JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–24

Learning Halfspaces Under Log-Concave Densities: Polynomial Approximations and Moment Matching

Daniel M. Kane DANKANE@MATH.STANFORD.EDU
Stanford

Adam Klivans KLIVANS@CS.UTEXAS.EDU
UT-Austin

Raghu Meka RAGHU@IAS.EDU
IAS

Abstract

We give the first polynomial-time algorithm for agnostically learning any function of a constant number of halfspaces with respect to any log-concave distribution (for any constant accuracy parameter). This result was not known even for the case of PAC learning the intersection of two halfspaces.

We give two very different proofs of this result. The first develops a theory of polynomial approximation for log-concave measures and constructs a low-degree ℓ_1 polynomial approximator for sufficiently smooth functions. The second uses techniques related to the classical moment problem to obtain sandwiching polynomials. Both approaches deviate significantly from known Fourier-based methods, where essentially all previous work required the underlying distribution to have some product structure.

Additionally, we show that in the smoothed-analysis setting, the above results hold with respect to distributions that have sub-exponential tails, a property satisfied by many natural and well-studied distributions in machine learning.

Keywords: Log-concave distributions, smoothed analysis, halfspaces, agnostic learning, Fourier analysis

1. Introduction: Learning Intersections of Halfspaces

Learning halfspaces is one of the core algorithmic tasks in machine learning and can be solved in the noiseless (PAC) model via efficient algorithms for linear programming. The simplest generalization of this problem– learning intersections of halfspaces– has attracted the attention of many researchers in theoretical computer science and statistics. Surprisingly, learning the intersection of even two halfspaces with respect to arbitrary distributions remains a challenging open problem.

It is well known that even simple classes of intersections of halfspaces cannot be learned with respect to arbitrary distributions (e.g., polygons in the plane have infinite VC-dimension), and a major open problem is to classify the types of distributions for which intersections of halfspaces are efficiently learnable. Along these lines, Blum and Kannan (1993) gave an algorithm for learning intersections of m halfspaces with respect to Gaussian distributions on $\mathbb{R}^n$ in time $n^{O(2^m)}$ (for any constant accuracy parameter). Vempala (2010b) improved on this work and gave a randomized algorithm for learning intersections of centered halfspaces with respect to any log-concave distribution on $\mathbb{R}^n$ in time roughly $(n/\varepsilon)^{O(m)}$ ("centered" means that each bounding hyperplane passes through

the mean of the distribution). In a beautiful follow-up paper, Vempala (2010a) used PCA to give an algorithm for learning the intersection of m halfspaces with respect to any Gaussian distribution in time $\mathsf{poly}(n) \cdot (m/\varepsilon)^{O(m)}$. We note that these results hold in the PAC model, and it is not clear if they succeed in the agnostic setting.

In the agnostic model, we are only aware of results that use the polynomial regression algorithm of Kalai et al. (2008). Klivans et al. (2004) (combined with the observations in Kalai et al.) gave an algorithm for learning *any function* of m halfspaces in time $n^{O(m^2/\varepsilon^2)}$ with respect to the uniform distribution on $\{-1,1\}^n$. Applying results on Gaussian surface area, Klivans et al. (2008) gave an algorithm for agnostically learning intersections of m halfspaces in time $n^{\mathsf{polylog}(m)/\varepsilon^2}$ with respect to any Gaussian distribution.

A major goal in this area has been to move beyond Gaussians and tackle the case when the underlying distribution is log-concave, as log-concave densities are a broad and widely-studied class of distributions. The Gaussian density is log-concave, and, in fact, *any* uniform distribution over a convex set is log-concave.

Kalai et al. (2008) give an algorithm for agnostically learning a *single* halfspace with respect to any log-concave distribution in time $n^{f(\varepsilon)}$ for some function f. The best known bound for f is currently $2^{O(1/\varepsilon^2)}$ (follows from Section 5 of Lubinsky (2007)). It is unclear how to extend the Kalai et al. analysis to work for the intersection of two halfspaces.

