and that for all $x_H \in V'$, $|w_H \cdot x_H - \theta| \leq \beta$. We now define the $|H|$-dimensional hyperplane $\mathbf{H}_H$ as $\mathbf{H}_H \stackrel{\text{def}}{=} \{x_H \in \mathbb{R}^{|H|} \mid w_H \cdot x_H = \theta\}$. As before, we note that $\|w_T\| < 1/2$ and hence $\|w_H\| > 1/2$. Hence, every point $x_H \in S' \cup V'$ is $2(\beta+\beta') \leq 4\beta'$ close to $\mathbf{H}_H$. As all points in $S' \cup V'$ are $4\beta'$ close to $\mathbf{H}_H$, we may now apply Theorem 44 over $\mathbb{R}^{|H|}$ to deduce the existence of an alternate hyperplane $\mathbf{H}'_H \stackrel{\text{def}}{=} \{x_H \in \mathbb{R}^{|H|} \mid v_H \cdot x_H = \nu\}$ that contains all points in S' and V'. The only condition we need to verify is that $4\beta'$ is at most

$$\left(\frac{2}{\alpha-\kappa}\cdot|H|^{5+\lfloor\log(|H|/(\alpha-\kappa))\rfloor}\cdot(2+\lfloor\log(|H|/(\alpha-\kappa))\rfloor!)\right)^{-1}.$$

As $\beta' = O((\beta\sqrt{\log(1/\kappa)})/\alpha)$, doing a calculation akin to the calculation in Case I (now using $|H| \leq K_0$) we get that the above inequality is true for

$$\beta \leq (\log(1/\kappa))^{-1/2}\cdot\alpha^{O(\log(1/\alpha))}\cdot\log\log(1/\kappa)^{-O(\log\log\log(1/\kappa))}$$

as long as the constant inside the $O(\cdot)$ notation are sufficiently large. (It is instructive to note here that it is Case II which is the "bottleneck" for our overall bound, in the sense that we require a stronger upper bound on β for Case II than for Case I.) It is now clear that the n-dimensional hyperplane $\mathbf{H}' = \{x \in \mathbb{R}^n \mid v_H \cdot x_H = \nu\}$ contains all the points in $S^* = (\phi_H)^{-1}(S')$ and the points in V, and has the claimed number of nonzero coordinates. This proves the Lemma in Case II and concludes the proof of Lemma 14.

C.3 Proof of Theorem 7.

As mentioned in the body of the paper, our proof is essentially a refined version of Theorem 4 of [Gol06] with two main modifications: one is that we generalize Goldberg's arguments to allow g to be a bounded function rather than a Boolean function, and the other is that we get rid of various factors of $\sqrt{n}$ which arise in the [Gol06] argument (and which would be prohibitively "expensive" for us). The key to getting rid of these factors is the following simple lemma:

LEMMA 30. *Let* $S \subseteq \{-1,1\}^n$ *and* $\mathcal{W} : S \to [0,2]$ *such that* $\sum_{x\in S}\mathcal{W}(x) = \delta 2^n$. *Also, let* $v \in \mathbb{R}^n$ *have* $\|v\| = 1$. *Then*

$$\sum_{x\in S}\mathcal{W}(x)\cdot|v\cdot x| = O(\delta\sqrt{\log(1/\delta)})\cdot 2^n.$$

PROOF. For any $x \in S$, let $D(x) \stackrel{\text{def}}{=} \mathcal{W}(x)/(\sum_{x\in S}\mathcal{W}(x))$. Clearly, D defines a probability distribution over S. By definition, $\mathbf{E}_{x\sim D}[|v\cdot x|] = (\sum_{x\in S}\mathcal{W}(x)\cdot|v\cdot x|)/(\sum_{x\in S}\mathcal{W}(x))$. Since $\sum_{x\in S}\mathcal{W}(x) = \delta\cdot 2^n$, to prove the lemma it suffices to show that $\mathbf{E}_{x\sim D}[|v\cdot x|] = O(\sqrt{\log(1/\delta)})$. Recall that for any non-negative random variable Y, we have the identity $\mathbf{E}[Y] = \int_{t\geq 0}\mathbf{Pr}[Y > t]\,dt$. Thus, we have

$$\mathbf{E}_{x\sim D}[|v\cdot x|] = \int_{t\geq 0}\mathbf{Pr}_{x\sim D}[|v\cdot x| > t]\,dt.$$

To bound this quantity, we exploit the fact that the integrand is concentrated. Indeed, by the Hoeffding bound we have that

$$\mathbf{Pr}_{x\sim\{-1,1\}^n}[|v\cdot x| > t] \leq 2e^{-t^2/2}.$$

This implies that the set $A = \{x \in \{-1,1\}^n : |v\cdot x| > t\}$ is of size at most $2e^{-t^2/2}2^n$. Since $\mathcal{W}(x) \leq 2$ for all $x \in S$, we have that $\sum_{x\in A\cap S}\mathcal{W}(x) \leq 4e^{-t^2/2}2^n$. This implies that $\mathbf{Pr}_{x\sim D}[|v\cdot x| > t] \leq (4/\delta)\cdot e^{-t^2/2}$. The following chain of inequalities completes the proof:

$$\begin{aligned}&\mathbf{E}_{x\sim D}\left[|v\cdot x|\right] = \\ &\int_{t=0}^{\sqrt{2\ln(1/\delta)}}\mathbf{Pr}_{x\sim D}[|w\cdot x| > t]\,dt + \int_{t\geq\sqrt{2\ln(1/\delta)}}\mathbf{Pr}_{x\sim D}[|v\cdot x| > t]\,dt \\ &\leq \sqrt{2\ln(1/\delta)} + \int_{t\geq\sqrt{2\ln(1/\delta)}}\mathbf{Pr}_{x\sim D}[|v\cdot x| > t]\,dt \\ &\leq \sqrt{2\ln(1/\delta)} + \int_{t\geq\sqrt{2\ln(1/\delta)}}\frac{4e^{-t^2/2}}{\delta}\,dt \\ &\leq \sqrt{2\ln(1/\delta)} + \int_{t\geq\sqrt{2\ln(1/\delta)}}\frac{4te^{-t^2/2}}{\delta}\,dt = \sqrt{2\ln(1/\delta)} + 4.\end{aligned}$$

□

We are now ready to prove Theorem 7.

PROOF OF THEOREM 7. Let $f : \{-1,1\}^n \to \{-1,1\}$ be an LTF and $g : \{-1,1\}^n \to [-1,1]$ be an arbitrary bounded function. Assuming that $\mathrm{dist}(f,g) = \epsilon$, we will prove that $d_{\mathrm{Chow}}(f,g) \geq \delta = \delta(\epsilon) \stackrel{\text{def}}{=} \epsilon^{\Theta(\log^2(1/\epsilon))}$.

Let us define $V_+ = \{x \in \{-1,1\}^n \mid f(x) = 1, g(x) < 1\}$ and $V_- = \{x \in \{-1,1\}^n \mid f(x) = -1, g(x) > -1\}$. Also, for every point $x \in \{-1,1\}^n$, we associate a weight $\mathcal{W}(x) = |f(x) - g(x)|$ and for a set S, we define $\mathcal{W}(S) \stackrel{\text{def}}{=} \sum_{x \in S} \mathcal{W}(x)$.

It is clear that $V_+ \cup V_-$ is the disagreement region between f and g and that therefore $\mathcal{W}(V_+) + \mathcal{W}(V_-) = \epsilon \cdot 2^n$. We claim that without loss of generality we may assume that $(\epsilon - \delta) \cdot 2^{n-1} \leq \mathcal{W}(V_+), \mathcal{W}(V_-) \leq (\epsilon + \delta) \cdot 2^{n-1}$. Indeed, if this condition is not satisfied, we have that $|\widehat{f}(0) - \widehat{g}(0)| > \delta$ which gives the conclusion of the theorem.

We record the following trivial fact which shall be used several times subsequently.

FACT 31. *For $\mathcal{W}$ as defined above, for all $X \subseteq \{-1,1\}^n$, $|X| \geq \mathcal{W}(X)/2$.*

We start by defining $V_+^0 = V_+$, $V_-^0 = V_-$ and $V^0 = V_+^0 \cup V_-^0$. The following simple proposition will be useful throughout the proof, since it characterizes the Chow distance between f and g (excluding the degree-0 coefficients) as the (normalized) Euclidean distance between two well-defined points in $\mathbb{R}^n$:

PROPOSITION 32. *Let $\mu_+ = \sum_{x \in V_+} \mathcal{W}(x) \cdot x$ and $\mu_- = \sum_{x \in V_-} \mathcal{W}(x) \cdot x$. Then $\sum_{i=1}^n (\widehat{f}(i) - \widehat{g}(i))^2 = 2^{-2n} \cdot \|\mu_+ - \mu_-\|^2$.*

PROOF. For $i \in [n]$ we have that $\widehat{f}(i) = \mathbf{E}[f(x)x_i]$ and hence $\widehat{f}(i) - \widehat{g}(i) = \mathbf{E}[(f(x) - g(x))x_i]$. Hence $2^n(\widehat{f}(i) - \widehat{g}(i)) = \sum_{x \in V_+} \mathcal{W}(x) \cdot x_i - \sum_{x \in V_-} \mathcal{W}(x) \cdot x_i = (\mu_+ - \mu_-) \cdot \mathbf{e}_i$ where $(\mu_+ - \mu_-) \cdot \mathbf{e}_i$ is the inner product of the vector $\mu_+ - \mu_-$ with the unit vector $\mathbf{e}_i$. Since $\mathbf{e}_1, \ldots, \mathbf{e}_n$ form a complete orthonormal basis for $\mathbb{R}^n$, it follows that

$$\|\mu_+ - \mu_-\|^2 = 2^{2n} \sum_{i \in [n]} (\widehat{f}(i) - \widehat{g}(i))^2$$

proving the claim. □

If $\eta \in \mathbb{R}^n$ has $\|\eta\| = 1$ then it is clear that $\|\mu_+ - \mu_-\| \geq (\mu_+ - \mu_-) \cdot \eta$. By Proposition 32, to lower bound the Chow distance $d_{\mathrm{Chow}}(f,g)$, it suffices to establish a lower bound on $(\mu_+ - \mu_-) \cdot \eta$ for a unit vector η of our choice.