To summarize, it was not known how to learn the intersection of two halfspaces with respect to log-concave distributions even in the noiseless (PAC) model.

1.1. Our Results

Here we give the first polynomial-time algorithm for agnostically learning intersections (or even arbitrary functions) of a constant number of halfspaces with respect to any log-concave distribution on $\mathbb{R}^n$(see Table 1.1 for the precise parameters):

Theorem 1 *Functions of m halfspaces are agnostically learnable with respect to any log-concave distribution on $\mathbb{R}^n$ in time $n^{O_{m,\varepsilon}(1)}$ where ε is the accuracy parameter.*

Admittedly, our dependence on the number of halfspaces m and the error parameter ε is not great, but we stress that no polynomial-time algorithm was known even for the intersection of two halfspaces. See Table 1.1 for a summary of previous work.

We extend the above result– in the *smoothed-analysis* setting– to hold with respect to arbitrary distributions with sub-exponential tail bounds. We first define the model of smoothed-complexity that we consider.

Definition 2 *Given a distribution $\mathcal{D}$ on $\mathbb{R}^n$, and a parameter $\sigma \in (0,1)$, let $\mathcal{D}(\sigma)$ be a perturbed distribution of $\mathcal{D}$ obtained by independently picking $X \leftarrow \mathcal{D}$, $Z \leftarrow \mathcal{N}(\mu,\Sigma)^n$ and outputting $X+Z$, where $\mu = X$, $\Sigma \succeq \sigma cov(X)$*[1].

That is, $\mathcal{D}(\sigma)$ is obtained by adding Gaussian noise to $\mathcal{D}$ and quantitatively, we want the variance of the noise in any direction to be comparable to (at least σ^2 times) the variance of $\mathcal{D}$ in the same direction. For instance, for $\mathcal{D}$ isotropic, perturbations by $\mathcal{N}(0,\sigma)^n$ would suffice. The latter corresponds more directly to the traditional smoothed-complexity setup, but we use the above definition as it is basis independent and allows for non-spherical Gaussian perturbations.

1. Here, $\succeq$ denotes the semi-definite ordering.

Concept Class	Distribution	Running Time	Model	Source
Intersections	Gaussian	$\mathsf{poly}(n) \cdot (m/\varepsilon)^m$	PAC	Vempala (2010a)
Intersections	Gaussian	$n^{(\mathsf{polylog}(m)/\varepsilon^{O(1)})}$	Agnostic	Klivans et al. (2008)
Intersections (centered)	Log-concave	$(n/\varepsilon)^m$	PAC	Vempala (2010b)
One halfspace	Log-concave	$n^{f(\varepsilon)}$	Agnostic	Kalai et al. (2008)
Arbitrary	Log-concave	$n^{\exp(O(m^4/\varepsilon^4))}$	Agnostic (convolution proof)	This work
Arbitrary	Log-concave	$n^{\exp((\log(1/\varepsilon))^{\tilde{O}(m)}/\varepsilon^4)}$	Agnostic (moment-matching)	This work
Arbitrary	Sub-exponential	$n^{\exp((\log(\log m/\sigma\varepsilon))^{\tilde{O}(m)}/\sigma^4\varepsilon^4)}$	Agnostic (σ-smoothed)	This work
Arbitrary	Sub-gaussian	$n^{(\log(\log m/\sigma\varepsilon))^{\tilde{O}(m)}/\sigma^4\varepsilon^4}$	Agnostic (σ-smoothed)	This work

Figure 1: Summary of recent work on learning intersections and arbitrary functions of m halfspaces

We define the smoothed-complexity of (agnostically) learning a concept class $\mathcal{C}$ under a distribution $\mathcal{D}$ to be the complexity of (agnostically) learning $\mathcal{C}$ under the perturbed distributions $\mathcal{D}(\sigma)$. This model first appears in the work of Blum and Dunagan (2002) (for the special case of spherical Gaussian perturbations) and we believe it to be a natural and practical extension of the traditional models of learning. For instance, the main motivating principle behind smoothed-analysis– that real data involves measurement error– is very much applicable here. Besides the work of Blum and Dunagan, there seems to be little known about learning in this model.

We say a distribution is sub-exponential (sub-gaussian) if every marginal (i.e., one-dimensional projection) of the distribution obeys a tail bound of the form $e^{-|z|}$ ($e^{-|z|^2}$, respectively). It is known that all log-concave distributions are sub-exponential. Sub-exponential and sub-gaussian densities are commonly studied in machine learning and statistics and model various real-word situations (see Buldygin and Kozachenko (2000) for instance). We show that for these types of distributions, our learning algorithms have polynomial smoothed-complexity (for constant σ):

Theorem 3 *Functions of m halfspaces are agnostically learnable with respect to any sub-exponential distribution on $\mathbb{R}^n$ in time $n^{O_{m,\varepsilon,\sigma}(1)}$ where ε is the accuracy parameter and σ is the perturbation parameter.*

We obtain much better parameters (in the constant hidden in $O_{m,\varepsilon,\sigma}(1)$) for the special case of sub-gaussian densities (see Theorem 26).

Blum and Dunagan were the first to study the smoothed complexity of learning halfspaces. They showed that for a single halfspace in the noiseless (in labels) setting, the perceptron algorithm converges quickly with high probability for examples perturbed by Gaussian noise. Their expected

running time, however, is infinite (and thus strictly speaking does not give bounds on the smoothed-complexity of the Perceptron algorithm).

To obtain our smoothed-analysis results, we prove that Gaussian perturbations provide enough *anti -concentration* for our polynomial approximation methods to work. We believe this connection will find additional applications related to the smoothed-complexity of learning Boolean functions.

1.2. Overview of Conceptual and Technical Contributions

In their seminal paper, Linial et al. Linial et al. (1993) introduced the polynomial approximation approach to learning theory. The core of their approach is to solve the following optimization problem: given a Boolean function f, minimize, over all polynomials p of degree at most d, the quantity $\mathbf{E}_{x\in\{-1,1\}^n}[(f-p)^2]$.

The algorithm is given uniformly random samples of the form $(x, f(x))$. Their "low-degree" algorithm approximately solves this optimization problem in time roughly $n^{O(d)}$. Later, the "sparse" algorithm of Kushilevitz and Mansour Kushilevitz and Mansour (1993) solved the same optimization problem but where the minimization is over all sparse polynomials, and the algorithm is allowed query access to the function f. These algorithms were developed in the context of PAC learning.

Kalai et al. Kalai et al. (2008) subsequently observed that in order to succeed in the agnostic framework of learning (we formally define agnostic learning in Section 2.1 but for now agnostic learning can be thought of as a model of PAC learning with adversarial noise), it suffices to approximately minimize $\mathbf{E}_{x\in\{-1,1\}^n}[|f-p|]$.

That is, minimizing with respect to the 1-norm rather than the 2-norm results in highly noise-tolerant learning algorithms. Finding efficient algorithms for directly minimizing the above expectation with respect to the 1-norm ("ℓ_1 minimization"), however, is more challenging than in the ℓ_2 case. The work of Kalai et al. Kalai et al. (2008) gives the analogue of the "low-degree" algorithm for ℓ_1 minimization and the work of Gopalan et al. Gopalan et al. (2008) gives the analogue of the "sparse" algorithm for ℓ_1 minimization.

Although, we have efficient algorithms that directly carry out ℓ_1 minimization for low-degree polynomials, proving the *existence* of good low-degree ℓ_1 approximators has required first finding a good low-degree ℓ_2 approximator (i.e., Fourier polynomial) and then applying the simple fact that $\mathbf{E}[|p|] \leq \sqrt{\mathbf{E}[p^2]}$. Directly analyzing the error of low-degree ℓ_1 approximators seems quite difficult. In our setting, for example, it is not even clear that the best low-degree ℓ_1 polynomial approximator is unique!