Before proceeding with the proof we fix some notation. For any line ℓ in $\mathbb{R}^n$ and point $x \in \mathbb{R}^n$, we let $\ell(x)$ denote the projection of the point x on the line ℓ. For a set $X \subseteq \mathbb{R}^n$ and a line ℓ in $\mathbb{R}^n$, $\ell(X) \stackrel{\text{def}}{=} \{\ell(x) : x \in X\}$. We use $\widehat{\ell}$ to denote the unit vector in the direction of ℓ (its orientation is irrelevant for us).

DEFINITION 33. *For a function $\mathcal{W} : \{-1,1\}^n \to [0,\infty)$, a set $X \subseteq \{-1,1\}^n$ is said to be (ϵ,ν)-*balanced *if $(\epsilon - \nu)2^{n-1} \leq \sum_{x \in X} \mathcal{W}(x) \leq (\epsilon + \nu)2^{n-1}$.*

Whenever we say that a set X is (ϵ,ν)-balanced, the associated function $\mathcal{W}$ is implicitly assumed to be the one defined at the start of the proof of Theorem 7. The following proposition will be very useful during the course of the proof.

PROPOSITION 34. *Let $X_1, X_2 \subseteq \{-1,1\}^n$ be (ϵ,ν)-balanced sets where $\nu \leq \epsilon/8$. Let ℓ be a line in $\mathbb{R}^n$ and $q \in \ell$ be a point on ℓ such that the sets $\ell(X_1)$ and $\ell(X_2)$ lie on opposite sides of q. Suppose that $S \stackrel{\text{def}}{=} \{x \mid x \in X_1 \cup X_2$ and $\|\ell(x) - q\| \geq \beta\}$. If $\sum_{x \in S} \mathcal{W}(x) \geq \gamma 2^n$, then for $\mu_1 = \sum_{x \in X_1} \mathcal{W}(x) \cdot x$ and $\mu_2 = \sum_{x \in X_2} \mathcal{W}(x) \cdot x$, we have*

$$|(\mu_1 - \mu_2) \cdot \widehat{\ell}| \geq (\beta\gamma - \nu\sqrt{2\ln(16/\epsilon)})2^n.$$

In particular, for $\nu\sqrt{2\ln(16/\epsilon)} \leq \beta\gamma/2$, we have $|(\mu_1 - \mu_2) \cdot \widehat{\ell}| \geq (\beta\gamma/2)2^n$.

PROOF. We may assume that the projection $\ell(x)$ of any point $x \in X_1$ on ℓ is of the form $q + \lambda_x \widehat{\ell}$ where $\lambda_x > 0$, and that the projection $\ell(x)$ of any point $x \in X_2$ on ℓ is of the form $q - \lambda_x \widehat{\ell}$ where $\lambda_x > 0$. We can thus write

$$\begin{aligned}(\mu_1 - \mu_2) \cdot \widehat{\ell} &= \sum_{x \in X_1} \mathcal{W}(x)(q \cdot \widehat{\ell} + \lambda_x) - \sum_{x \in X_2} \mathcal{W}(x)(q \cdot \widehat{\ell} - \lambda_x) \\ &= (\mathcal{W}(X_1) - \mathcal{W}(X_2))\, q \cdot \widehat{\ell} + \sum_{x \in X_1 \cup X_2} \mathcal{W}(x) \cdot \lambda_x.\end{aligned}$$

By the triangle inequality we have

$$\left|(\mu_1 - \mu_2) \cdot \widehat{\ell}\right| \geq \sum_{x \in X_1 \cup X_2} \mathcal{W}(x) \cdot \lambda_x - |q \cdot \widehat{\ell}|\,|(\mathcal{W}(X_1) - \mathcal{W}(X_2))|$$

so it suffices to bound each term separately. For the first term we can write

$$\sum_{x \in X_1 \cup X_2} \mathcal{W}(x) \cdot \lambda_x \geq \sum_{x \in S} \mathcal{W}(x) \cdot \lambda_x \geq \beta\gamma 2^n.$$

To bound the second term, we first recall that (by assumption) $|\mathcal{W}(X_1) - \mathcal{W}(X_2)| \leq \nu 2^n$. Also, we claim that $|q \cdot \widehat{\ell}| < \sqrt{2\ln(16/\epsilon)}$. This is because otherwise the function defined by $g(x) = \mathrm{sign}(x \cdot \widehat{\ell} - q \cdot \widehat{\ell})$ will be $\epsilon/8$ close to a constant function on $\{-1,1\}^n$. In particular, at least one of $|X_1|, |X_2|$ must be at most $(\epsilon/8)2^n$. However, by Fact 31, for $i = 1, 2$ we have that $|X_i| \geq \mathcal{W}(X_i)/2 \geq (\epsilon/4 - \nu/4)2^n > (\epsilon/8)2^n$ resulting in a contradiction. Hence it must be the case that $|q \cdot \widehat{\ell}| < \sqrt{2\ln(16/\epsilon)}$. This implies that $|(\mu_1 - \mu_2) \cdot \widehat{\ell}| \geq (\beta\gamma - \nu\sqrt{2\ln(16/\epsilon)})2^n$ and the proposition is proved. □

We consider a separating hyperplane $\mathbf{A}_0$ for f and assume (without loss of generality) that $\mathbf{A}_0$ does not contain any points of the unit hypercube $\{-1,1\}^n$. Let $\mathbf{A}_0 = \{x \in \mathbb{R}^n \mid w \cdot x = \theta\}$, where $\|w\| = 1$, $\theta \in \mathbb{R}$ and $f(x) = \mathrm{sign}(w \cdot x - \theta)$.

Consider a line ℓ_0 normal to $\mathbf{A}_0$, so w is the unit vector defining the direction of ℓ_0 that points to the halfspace $f^{-1}(1)$. As stated before, the exact orientation of ℓ_0 is irrelevant to us and the choice of orientation here is arbitrary. Let $q_0 \in \mathbb{R}^n$ be the intersection point of ℓ_0 and $\mathbf{A}_0$. Then we can write the line ℓ_0 as $\ell_0 = \{p \in \mathbb{R}^n \mid p = q_0 + \lambda w, \lambda \in \mathbb{R}\}$.

Define $\beta \stackrel{\text{def}}{=} \epsilon^{O(\log(1/\epsilon))}$ and consider the set of points

$$S_0 = \{x : x \in V^0 \mid \|\ell_0(x) - q_0\| \geq \beta\}.$$

The following claim states that if $\mathcal{W}(S_0)$ is not very small, we get the desired lower bound on the Chow distance.

CLAIM 35. *Suppose that $\mathcal{W}(S_0) \geq \gamma_0 \cdot 2^n$ where $\gamma_0 \stackrel{\text{def}}{=} \beta^{4\log(1/\epsilon)-2}$. ϵ. Then $d_{\mathrm{Chow}}(f,g) \geq \delta$.*

PROOF. To prove the desired lower bound, we will apply Proposition 32. Consider projecting every point in V^0 on the line ℓ_0. Observe that the projections of V_+^0 are separated from the projections

of V^0_- by the point q_0. Also, we recall that the sets V^0_+ and V^0_- are (ϵ, δ) balanced. Thus, if we define $\mu_+ = \sum_{x \in V^0_+} \mathcal{W}(x) \cdot x$ and $\mu_- = \sum_{x \in V^0_-} \mathcal{W}(x) \cdot x$, we can apply Proposition 34 to get that $|(\mu_+ - \mu_-) \cdot w| \geq (\beta\gamma_0 - \delta\sqrt{2\ln(16/\epsilon)})2^n \geq \delta 2^n$. This implies that $\|\mu_+ - \mu_-\|^2 \geq \delta^2 2^{2n}$ and using Proposition 32, this proves that $d_{\mathrm{Chow}}(f, g) \geq \delta$. $\square$

If the condition of Claim 35 is not satisfied, then we have that $\mathcal{W}(V^0 \setminus S_0) \geq (\epsilon - \gamma_0)2^n$. By Fact 31, we have $|V^0 \setminus S_0| \geq (\epsilon - \gamma_0)2^{n-1}$. We now apply Lemma 14 to obtain another hyperplane $\mathbf{A}_1$ which passes through all but $\kappa_1 \cdot 2^n$ points ($\kappa_1 \stackrel{\text{def}}{=} \gamma_0/2$) in $V^0 \setminus S_0$. We note that the condition of the lemma is satisfied, as $\log(1/\kappa_1) = \mathrm{poly}(\log(1/\epsilon))$ and $|V^0 \setminus S_0| > (\epsilon/4) \cdot 2^n$.

From this point onwards, our proof uses a sequence of $\lfloor \log(1/\epsilon) \rfloor$ cases. To this end, we define $\gamma_j = \beta^{4\log(1/\epsilon) - 2(j+1)} \cdot \epsilon$. At the beginning of case j, we will have an affine space A_j of dimension $n - j$ such that $\mathcal{W}(V^0 \cap A_j) \geq (\epsilon - 2(\sum_{\ell=0}^{j-1} \gamma_\ell))2^n$. We note that this is indeed satisfied at the beginning of case 1. To see this, recall that $\mathcal{W}(V^0 \setminus S_0) > (\epsilon - \gamma_0)2^n$. Also, we have that

$$\begin{aligned} W((V^0 \setminus S_0) \setminus (V^0 \cap \mathbf{A}_1)) &\leq 2|(V^0 \setminus S_0) \setminus (V^0 \cap \mathbf{A}_1)| \\ &\leq 2\kappa_1 2^n = \gamma_0 2^n. \end{aligned}$$

These together imply that $\mathcal{W}(V^0 \cap \mathbf{A}_1) \geq (\epsilon - 2\gamma_0)2^n$ confirming the hypothesis for $j = 1$.

We next define $V^j = V^0 \cap A_j$, $V^j_+ = V^j \cap V_+$ and $V^j_- = V^j \cap V_-$. Similarly, define $\Delta^j_+ = V^0_+ \setminus V^j_+$ and $\Delta^j_- = V^0_- \setminus V^j_-$. Let $A'_{j+1} = A_j \cap \mathbf{A}_0$. Note that $A_j \not\subseteq \mathbf{A}_0$. This is because A_j contains points from $\{-1, 1\}^n$ as opposed to $\mathbf{A}_0$ which does not. Also, A_j is not contained in a hyperplane parallel to $\mathbf{A}_0$ because A_j contains points of the unit hypercube lying on either side of $\mathbf{A}_0$. Hence it must be the case that $\dim(A'_{j+1}) = n - (j+1)$. Let ℓ_j be a line orthogonal to A'_{j+1} which is parallel to A_j. Again, we observe that the direction of ℓ_j is unique.

We next observe that all points in A'_{j+1} project to the same point in ℓ_j, which we call q_j. Let us define $\Lambda^j_+ = \ell_j(V^j_+)$ and $\Lambda^j_- = \ell_j(V^j_-)$. We state the following important observation.

OBSERVATION 36. *The sets Λ^j_+ and Λ^j_- are separated by q_j.*

Next, we define S_j as :

$$S_j = \{x \in V^j \mid \|\ell_j(x) - q_j\|_2 \geq \beta\}.$$

The next claim is analogous to Claim 35. It says that if $\mathcal{W}(S_j)$ is not too small, then we get the desired lower bound on the Chow distance. The proof is slightly more technical and uses Lemma 30.

CLAIM 37. *For $j \leq \log(8/\epsilon)$, suppose that $\mathcal{W}(S_j) \geq \gamma_j \cdot 2^n$ where γ_j is as defined above. Then $d_{\mathrm{Chow}}(f, g) \geq \delta$.*

PROOF. We start by observing that

$$\left(\epsilon - 4\sum_{\ell=0}^{j-1} \gamma_\ell\right) 2^{n-1} \leq \mathcal{W}(V^j_+), \mathcal{W}(V^j_-) \leq (\epsilon + \delta)2^{n-1}.$$

The upper bound is obvious because $V^j_+ \subseteq V^0_+$ and $V^j_- \subseteq V^0_-$ and the range of $\mathcal{W}$ is non-negative. To see the lower bound, note that $\mathcal{W}(V^0 \setminus V^j) \leq 2(\sum_{\ell=0}^{j-1} \gamma_\ell)2^n$. As $V^0_+ \setminus V^j_+$ and $V^0_- \setminus V^j_-$ are both contained in $V^0 \setminus V^j$, we get the stated lower bound. We also note that

$$\begin{aligned} 2\left(\sum_{\ell=0}^{j-1} \gamma_\ell\right) 2^n &= 2\left(\sum_{\ell=0}^{j-1} \beta^{4\log(1/\epsilon) - 2\ell - 2}\right) 2^n \\ &\leq 4\beta^{4\log(1/\epsilon) - 2j} 2^n. \end{aligned}$$

This implies that the sets V^j_+ and V^j_- are $(\epsilon, 4\beta^{4\log(1/\epsilon) - 2j} + \delta)$ balanced. In particular, using that $\delta \leq 4\beta^{4\log(1/\epsilon) - 2j}$, we can say that the sets V^j_+ and V^j_- are $(\epsilon, 8\beta^{4\log(1/\epsilon) - 2j})$-balanced. We also observe that for $j \leq \log(8/\epsilon)$, we have that $8\beta^{4\log(1/\epsilon) - 2j} \leq \epsilon/8$. Let us define $\mu^j_+ = \sum_{x \in V^j_+} \mathcal{W}(x) \cdot x$ and $\mu^j_- = \sum_{x \in V^j_-} \mathcal{W}(x) \cdot x$. An application of Proposition 34 yields that $|(\mu^j_+ - \mu^j_-) \cdot \widehat{\ell_j}| \geq (\beta\gamma_j - 8\beta^{4\log(1/\epsilon) - 2j}\sqrt{2\ln(16/\epsilon)})2^n$.

We now note that

$$(\mu_+ - \mu_-) \cdot \widehat{\ell_j} = (\mu^j_+ - \mu^j_-) \cdot \widehat{\ell_j} + \left(\sum_{x \in \Delta^j_+} \mathcal{W}(x) - \sum_{x \in \Delta^j_-} \mathcal{W}(x)\right) \cdot \widehat{\ell_j}.$$

Defining $\mu'^j_+ = \sum_{x \in \Delta^j_+} \mathcal{W}(x) \cdot x$ and $\mu'^j_- = \sum_{x \in \Delta^j_-} \mathcal{W}(x) \cdot x$, the triangle inequality implies that

$$\left|(\mu_+ - \mu_-) \cdot \widehat{\ell_j}\right| \geq \left|\left(\mu^j_+ - \mu^j_-\right) \cdot \widehat{\ell_j}\right| - \left|\mu'^j_+ \cdot \widehat{\ell_j}\right| - \left|\mu'^j_- \cdot \widehat{\ell_j}\right|.$$

Using Lemma 30 and that $\mathcal{W}(\Delta^j_+), \mathcal{W}(\Delta^j_-) \leq \mathcal{W}(V^0 \setminus V^j) \leq 8\beta^{4\log(1/\epsilon) - 2j} \cdot 2^n$, we get that

$$\begin{aligned} \left|\mu'^j_+ \cdot \widehat{\ell_j}\right| &= \sum_{x \in \Delta^j_+} \mathcal{W}(x) \cdot x \cdot \widehat{\ell_j} \\ &= O\left(|\Delta^j_+| \cdot \sqrt{\log(2^n/|\Delta^j_+|)}\right) \\ &= O\left(\beta^{4\log(1/\epsilon) - 2j} \cdot \log^{3/2}(1/\epsilon) \cdot 2^n\right) \end{aligned}$$

and similarly

$$\begin{aligned} \left|\mu'^j_- \cdot \widehat{\ell_j}\right| &= \sum_{x \in \Delta^j_-} \mathcal{W}(x) \cdot x) \cdot \widehat{\ell_j} \\ &= O\left(|\Delta^j_-| \cdot \sqrt{\log(2^n/|\Delta^j_-|)}\right) \\ &= O\left(\beta^{4\log(1/\epsilon) - 2j} \cdot \log^{3/2}(1/\epsilon) \cdot 2^n\right). \end{aligned}$$

This implies that

$$\begin{aligned} \left|(\mu_+ - \mu_-) \cdot \widehat{\ell_j}\right| \geq (\beta\gamma_j - 8\beta^{4\log(1/\epsilon) - 2j}\sqrt{2\ln(8/\epsilon)})2^n \\ -O\left(\beta^{4\log(1/\epsilon) - 2j} \cdot \log^{3/2}(1/\epsilon) \cdot 2^n\right). \end{aligned}$$

Plugging in the value of γ_j, we see that for ϵ smaller than a sufficiently small constant, we have that

$$\left|(\mu_+ - \mu_-) \cdot \widehat{\ell_j}\right| \geq \beta\gamma_j 2^{n-1}.$$

An application of Proposition 32 finally gives us that

$$d_{\mathrm{Chow}}(f, g) \geq 2^{-n}\|\mu_+ - \mu_-\| \geq 2^{-n}(\mu_+ - \mu_-) \cdot \widehat{\ell_j} = \beta\gamma_j/2 \geq \delta$$

which establishes the Claim. $\square$

If the hypothesis of Claim 37 fails, then we construct an affine space A_{j+1} of dimension $n - j - 1$ such that $\mathcal{W}(V^0 \cap A_{j+1}) \geq (\epsilon - 2\sum_{\ell=0}^{j} \gamma_\ell)2^n$ as described next. We recall that $U = \cup_{i=1}^n \mathbf{e}_i \cup \mathbf{0}$. It is obvious there is some subset $Y_j \subseteq U$ such that $|Y_j| = j$ and $\mathrm{span}(A_j \cup Y_j) = \mathbb{R}^n$. Now, let us define $\mathbf{H}'_j \stackrel{\text{def}}{=} \mathrm{span}(Y_j \cup A'_{j+1})$. Clearly, $\mathbf{H}'_j$ is a hyperplane and every point $x \in (V^0 \cap A_j) \setminus S_j$ is at a distance at most β from H'_j. This is because every $x \in (V^0 \cap A_j) \setminus S_j$ is at a distance at most β from A'_{j+1} and $A'_{j+1} \subset \mathbf{H}'_j$. Also, note that all $x \in Y_j$ lie on $\mathbf{H}'_j$.