The main conceptual contribution is the first framework for *directly* proving the existence of low-degree ℓ_1 approximating polynomials for Boolean functions (in fact, we also obtain *sandwiching* polynomials). One benefit of our approach is that we do not require the underlying distribution to be product (essentially all of the techniques involving the discrete Fourier polynomial require some sort of product structure). As such, in this work, we are able to reason about approximating Boolean functions with respect to interesting non-product distributions, such as log-concave densities.

In fact, we give two very different frameworks and two very different proofs within these frameworks for establishing the existence of such approximating polynomials. The proofs were obtained independently by different sets of authors and are merged in this submission. To compare the two proofs, the first– a "convolution proof"– has better dependence on the relevant parameters; the second– a "moment-matching proof"– may be applicable to a wider class of distributions and gives *sandwiching approximations*. In the Appendix we show how to use "moment-matching" to learn with respect to interesting non-product distributions on the hypercube.

In the following descriptions, we assume we are trying to show polynomial approximations for $f : \mathbb{R}^n \to \{0,1\}$, where $f = g(h_1(x), \ldots, h_m(x))$, where $g : \{0,1\}^m \to \{0,1\}$ is an arbitrary Boolean function and $h_1, \ldots, h_m : \mathbb{R}^n \to \{0,1\}$ are halfspaces.

A "Convolution" Proof. In our first method of proof, in Section 3, we begin by projecting our measure down to $\mathbb{R}^m$ as defined by the halfspaces in question. We then approximate the function f by first finding a polynomial approximation to some kernel, ρ, and then approximating f by its convolution with ρ.

Much of the effort is spent finding an appropriate ρ. It is noted that subexponential tail bounds mean that any polynomial with tight bounds on its higher-order derivatives is well-approximated by a sufficiently high degree Taylor polynomial with respect to the distribution of interest. While such functions cannot be tightly concentrated enough to produce an appropriate kernel, we show that products of such functions may also be polynomially approximated, and in particular note that a high power of such a function will yield an appropriate kernel ρ.

A "Moment-Matching" Proof. The second proof, found in Section 4, uses ideas from probability theory and linear programming to give a framework for proving the existence of sandwiching polynomials (it is easy to see that sandwiching polynomials are stronger than ℓ_1 approximators). The main technical contribution is to show how to use a set of powerful theorems from the study of the *classical moment problem* to apply our framework to functions of halfspaces. At a high level, our approach makes crucial use of the following consequence of strong duality for semi-infinite linear programs: let $\mathcal{D}$ be a distribution and let $\mathcal{D}_k$ be any distribution where all moments of order less than or equal to k match those of $\mathcal{D}$. If $\mathbf{E}_{\mathcal{D}}[f]$ is "close" to $\mathbf{E}_{\mathcal{D}_k}[f]$ then f has a low-degree sandwiching polynomials with respect to D. The question then becomes how to analyze the bias of a Boolean function where only the low-order moments of a distribution have been specified. We show how to use several deep results from probability to answer this question in Sections 4.3 and 4.4.

We show that the moment-matching approach also has some interesting applications for learning with respect to distributions on the discrete cube $\{-1,+1\}^n$. Due to lack of space, we defer this section to the Appendix.

2. Preliminaries

Agnostic Learning. We recall the model of agnostically learning a concept class $\mathcal{C}$ Haussler (1992), Kearns et al. (1994). In this scenario there is an unknown distribution $\mathcal{D}$ over $\mathbb{R}^n \times \{-1,1\}$ with marginal distribution over $\mathbb{R}^n$ denoted $\mathcal{D}_X$.