Note that $\mathcal{W}((V^0 \cap A_j) \setminus S_j) \geq (\epsilon - 2\sum_{\ell=0}^{j-1}\gamma_\ell - \gamma_j)2^n$. As prior calculation has shown, for $j \leq \log(8/\epsilon)$ we have $\mathcal{W}((V^0 \cap A_j) \setminus S_j) \geq (\epsilon - 2\sum_{\ell=0}^{j-1}\gamma_\ell - \gamma_j)2^n \geq (\epsilon/2)2^n$. Using Fact 31, we get that $|(V^0 \cap A_j) \setminus S_j| \geq (\epsilon/4)2^n$. Thus, putting $\kappa_j = \gamma_j/2$ and applying Lemma 14, we get a new hyperplane $\mathbf{H}_j$ such that $|((V^0 \cap A_j) \setminus S_j) \setminus (\mathbf{H}_j \cap V^0)| \leq (\gamma_j/2) \cdot 2^n$. Using that the range of $\mathcal{W}$ is bounded by 2, we get $\mathcal{W}(((V^0 \cap A_j) \setminus S_j) \setminus (\mathbf{H}_j \cap V^0)) \leq \gamma_j \cdot 2^n$. Thus, we get that $\mathcal{W}(\mathbf{H}_j \cap V^0 \cap A_j) \geq (\epsilon - 2\sum_{\ell=0}^{j}\gamma_\ell)2^n$. Also, $Y_j \subset \mathbf{H}_j$.

Let us now define $A_{j+1} = A_j \cap \mathbf{H}_j$. It is clear that $\mathcal{W}(A_{j+1} \cap V^0) \geq (\epsilon - 2\sum_{\ell=0}^{j}\gamma_\ell)2^n$. Also, $\dim(A_{j+1}) < \dim(A_j)$. To see this, assume for contradiction that $\dim(A_j) = \dim(A_{j+1})$. This means that $A_j \subseteq \mathbf{H}_j$. Also, $Y_j \subset \mathbf{H}_j$. This means that $\mathrm{span}(A_j \cup Y_j) \subset \mathbf{H}_j$. But $\mathrm{span}(A_j \cup Y_j) = \mathbb{R}^n$ which cannot be contained in $\mathbf{H}_j$. Thus we have that $\dim(A_{j+1}) = \dim(A_j) - 1$.

Now we observe that taking $j = \lfloor \log(8/\epsilon) \rfloor$, we have a subspace A_j of dimension $n - j$ which has $\mathcal{W}(A_j \cap V^0) \geq (\epsilon - 2\sum_{\ell=0}^{j-1}\gamma_\ell)2^n > (\epsilon/2)2^n$. By Fact 31, we have that $|A_j \cap V^0| \geq (\epsilon/4)2^n$. However, by Fact 25, a subspace of dimension $n - j$ can contain at most 2^{n-j} points of $\{-1, 1\}^n$. Since $j = \lfloor \log(8/\epsilon) \rfloor$, this leads to a contradiction. That implies that the number of cases must be strictly less than $\lfloor \log(8/\epsilon) \rfloor$. In particular, for some $j < \lfloor \log(8/\epsilon) \rfloor$, it must be the case that $|S_j| \geq \gamma_j 2^n$. For this j, by Claim 37, we get a lower bound of δ on $d_{\mathrm{Chow}}(f, g)$. This concludes the proof of Theorem 7.

D. PROOF OF THEOREM 10

In this section we give a proof of our main algorithmic result (Theorem 10).

We build g through the following iterative process. Let $g'_0 \equiv 0$ and let $g_0 = P_1(g'_0)$. Given g_t, we compute the Chow parameters of g_t to accuracy $\epsilon/(4\sqrt{n+1})$ and let $(\beta_0, \beta_1, \ldots, \beta_n)$ denote the results. For each $0 \leq i \leq n$ we define $\tilde{g}_t(i)$ to be the closest value to β_i that ensures that $\alpha_i - \beta_i$ is an integer multiple of $\epsilon/(2\sqrt{n+1})$. Let $\tilde{\chi}_{g_t} = (\tilde{g}_t(0), \ldots, \tilde{g}_t(n))$ denote the resulting vector of coefficients. Note that

$$\|\tilde{\chi}_{g_t} - \vec{\chi}_{g_t}\| \leq \sqrt{\sum_{i \in [0..n]} (\epsilon/(2\sqrt{n+1}))^2} = \epsilon/2.$$

If $\rho \triangleq \|\vec{\alpha} - \tilde{\chi}_{g_t}\| \leq 4\epsilon$ then we stop and output g_t. By triangle inequality,

$$\begin{aligned} \|\vec{\chi}_f - \vec{\chi}_{g_t}\| &\leq \|\vec{\chi}_f - \vec{\alpha}\| + \|\vec{\alpha} - \tilde{\chi}_{g_t}\| + \|\tilde{\chi}_{g_t} - \vec{\chi}_{g_t}\| \\ &\leq \epsilon(1 + 4 + 1/2) < 6\epsilon, \end{aligned}$$

in other words g_t satisfies the claimed condition.

Otherwise (when $\rho > 4\epsilon$), let $g'_{t+1} = g'_t + h_t/2$ and $g_{t+1} = P_1(g'_{t+1})$ for

$$h_t \triangleq \sum_{i \in [0..n]} (\alpha_i - \tilde{g}_t(i)) x_i.$$

Note that this is equivalent to adding the vector $(\vec{\alpha} - \tilde{\chi}_{g_t})/2$ to the degree 0 and 1 Fourier coefficients of g'_t (which are also the components of the vector representing g_t).

To prove the convergence of this process we define a potential function at step t as

$$\begin{aligned} E(t) &= \mathbf{E}[(f - g_t)^2] + 2\,\mathbf{E}[(f - g_t)(g_t - g'_t)] \\ &= \mathbf{E}[(f - g_t)(f - 2g'_t + g_t)]. \end{aligned}$$

The key claim of this proof is that

$$E(t+1) - E(t) \leq -2\epsilon^2.$$

To prove this claim we first prove that

$$\mathbf{E}[(f - g_t)h_t] \geq \rho(\rho - \frac{3}{2}\epsilon). \tag{1}$$

To prove equation (1) we observe that, by Cauchy-Schwartz inequality,

$$\begin{aligned} \mathbf{E}[(f - g_t)h_t] &= \sum_{i \in [0..n]} (\hat{f}(i) - \hat{g}_t(i))(\alpha_i - \tilde{g}_t(i)) \\ &= \sum_{i \in [0..n]} \Big[(\hat{f}(i) - \alpha_i)(\alpha_i - \tilde{g}_t(i)) + \\ &\quad (\tilde{g}_t(i) - \hat{g}_t(i))(\alpha_i - \tilde{g}_t(i)) + (\alpha_i - \tilde{g}_t(i))^2 \Big] \\ &\geq -\rho\epsilon - \rho\epsilon/2 + \rho^2 \geq \rho^2 - \frac{3}{2}\rho\epsilon. \end{aligned}$$

In addition, by Parseval's identity,

$$\mathbf{E}[h_t^2] = \sum_{i \in [0..n]} (\alpha_i - \tilde{g}_t(i))^2 = \rho^2\,. \tag{2}$$

Now,

$$E(t+1) - E(t) =$$

$$\begin{aligned} &= \mathbf{E}[(f - g_{t+1})(f - 2g'_{t+1} + g_{t+1})] - \mathbf{E}[(f - g_t)(f - 2g'_t + g_t)] \\ &= \mathbf{E}\left[(f - g_t)(2g'_t - 2g'_{t+1}) + (g_{t+1} - g_t)(2g'_{t+1} - g_t - g_{t+1})\right] \\ &= -\mathbf{E}[(f - g_t)h_t] + \mathbf{E}\left[(g_{t+1} - g_t)(2g'_{t+1} - g_t - g_{t+1})\right] \end{aligned} \tag{3}$$

To upper-bound the expression $\mathbf{E}\left[(g_{t+1} - g_t)(2g'_{t+1} - g_t - g_{t+1})\right]$ we prove that for every point $x \in \{-1, 1\}^n$,

$$(g_{t+1}(x) - g_t(x))(2g'_{t+1}(x) - g_t(x) - g_{t+1}(x)) \leq h_t(x)^2/2.$$

We first observe that

$$|g_{t+1}(x) - g_t(x)| = |P_1(g'_t(x) + h_t(x)/2) - P_1(g'_t(x))| \leq |h_t(x)/2|$$

(a projection operation does not increase the distance). Now

$$\begin{aligned} &|2g'_{t+1}(x) - g_t(x) - g_{t+1}(x)| \leq \\ &|g'_{t+1}(x) - g_t(x)| + |(g'_{t+1}(x) - g_{t+1}(x)|. \end{aligned}$$

The first part $|g'_{t+1}(x) - g_t(x)| = |h_t(x)/2 + g'_t(x) - g_t(x)| \leq |h_t(x)/2|$ unless $g'_t(x) - g_t(x) \neq 0$ and $g'_t(x) - g_t(x)$ has the same sign as $h_t(x)$. However, in this case $g_{t+1}(x) = g_t(x)$ and as a result $(g_{t+1}(x) - g_t(x))(2g'_{t+1}(x) - g_t(x) - g_{t+1}(x)) = 0$. Similarly, $|g'_{t+1}(x) - g_{t+1}(x)| \leq |h_t(x)/2|$ unless $g_{t+1}(x) = g_t(x)$. Altogether we obtain that

$$\begin{aligned} &(g_{t+1}(x) - g_t(x))(2g'_{t+1}(x) - g_t(x) - g_{t+1}(x)) \leq \\ &\max\{0, |h_t(x)/2|(|h_t(x)/2| + |h_t(x)/2|)\} = h_t(x)^2/2. \end{aligned}$$

This implies that

$$\mathbf{E}\left[(g_{t+1} - g_t)(2g'_{t+1} - g_t - g_{t+1})\right] \leq \mathbf{E}[h_t^2]/2 = \rho^2/2. \tag{4}$$

By substituting equations (1) and (4) into equation (3), we obtain the claimed decrease in the potential function

$$E(t+1) - E(t) \leq -\rho^2 + \frac{3}{2}\rho\epsilon + \rho^2/2 = -(\rho - 3\epsilon)\rho/2 \leq -2\epsilon^2.$$

We now observe that

$$E(t) = \mathbf{E}[(f - g_t)^2] + 2\,\mathbf{E}[(f - g_t)(g_t - g'_t)] \geq 0$$

for all t. This follows from noting that for every x and $f(x) \in \{-1,1\}$, either $f(x) - P_1(g'_t(x))$ and $P_1(g'_t(x)) - g'_t(x)$ have the same sign or one of them equals zero. Therefore

$$\mathbf{E}[(f - g_t)(g_t - g'_t)] \geq 0$$

(and, naturally, $\mathbf{E}[(f-g_t)^2] \geq 0$). It is easy to see that $E(0) = 1$ and therefore this process will stop after at most $1/(2\epsilon^2)$ steps.

We now establish the claimed weight bound on the LBF output by the algorithm and the bound on the running time. Let T denote the number of iterations of the algorithm. By our construction, the function $g_T = P_1(\sum_{t \leq T} h_t/2)$ is an LBF represented by weight vector $\vec{w}$ such that $w_i = \sum_{j \leq T}(\alpha_i - \tilde{g}_j(i))/2$. Our rounding of the estimates of Chow parameters of g_t ensures that each of $(\alpha_i - \tilde{g}_j(i))/2$ is a multiple of $\kappa = \epsilon/(4\sqrt{n+1})$. Hence g_T can be represented by vector $\vec{w} = \kappa\vec{v}$, where vector $\vec{v}$ has only integer components. At every step j,

$$\sqrt{\sum_{i \in [0..n]} (\alpha_i - \tilde{g}_j(i))^2} \leq 2 + \epsilon + \epsilon/2 = O(1).$$

Therefore, by triangle inequality, $\|\vec{w}\| = O(\epsilon^{-2})$ and hence $\|\vec{v}\| = \|\vec{w}\|/\kappa = O(\sqrt{n}/\epsilon^3)$.

The running time of the algorithm is essentially determined by finding $\tilde{\chi}_{g_t}$ in each step t. Finding $\tilde{\chi}_{g_t}$ requires estimating each $\hat{g}_t(i) = \mathbf{E}[g_t(x) \cdot x_i]$ to accuracy $\epsilon/(4\sqrt{n+1})$. Chernoff bounds imply that, by using the empirical mean of $g_t(x) \cdot x_i$ on $O((n/\epsilon^2) \cdot \log(n/(\epsilon\delta))$ random points as our estimate of $\hat{g}_t(i)$ we can ensure that, with probability at least $1-\delta$, the estimates are within $\epsilon/(4\sqrt{n+1})$ of the true values for all $n+1$ Chow parameters of g_t for every $t \leq T = O(\epsilon^{-2})$.

Evaluating g_t on any point $x \in \{-1,1\}^n$ takes $O(n)$ time and we need to evaluate it on $O((n/\epsilon^2) \cdot \log(n/(\epsilon\delta))$ points in each of $O(\epsilon^{-2})$ steps. This gives us the claimed total running time bound of $\tilde{O}(n^2\epsilon^{-4}\log(1/\delta))$.

D.1 Faster algorithms for small-weight LTFs

In this section, we restate Theorem 16 followed by its proof.

THEOREM 16. *Let $f = \mathrm{sign}(\sum_{i=1}^n w_i x_i - \theta)$ be an LTF with integer weights w_i such that $W \stackrel{def}{=} \sum_{i=1}^n |w_i| = \mathrm{poly}(n)$. Fix $0 < \epsilon, \delta < 1/2$. Write $\vec{\chi}_f$ for the Chow vector of f and assume that $\vec{\alpha} \in \mathbb{R}^{n+1}$ is a vector satisfying $\|\vec{\alpha} - \vec{\chi}_f\| \leq \epsilon/(12W)$. Then, there is an algorithm $\mathcal{A}'$ with the following property: Given as input $\vec{\alpha}$, ϵ and δ, $\mathcal{A}'$ performs $\mathrm{poly}(n/\epsilon) \cdot \log(1/\delta)$ bit operations and outputs the (weights-based) representation of an LTF f^* which with probability at least $1-\delta$ satisfies $\mathrm{dist}(f, f^*) \leq \epsilon$.*

PROOF. As stated before, both the algorithm and proof of the above theorem are identical to the ones in Theorem 15. The details follow.

Given a vector $\vec{\alpha} \in \mathbb{R}^{n+1}$ satisfying $\Delta := \|\vec{\alpha} - \vec{\chi}_f\| \leq \epsilon/(12W)$, where f is the unknown LTF, we run algorithm `ChowReconstruct` on input $\vec{\alpha}$. The algorithm runs in time $\mathrm{poly}(1/\Delta) \cdot \tilde{O}(n^2) \cdot \log(1/\delta)$, which is $\mathrm{poly}(n/\epsilon) \cdot \log(1/\delta)$ by our assumption on W, and outputs an LBF g such that with probability at least $1-\delta$, $d_{\mathrm{Chow}}(f,g) \leq 6\Delta \leq \epsilon/(2W)$. At this point, we need to apply the following simple structural result of [BDJ+98]:

FACT 38. *Let $f = \mathrm{sign}(\sum_{i=1}^n w_i x_i - \theta)$ be an LTF with integer weights w_i, where $W \stackrel{def}{=} \sum_{i=1}^n |w_i|$, and $g : \{-1,1\}^n \to [-1,1]$ be an arbitrary bounded function. Fix $0 < \epsilon < 1/2$. If $d_{\mathrm{Chow}}(f,g) \leq \epsilon/W$, then $\mathrm{dist}(f,g) \leq \epsilon$.*

The above fact implies that, with probability at least $1-\delta$, the LBF g output by the algorithm satisfies $\mathrm{dist}(f,g) \leq \epsilon/2$. If $g(x) = P_1(v_0 + \sum_{i=1}^n v_i x_i)$, we similarly have that the LTF $f^*(x) = \mathrm{sign}(v_0 + \sum_{i=1}^n v_i x_i)$ has $\mathrm{dist}(f, f^*) \leq \epsilon$. This completes the proof. □

E. APPLICATIONS TO COMPUTATIONAL LEARNING THEORY.

In this section we show that our approach yields a range of interesting algorithmic applications in learning theory.

E.0.1 Learning threshold functions in the 1-RFA model.

Ben-David and Dichterman [BDD98] introduced the "Restricted Focus of Attention" (RFA) learning framework to model the phenomenon (common in the real world) of a learner having incomplete access to examples. We focus here on the uniform-distribution "1-RFA" model. In this setting each time the learner is to receive a labeled example, it first specifies an index $i \in [n]$; then an n-bit string x is drawn from the uniform distribution over $\{-1,1\}^n$ and the learner is given $(x_i, f(x))$. So for each labeled example, the learner is only shown the i-th bit of the example along with the label.

Birkendorf et al. [BDJ+98] asked whether LTFs can be learned in the uniform distribution 1-RFA model, and showed that a sample of $O(n \cdot W^2 \cdot \log(\frac{n}{\delta})/\epsilon^2)$ many examples is information-theoretically sufficient for learning an unknown threshold function with integer weights w_i that satisfy $\sum_i |w_i| \leq W$. The results of Goldberg [Gol06] and Servedio [Ser07] show that samples of size $(n/\epsilon)^{O(\log(n/\epsilon)\log(1/\epsilon))}$ and $\mathrm{poly}(n) \cdot 2^{\tilde{O}(1/\epsilon^2)}$ respectively are information-theoretically sufficient for learning an arbitrary LTF to accuracy ϵ, but none of these earlier results gave a computationally efficient algorithm. [OS11] gave the first algorithm for this problem; as a consequence of their result for the Chow Parameters Problem, they gave an algorithm which learns LTFs to accuracy ϵ and confidence $1-\delta$ in the uniform distribution 1-RFA model, running in $2^{2^{\tilde{O}(1/\epsilon^2)}} \cdot n^2 \cdot \log n \cdot \log(\frac{n}{\delta})$ bit operations. As a direct consequence of Theorem 1, we obtain a much more time efficient learning algorithm for this learning task. We now restate Theorem 17 here for convenience.

THEOREM 17. *There is an algorithm which performs $\tilde{O}(n^2) \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))} \cdot \log(\frac{1}{\delta})$ bit-operations and properly learns LTFs to accuracy ϵ and confidence $1-\delta$ in the uniform distribution 1-RFA model.*

E.0.2 Agnostic-type learning.

In this section we show that a variant of our main algorithm gives a very fast "agnostic-type" algorithm for learning LTFs under the uniform distribution.