Let $\mathsf{opt} \stackrel{\text{def}}{=} \inf_{f \in \mathcal{C}} \mathbf{Pr}_{(x,y)\sim\mathcal{D}}[f(x) \neq y]$; i.e. opt is the minimum error of any function from $\mathcal{C}$ in predicting the labels y. The learner must output a hypothesis whose error is within ε of opt:

Definition 4 *Let $\mathcal{D}$ be an arbitrary distribution on $\mathbb{R}^n \times \{-1,1\}$ whose marginal over $\mathbb{R}^n$ is $\mathcal{D}_X$, and let $\mathcal{C}$ be a class of Boolean functions $f : \mathbb{R}^n \to \{-1,1\}$. We say that algorithm B is an agnostic learning algorithm for $\mathcal{C}$ with respect to $\mathcal{D}$ if the following holds: for any $\mathcal{D}$ as described above, if B is given access to a set of labeled examples (x,y) drawn from $\mathcal{D}$, then with probability at least $1-\delta$ algorithm B outputs a hypothesis $h : \mathbb{R}^n \to \{-1,1\}$ such that $\mathbf{Pr}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \leq \mathsf{opt} + \varepsilon$.*

Note that PAC learning is a special case of agnostic learning (the case when $\mathsf{opt} = 0$).

The "L_1 Polynomial Regression Algorithm" due to Kalai et al. (2008) shows that one can *agnostically* learn any concept class that can be approximated by low-degree polynomials:

Theorem 5 (Kalai et al. (2008)) *Fix $\mathcal{D}$ on $X \times \mathbb{R}$ and let $f \in \mathcal{C}$. Assume there exists a polynomial p of degree d such that $\mathbf{E}_{x\sim\mathcal{D}_X}[|f(x) - p(x)|] < \varepsilon$ where $\mathcal{D}_X$ is the marginal distribution on X. Then, with probability $1-\delta$, the L_1 Polynomial Regression Algorithm outputs a hypothesis h such that $Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \leq opt + \varepsilon$ in time $\mathrm{poly}(n^d/\varepsilon, \log(1/\delta))$.*

Throughout, we suppress the $\mathrm{poly}(\log(1/\delta))$ dependence on δ.

3. The Convolution Proof

The bulk of our work will be to show that if $f : \mathbb{R}^n \to [-1, 1]$ is an function of m halfspaces and if $\mathcal{D}$ is a log-concave distribution on $\mathbb{R}^n$ there there exists some polynomial p of bounded degree that approximates f to within ε in $L^1(\mathcal{D})$. For such a function f, it is clear that $f(x)$ depends only on $\langle x, w_i\rangle$ for some m vectors w_i. If we choose a polynomial p depending only on these linear functions, we can project our problem about a polynomial approximation in $\mathbb{R}^n$ to a problem about approximation in $\mathbb{R}^m$. By Theorem 5.1 of Lovász and Vempala (2003), the projection of the measure $\mathcal{D}$ onto $\mathbb{R}^m$ is itself log-concave, and thus it suffices to solve this problem in the special case of $n = m$. By applying an appropriate affine transformation we may assume that our distribution $\mathcal{D}$ has mean 0 and covariance matrix given by the identity. In this Section, we develop a theory to show the existence of such approximations. In particular we show:

Theorem 6 *Let $m > 0$ be an integer. Let $\mathcal{D}$ be a log-concave distribution on $\mathbb{R}^m$ with mean 0 and identity covariance matrix. Let $f : \mathbb{R}^m \to [-1, 1]$ be a function. The for any $\varepsilon, \delta > 0$, there exists a polynomial p on $\mathbb{R}^m$ of degree at most $d = \exp\left(O\left(m^4\delta^{-4} + \log^2(\varepsilon^{-1})\right)\right)$ so that*

$$\mathbf{E}_{X\sim\mathcal{D}}[|f(X) - p(X)|] \leq \varepsilon + 2\mu_{\mathcal{D}}(S_\delta)$$

where $S_\delta := \{x \in \mathbb{R}^m | \exists y \in \mathbb{R}^m : |x - y| < \delta$ and $f(x) \neq f(y)\}$.

Before we begin the proof of Theorem 6, we will need to recall some results on log-concave distributions.

Theorem 7 *[Theorem 5.1 of Lovász and Vempala (2003)] Any projection of a log-concave distribution is log-concave.*

It should be noted that the above result also follows easily from the Prékopa-Leindler Inequality.