Let us briefly review the uniform distribution agnostic learning model [KSS94] in our context. Let $f : \{-1,1\}^n \to \{-1,1\}$ be an arbitrary boolean function. We write $\mathsf{opt} = \mathrm{dist}(f, \mathcal{H}) \stackrel{\text{def}}{=} \min_{h \in \mathcal{H}} \mathbf{Pr}_x[h(x) \neq f(x)]$, where $\mathcal{H}$ denotes the class of LTFs. A uniform distribution agnostic learning algorithm is given uniform random examples labeled according to an arbitrary f and outputs a hypothesis h satisfying $\mathrm{dist}(h,f) \leq \mathsf{opt} + \epsilon$.

The only efficient algorithm for learning LTFs in this model [KKMS05] is non-proper and runs in time $n^{O(1/\epsilon^4)}$. This motivates the design of more efficient algorithms with potentially relaxed guarantees. [OS11] give an "agnostic-type" algorithm, that guarantees $\mathrm{dist}(h,f) \leq \mathsf{opt}^{\Omega(1)} + \epsilon$ and runs in time $\mathrm{poly}(n) \cdot 2^{\mathrm{poly}(1/\epsilon)}$. In

contrast, we give an algorithm that is significantly more efficient, but has a relaxed error guarantee. Theorem 18 from Section 4.3 is restated here followed by its proof.

THEOREM 18. *There is an algorithm $\mathcal{B}$ with the following performance guarantee: Let f be any Boolean function and let $\mathsf{opt} = \mathrm{dist}(f, \mathcal{H})$. Given $0 < \epsilon, \delta < 1/2$ and access to independent uniform examples $(x, f(x))$, algorithm $\mathcal{B}$ outputs the (weights-based) representation of an LTF f^* which with probability $1 - \delta$ satisfies $\mathrm{dist}(f^*, f) \leq 2^{-\Omega(\sqrt[3]{\log(1/\mathsf{opt})})} + \epsilon$. The algorithm performs $\tilde{O}(n^2) \cdot (1/\epsilon)^{O(\log^2(1/\epsilon))} \cdot \log(1/\delta)$ bit operations.*

PROOF. We describe the algorithm $\mathcal{B}$ in tandem with a proof of correctness. We start by estimating each Chow parameter of f (using the random labeled examples) to accuracy $O(\kappa(\epsilon)/\sqrt{n})$; we thus compute a vector $\vec{\alpha} \in \mathbb{R}^{n+1}$ that satisfies $\Delta := \|\vec{\alpha} - \vec{\chi}_f\| \leq \kappa(\epsilon)$. We then run algorithm `ChowReconstruct` (from Theorem 10) on input $\vec{\alpha}$. The algorithm runs in time $\mathrm{poly}(1/\Delta) \cdot \tilde{O}(n^2) \cdot \log(1/\delta)$ and outputs an LBF g such that with probability at least $1 - \delta$ we have $d_{\mathrm{Chow}}(f, g) \leq 6\Delta \leq 6\kappa(\epsilon)$. By assumption, there exists an LTF h^* such that $\mathrm{dist}(h^*, f) \leq \mathsf{opt}$. By Fact 6 we get $d_{\mathrm{Chow}}(h^*, f) \leq 2\sqrt{\mathsf{opt}}$. An application of the triangle inequality now gives $d_{\mathrm{Chow}}(g, h^*) \leq 2\sqrt{\mathsf{opt}} + 4\kappa(\epsilon)$. By Theorem 7, we thus obtain $\mathrm{dist}(g, h^*) \leq 2^{-\Omega(\sqrt[3]{\log(1/\mathsf{opt})})} + \epsilon/2$. Writing the LBF g as $g(x) = P_1(v_0 + \sum_{i=1}^n v_i x_i)$, we similarly have that $f^*(x) = \mathrm{sign}(v_0 + \sum_{i=1}^n v_i x_i)$ has $\mathrm{dist}(f, f^*) \leq 2^{-\Omega(\sqrt[3]{\log(1/\mathsf{opt})})} + \epsilon$. It is easy to see that the running time is dominated by the second step and the proof of Theorem 18 is complete. □

F. NEAR-OPTIMALITY OF LEMMA 14

The following lemma shows that in any statement like Lemma 14 in which the hyperplane $\mathbf{H}'$ passes through *all* the points in S, the distance bound on β can be no larger than $n^{-1/2}$ as a function of n. This implies that the result obtained by taking $\kappa = 1/2^{n+1}$ in Lemma 14, which gives a distance bound of $n^{-(1/2+o(1))}$ as a function of n, is optimal up to the $o(1)$ in the exponent.

LEMMA 39. *Fix $\epsilon > 8n^{-1/2}$. There is a hyperplane $\mathbf{H} \in \mathbb{R}^n$ and a set $S \subseteq \{-1, 1\}^n$ such that $|S| \geq \frac{\epsilon}{8} 2^n$ and the following properties both hold:*

- *For every $x \in S$ we have $d(x, \mathbf{H}) \leq 2\epsilon n^{-1/2}$; and*
- *There is no hyperplane $\mathbf{H}'$ which passes through all the points in S.*

PROOF. Without loss of generality, let us assume $K = 4/\epsilon^2$ is an even integer; note that by assumption $K < n/2$. Now let us define the hyperplane $\mathbf{H}$ by

$$\mathbf{H} = \left\{ x \in \mathbb{R}^n : (x_1 + \ldots + x_K) + \frac{2(x_{K+1} + \ldots + x_n)}{(n-K)} = 0 \right\}$$

Let us define $S = \{x \in \{-1, 1\}^n : d(x, \mathbf{H}) \leq 4/\sqrt{K(n-K)}\}$. It is easy to verify that every $x \in S$ indeed satisfies $d(x, \mathbf{H}) \leq 2\epsilon n^{-1/2}$ as claimed. Next, let us define A as follows:

$$A = \{x \in \{-1, 1\}^n : x_1 + \ldots + x_K = 0$$

and

$$|x_{K+1} + \ldots + x_n| \leq 2\sqrt{n-K}\}.$$

It is easy to observe that $A \subseteq S$. Also, we have

$$\mathbf{Pr}_{x_1, \ldots, x_K}[x_1 + \ldots + x_K = 0] \geq (2\sqrt{K})^{-1}$$

and

$$\mathbf{Pr}_{x_{K+1}, \ldots, x_n}[|x_{K+1} + \ldots + x_n| \leq 2\sqrt{n-K}] \geq 1/2.$$

Hence we have that $|S| \geq \epsilon 2^n/8$. We also observe that the point $z \in \{-1, 1\}^n$ defined as

$$z := (1, 1, \underbrace{1, -1, \ldots, 1, -1}_{K-2}, -1, \ldots, -1) \tag{5}$$

(whose first two coordinates are 1, next $K-2$ coordinates alternate between 1 and -1, and final $n - K$ coordinates are -1) lies on $\mathbf{H}$ and hence $z \in S$.

We next claim that the dimension of the affine span of the points in $A \cup z$ is n. This obviously implies that there is no hyperplane which passes through all points in $A \cup z$, and hence no hyperplane which passes through all points in S. Thus to prove the lemma it remains only to prove the following claim:

CLAIM 40. *The dimension of the affine span of the elements of $A \cup z$ is n.*

To prove the claim, we observe that if we let Y denote the affine span of elements in $A \cup z$ and Y' denote the linear space underlying Y, then it suffices to show that the dimension of Y' is n. Each element of Y' is obtained as the difference of two elements in Y.

First, let $y \in \{-1, 1\}^n$ be such that

$$\sum_{i \leq K} y_i = \sum_{K+1 \leq i \leq n} y_i = 0.$$

Let $y^{\oplus i} \in \{-1, 1\}^n$ be obtained from y by flipping the i-th bit. For each $i \in \{K+1, \ldots, n\}$ we have that y and $y^{\oplus i}$ are both in A, so subtracting the two elements, we get that the basis vector e_i belongs to Y' for each $i \in \{K+1, \ldots, n\}$.

Next, let $i \neq j \leq K$ be positions such that $y_i = 1$ and $y_j = -1$. Let y^{ij} denote the vector which is the same as y except that the signs are flipped at coordinates i and j. Since y^{ij} belongs to A, by subtracting y from y^{ij} we get that for every vector e_{ij} ($i \neq j \leq K$) which has 1 in coordinate i, -1 in coordinate j, and 0 elsewhere, the vector e_{ij} belongs to Y'.

The previous two paragraphs are easily seen to imply that the linear space Y' contains all vectors $x \in \mathbb{R}^n$ that satisfy the condition $x_1 + \cdots + x_K = 0$. Thus to show that the dimension of Y' is n, it suffices to exhibit any vector in Y' that does not satisfy this condition. But it is easy to see that the vector $y - z$ (where z is defined in (5)) is such a vector. This concludes the proof of the claim and of Lemma 39. □

G. USEFUL VARIANTS OF GOLDBERG'S THEOREMS

For technical reasons we require an extension of Theorem 13 (Theorem 3 of [Gol06]) which roughly speaking is as follows: the hypothesis is that not only does the set $S \subset \{-1, 1\}^n$ lie close to hyperplane $\mathbf{H}$ but so also does a (small) set R of points in $\{0, 1\}^n$; and the conclusion is that not only does "almost all" of S (the subset S^*) lie on $\mathbf{H}'$ *but so also does all of R*. To obtain this extension we need a corresponding extension of an earlier result of Goldberg (Theorem 2 of [Gol06]), which he uses to prove his Theorem 3; similar to our extension of Theorem 13 our extension of Theorem 2 of [Gol06] deals with points from both $\{-1, 1\}^n$ and $\{0, 1\}^n$. The simplest approach we have found to obtain our desired extension of Theorem 2 of [Gol06] uses the "Zeroth Inverse Theorem" of Tao and Vu [TV09]. We begin with a useful definition from their paper:

DEFINITION 41. *Given a vector $w = (w_1, \ldots, w_k)$ of real values, the* cube *$S(w)$ is the subset of $\mathbb{R}$ defined as* [2]

$$S(w) = \left\{ \sum_{i=1}^{k} \epsilon_i w_i : (\epsilon_1, \ldots, \epsilon_n) \in \{-1, 0, 1\}^n \right\}.$$

The "Zeroth Inverse Theorem" of [TV09] is as follows:

THEOREM 42. *Suppose $w \in \mathbb{R}^n$, $d \in \mathbb{N}$ and $\theta \in \mathbb{R}$ satisfy $\mathbf{Pr}_{x \in \{-1,1\}^n}[w \cdot x = \theta] > 2^{-d-1}$. Then there exists a d-element subset $A = \{i_1, \ldots, i_d\} \subset [n]$ such that for $v = (w_{i_1}, \ldots, w_{i_d})$ we have $\{w_1, \ldots, w_n\} \subseteq S(v)$.*

For convenience of the reader, we include the proof here.

PROOF OF THEOREM 42. Towards a contradiction, assume that there is no $v = (w_{i_1}, \ldots, w_{i_d})$ such that $\{w_1, \ldots, w_n\} \subseteq S(v)$. Then an obvious greedy argument shows that there are distinct integers $i_1, \ldots, i_{d+1} \in [n]$ such that $w_{i_1}, \ldots, w_{i_{d+1}}$ is *dissociated*, i.e. there does not exist $j \in [n]$ and $\epsilon_i \in \{-1, 0, 1\}$ such that $w_j = \sum_{i \neq j} \epsilon_i w_i$.

Let $v = (w_{i_1}, \ldots, w_{i_{d+1}})$. By an averaging argument, it is easy to see that if $\mathbf{Pr}_{x \in \{-1,1\}^n}[w \cdot x = \theta] > 2^{-d-1}$, then $\exists \nu \in \mathbb{R}$ such that $\mathbf{Pr}_{x \in \{-1,1\}^{d+1}}[v \cdot x = \nu] > 2^{-d-1}$. By the pigeon hole principle, this means that there exist $x, y \in \{-1, 1\}^{d+1}$ such that $x \neq y$ and $v \cdot ((x - y)/2) = 0$. Since entries of $(x - y)/2$ are in $\{-1, 0, 1\}$, and not all the entries in $(x - y)/2$ are zero, this means that v is not dissociated resulting in a contradiction. □

Armed with this result, we now prove the extension of Goldberg's Theorem 2 that we will need later:

THEOREM 43. *Let $w \in \mathbb{R}^n$ have $\|w\|_2 = 1$ and let $\theta \in \mathbb{R}$ be such that $\mathbf{Pr}_{x \in \{-1,1\}^n}[w \cdot x = \theta] = \alpha$. Let $\mathbf{H}$ denote the hyperplane $\mathbf{H} = \{x \in \mathbb{R}^n \mid w \cdot x = \theta\}$. Suppose that $\mathrm{span}(\mathbf{H} \cap (\{-1, 1\}^n \cup \{0, 1\}^n)) = \mathbf{H}$, i.e. the affine span of the points in $\{-1, 1\}^n \cup \{0, 1\}^n$ that lie on $\mathbf{H}$ is $\mathbf{H}$. Then all entries of w are integer multiples of $f(n, \alpha)^{-1}$, where*

$$f(n, \alpha) \leq (2n)^{\lfloor \log(1/\alpha) \rfloor + 3/2} \cdot (\lfloor \log(1/\alpha) \rfloor)!$$

PROOF. We first observe that $w \cdot (x - y) = 0$ for any two points x, y that both lie on $\mathbf{H}$. Consider the system of homogeneous linear equations in variables $w'_1, \ldots, w'_n$ defined by

$$w' \cdot (x - y) = 0 \qquad \text{for all } x, y \in \mathbf{H} \cap (\{-1, 1\}^n \cup \{0, 1\}^n). \quad (6)$$

Since $\mathrm{span}(\mathbf{H} \cap (\{-1, 1\}^n \cup \{0, 1\}^n))$ is by assumption the entire hyperplane $\mathbf{H}$, the system (6) must have rank $n - 1$; in other words, every solution w' that satisfies (6) must be some rescaling $w' = cw$ of the vector w defining $\mathbf{H}$.

Let A denote a subset of $n - 1$ of the equations comprising (6) which has rank $n - 1$ (so any solution to A must be a vector $w' = cw$ as described above). We note that each coefficient in each equation of A lies in $\{-2, -1, 0, 1, 2\}$. Let us define $d = \lfloor \log(1/\alpha) \rfloor + 1$. By Theorem 42, there is some $w_{i_1}, \ldots, w_{i_{d'}}$ with $d' \leq d$ such that for $v \stackrel{\text{def}}{=} (w_{i_1}, \ldots, w_{i_{d'}})$, we have $\{w_1, \ldots, w_n\} \subseteq S(v)$; in other words, for all $j \in [n]$ we have $w_j = \sum_{\ell=1}^{d'} \epsilon_{\ell,j} w_{i_\ell}$ where each $\epsilon_{\ell,j}$ belongs to $\{-1, 0, 1\}$. Substituting these relations into the system A, we get a new system of homogenous linear equations, of rank $d' - 1$, in the variables $w'_{i_1}, \ldots, w'_{i_{d'}}$, where all coefficients of all variables in all equations of the system are integers of magnitude at most $2n$.

Let M denote a subset of $d' - 1$ equations from this new system which has rank $d' - 1$. In other words, viewing M as a $d' \times (d' - 1)$ matrix, we have the equation $M \cdot v^T = 0$ where all entries in the matrix M are integers in $[-2n, 2n]$. Note that at least one of the values $w_{i_1}, \ldots, w_{i_{d'}}$ is non-zero (for if all of them were 0, then since $\{w_1, \ldots, w_n\} \subseteq S(v)$ it would have to be the case that $w_1 = \cdots = w_n = 0$.). Without loss of generality we may suppose that w_{i_1} has the largest magnitude among $w_{i_1}, \ldots, w_{i_{d'}}$. We now fix the scaling constant c, where $w' = cw$, to be such that $w'_{i_1} = 1$. Rearranging the system $M(cv)^T = M(1, w'_{i_2}, \ldots, w'_{i_{d'}})^T = 0$, we get a new system of $d' - 1$ linear equations $M'(w'_{i_2}, \ldots, w'_{i_{d'}})^T = b$ where M' is a $(d' - 1) \times (d' - 1)$ matrix whose entries are integers in $[-2n, 2n]$ and b is a vector whose entries are integers in $[-2n, 2n]$.

We now use Cramer's rule to solve the system

$$M'(w'_{i_2}, \ldots, w'_{i_{d'}})^T = b.$$

This gives us that $w'_{i_j} = \det(M'_j)/\det(M')$ where M'_j is the matrix obtained by replacing the j^{th} column of M' by b. So each w'_{i_j} is an integer multiple of $1/\det(M')$ and is bounded by 1 (by our earlier assumption about w_{i_1} having the largest magnitude). Since $\{w'_1, \ldots, w'_n\} \subseteq S(v)$, we get that each value w'_i is an integer multiple of $1/\det(M')$, and each $|w'_i| \leq n$. Finally, since M' is a $(d' - 1) \times (d' - 1)$ matrix where every entry is an integer of magnitude at most $2n$, we have that $|\det(M')| \leq (2n)^{d'-1} \cdot (d' - 1)! \leq (2n)^{d-1} \cdot (d - 1)!$. Moreover, the ℓ_2 norm of the vector w' is bounded by $n^{3/2}$. So renormalizing (dividing by c) to obtain the unit vector w back from $w' = cw$, we see that every entry of w is an integer multiple of $1/N$, where N is a quantity at most $(2n)^{d+1/2} \cdot d!$. Recalling that $d = \lfloor \log(1/\alpha) \rfloor + 1$, the theorem is proved. □

We next prove the extension of Theorem 3 from [Gol06] that we require. The proof is almost identical to the proof in [Gol06] except for the use of Theorem 43 instead of Theorem 2 from [Gol06] and a few other syntactic changes. For the sake of clarity and completeness, we give the complete proof here.

THEOREM 44. *Given any hyperplane $\mathbf{H}$ in $\mathbb{R}^n$ whose β-neighborhood contains a subset S of vertices of $\{-1, 1\}^n$ where $S = \alpha \cdot 2^n$, there exists a hyperplane which passes through all the points of $(\{-1, 1\}^n \cup \{0, 1\}^n)$ that are contained in the β-neighborhood of $\mathbf{H}$ provided that*

$$0 \leq \beta \leq \left((2/\alpha) \cdot n^{5 + \lfloor \log(n/\alpha) \rfloor} \cdot (2 + \lfloor \log(n/\alpha) \rfloor)! \right)^{-1}.$$

Before giving the proof, we note that the hypothesis of our theorem is the same as the hypothesis of Theorem 3 of [Gol06]. The only difference in the conclusion is that while Goldberg proves that all points of $\{-1, 1\}^n$ in the β-neighborhood of $\mathbf{H}$ lie on the new hyperplane, we prove this for all the points of $(\{-1, 1\}^n \cup \{0, 1\}^n)$ in the β-neighborhood of $\mathbf{H}$.

PROOF. Let $\mathbf{H} = \{x \mid w \cdot x - t = 0\}$ with $\|w\| = 1$. Also, let $S = \{x \in \{-1, 1\}^n \mid d(x, \mathbf{H}) \leq \beta\}$ and $S' = \{x \in (\{-1, 1\}^n \cup \{0, 1\}^n) \mid d(x, \mathbf{H}) \leq \beta\}$. For any $x \in S'$ we have that $w \cdot x \in [t - \beta, t + \beta]$. Following [Gol06] we create a new weight vector $w' \in \mathbb{R}^n$ by rounding each coordinate w_i of w to the nearest integer multiple of β (rounding up in case of a tie). Since every $x \in S'$ has entries from $\{-1, 0, 1\}$, we can deduce that for any $x \in S'$, we

[2] In [TV09] the cube is defined only allowing $\epsilon_i \in \{-1, 1\}$ but this is a typographical error; their proof uses the $\epsilon_i \in \{-1, 0, 1\}$ version that we state.

have

$$t-\beta-n\beta/2 < w\cdot x-n\beta/2 < w'\cdot x < w\cdot x+n\beta/2 \leq t+\beta+n\beta/2.$$

Thus for every $x \in S'$, the value $w' \cdot x$ lies in a semi-open interval of length $\beta(n+2)$; moreover, since it only takes values which are integer multiples of β, there are at most $n+2$ possible values that $w' \cdot x$ can take for $x \in S'$. Since $S \subset S'$ and $|S| \geq \alpha 2^n$, there must be at least one value $t' \in (t-n\beta/2-\beta, t+n\beta/2+\beta]$ such that at least $\alpha 2^n/(n+2)$ points in S lie on the hyperplane $\mathbf{H}_1$ defined as $\mathbf{H}_1 = \{x : w' \cdot x = t'\}$. We also let $A_1 = \mathrm{span}\{x \in S' : w' \cdot x = t'\}$. It is clear that $A_1 \subset \mathbf{H}_1$. Also, since at least $\alpha 2^n/(n+2)$ points of $\{-1,1\}^n$ lie on A_1, by Fact 25 we get that $\dim(A_1) \geq n - \log(n+2) - \log(1/\alpha)$.

It is easy to see that $\|w' - w\| \leq \sqrt{n}\beta/2$, which implies that $\|w'\| \geq 1 - \sqrt{n}\beta/2$. Note that for any $x \in S'$ we have $|w' \cdot x - t'| \leq (n+2)\beta$. Recalling Fact 12, we get that for any $x \in S'$ we have $d(x, \mathbf{H}_1) \leq (\beta(n+2))/(1-\sqrt{n}\beta/2)$. Since $\sqrt{n}\beta \ll 1$, we get that $d(x, \mathbf{H}_1) \leq 2n\beta$ for every $x \in S'$.

At this point our plan for the rest of the proof of Theorem 44 is as follows: First we will construct a hyperplane $\mathbf{H}_k$ (by an inductive construction) such that $\mathrm{span}(\mathbf{H}_k \cap (\{-1,1\}^n \cup \{0,1\}^n)) = \mathbf{H}_k$, $A_1 \subseteq \mathbf{H}_k$, and all points in S' are very close to $\mathbf{H}_k$ (say within Euclidean distance γ). Then we will apply Theorem 43 to conclude that any point $\{-1,1\}^n \cup \{0,1\}^n$ which is not on $\mathbf{H}_k$ must have Euclidean distance at least some γ' from $\mathbf{H}_k$. If $\gamma' > \gamma$ then we can infer that every point in S' lies on $\mathbf{H}_k$, which proves the theorem. We now describe the construction that gives $\mathbf{H}_k$.

If $\dim(A_1) = n-1$, then we let $k = 1$ and stop the process, since as desired we have $\mathrm{span}(\mathbf{H}_k \cap (\{-1,1\}^n \cup \{0,1\}^n)) = \mathbf{H}_k$, $A_1 = H_k$, and $d(x, \mathbf{H}_k) \leq 2n\beta$ for every $x \in S'$. Otherwise, by an inductive hypothesis, we may assume that for some $j \geq 1$ we have an affine space A_j and a hyperplane $\mathbf{H}_j$ such that

- $A_1 \subseteq A_j \subsetneq \mathbf{H}_j$;
- $\dim(A_j) = \dim(A_1) + j - 1$, and
- for all $x \in S'$ we have $d(x, \mathbf{H}_j) \leq 2^j n\beta$.

Using this inductive hypothesis, we will construct an affine space A_{j+1} and a hyperplane $\mathbf{H}_{j+1}$ such that $A_1 \subset A_{j+1} \subseteq \mathbf{H}_{j+1}$, $\dim(A_{j+1}) = \dim(A_1) + j$, and for all $x \in S'$ we have

$$d(x, \mathbf{H}_{j+1}) \leq 2^{j+1} n\beta.$$

If $A_{j+1} = \mathbf{H}_{j+1}$, we stop the process, else we continue.

We now describe the inductive construction. Since $A_j \subsetneq \mathbf{H}_j$, there must exist an affine subspace A'_j such that $A_j \subseteq A'_j \subsetneq \mathbf{H}_j$ and $\dim(A'_j) = n-2$. Let x_j denote $\arg\max_{x \in S'} d(x, A'_j)$. (We assume that $\max_{x \in S'} d(x, A'_j) > 0$; if not, then choose x_j to be an arbitrary point in $\{-1,1\}^n$ not lying on A'_j. In this case, the properties of the inductive construction will trivially hold.) Define $\mathbf{H}_{j+1} = \mathrm{span}(A'_j \cup x_j)$. It is clear that $\mathbf{H}_{j+1}$ is a hyperplane. We claim that for $x \in S'$ we have

$$d(x, \mathbf{H}_{j+1}) \leq d(x, \mathbf{H}_j) + d(x_j, \mathbf{H}_j) \leq 2^j n\beta + 2^j n\beta = 2^{j+1} n\beta.$$

To see this, observe that without loss of generality we may assume that $\mathbf{H}_j$ passes through the origin and thus A'_j is a linear subspace. Thus we have that $\|x_{\perp A'_j}\| \leq \|(x_j)_{\perp A'_j}\|$ for all $x \in S'$, where for a point $z \in \mathbb{R}^n$ we write $z_{\perp A'_j}$ to denote the component of x orthogonal to A'_j. Let $r = \|x_{\perp A'_j}\|$ and $r_1 = \|x_{j,\perp A'_j}\|$, where $r_1 \geq r$. Let θ denote the angle that $x_{\perp A'_j}$ makes with $\mathbf{H}_j$ and let ϕ denote the angle that $x_{\perp A'_j}$ makes with $(x_j)_{\perp A'_j}$. Then it is easy to see that $d(x, \mathbf{H}_{j+1}) = |r \cdot \sin(\theta - \phi)|$, $d(x, \mathbf{H}_j) = |r \cdot \sin(\theta)|$ and $d(x_j, \mathbf{H}_j) = |r_1 \cdot \sin(\phi)|$. Thus, we only need to check that if $r_1 \geq r$, then $|r \cdot \sin(\theta - \phi)| \leq |r \cdot \sin(\theta)| + |r_1 \cdot \sin(\phi)|$ which is trivial to check.

Let $A_{j+1} = \mathrm{span}(A_j \cup x_j)$ and note that $A_1 \subset A_{j+1} \subseteq \mathbf{H}_{j+1}$ and $\dim(A_{j+1}) = \dim(A_j) + 1$. As shown above, for all $x \in S'$ we have $d(x, \mathbf{H}_{j+1}) \leq 2^{j+1} n\beta$. This completes the inductive construction.

Since $\dim(A_1) \geq n - \log(n+2) - \log(1/\alpha)$, the process must terminate for some $k \leq \log(n+2) + \log(1/\alpha)$. When the process terminates, we have a hyperplane $\mathbf{H}_k$ satisfying the following properties:

- $\mathrm{span}(\mathbf{H}_k \cap (\{-1,1\}^n \cup \{0,1\}^n)) = \mathbf{H}_k$; and
- $|\mathbf{H}_k \cap S| \geq \alpha 2^n/(n+2)$; and
- for all $x \in S'$ we have $d(x, \mathbf{H}_k) \leq 2^k n\beta \leq (1/\alpha)n(n+2)\beta$.

We can now apply Theorem 43 to the hyperplane $\mathbf{H}_k$ to get that if $\mathbf{H}_k = \{x \mid v \cdot x - \nu = 0\}$ with $\|v\| = 1$, then all the entries of v are integral multiples of a quantity E^{-1} where

$$E \leq (2n)^{\lfloor \log((n+2)/\alpha) \rfloor + 3/2} \cdot (\lfloor \log((n+2)/\alpha) \rfloor)!.$$

Consequently $v \cdot x$ is an integral multiple of E^{-1} for every $x \in (\{-1,1\}^n \cup \{0,1\}^n)$. Since there are points of $\{-1,1\}^n$ on $\mathbf{H}_k$, it must be the case that ν is also an integral multiple of E. So if any $x \in (\{-1,1\}^n \cup \{0,1\}^n)$ is such that $d(x, \mathbf{H}_k) < E$, then $d(x, \mathbf{H}_k) = 0$ and hence x actually lies on $\mathbf{H}_k$. Now recall that for any $x \in S'$ we have $d(x, \mathbf{H}_k) \leq (n/\alpha)(n+2)\beta$. Our upper bound on β from the theorem statement ensures that $(n/\alpha)(n+2)\beta < E^{-1}$, and consequently every $x \in S'$ must lie on $\mathbf{H}_k$, proving the theorem. $\square$