Lemma 8 *[Lemma 6 of Lovász and Vempala (2003)] Let ν be a log-concave measure on $\mathbb{R}$ with mean 0 and variance 1. Then the probability density function of ν is bounded above by $e^{-|x|/16}$.*

Corollary 9 *Let $\mathcal{D}$ be a log-concave distribution on $\mathbb{R}^m$ with mean 0 and identity covariance matrix, then the probability that $|X| > R$ for $X \sim \mathcal{D}$ is at most $2^{O(m)}e^{-R/32}$.*

The proof of Corollary 9 is standard and deferred to the full version. Throughout this section we shall use the notation that if $g : \mathbb{R}^m \to \mathbb{R}$ is a function, then $|g|_1$ denotes the L^1 norm of g with respect to $\mathcal{D}$: $|g|_1 = \mathbf{E}_{X\sim\mathcal{D}}[|g(X)|]$.

We are now ready to begin to develop a theory of polynomial approximation with respect to log-concave measures. We start by introducing a class of functions that will turn out to be easy to approximate.

Definition 1 *We say that a function $f : \mathbb{R}^m \to \mathbb{R}$ is c-smooth (for some $c > 0$), if for every non-negative integer k, and every unit vector x we have that $|D_x^k f|_\infty \leq c^k$. Where $D_x^k f$ is the k^{th} order directional derivative of f in the x-direction.*

The class of c-smooth functions will be important because such functions can be well approximated by their Taylor polynomials.

Lemma 10 *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a c-smooth function. Then for every non-negative integer d, there exists a degree-d polynomial p so that $|f(x) - p(x)| \leq (c|x|)^{d+1}/(d+1)!$.*

Proof We choose p to be the Taylor polynomial of f around 0 of degree-d. In order to prove the desired bound, we let $x = ty$ for y a unit vector. Define the function $g : \mathbb{R} \to \mathbb{R}$ by $g(s) = f(sy)$. It is clear that the polynomial $p(sy)$ is the degree-d Taylor polynomial of g. Therefore,

$$|f(x) - p(x)| = |g(s) - p(sy)| \leq \frac{s^{d+1}}{(d+1)!}|g^{(d+1)}|_\infty \leq \frac{|x|^{d+1}}{(d+1)!}|D_x^{d+1} f|_\infty \leq \frac{(c|x|)^{d+1}}{(d+1)!}.$$

■

Our construction will depend critically on using the above result to find good approximations for powers of a c-smooth function.

Lemma 11 *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a c-smooth function. Let k, N be positive integers. There exists a polynomial p of degree at most $k2^N$ so that $|(f(x))^N - p(x)| \leq \sum_{j=0}^{n-1} 2^N (2c|x|)^{2^j k}/(2^j k)!$.*

Proof deferred to the Appendix Section A.1

Corollary 12 *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a c-smooth function for $c < \frac{1}{256}$. Let $\mathcal{D}$ be a log-concave distribution on $\mathbb{R}^m$ with mean 0 and identity covariance matrix. Let k, N be positive integers. There exists a polynomial p of degree at most $k2^N$ so that $|f^N - p|_{1,\mathcal{D}} \leq 2^N 2^{O(m)} 2^{-k}$.*

Proof deferred to the Appendix Section A.2

The great advantage of Lemma 11 and Corollary 12 is that although f being c-smooth will necessitate that f is not sharply concentrated at any point, f^N may well have this property. In fact, we will be able to approximate our desired function by its convolution with f^N for an appropriate power N. Before we can do this, we will want to find some useful c-smooth functions.

Lemma 13 *Let b be a function supported on the ball of radius c about the origin with $|b|_1 \leq 1$. Then the Fourier transform, $\hat{b}$ of b is c-smooth.*

Proof If x is a unit vector, then $D_x^k(\hat{b})$ is given by the Fourier transform of $(ix)^k b$. Since $|(ix)^k b|_1 \leq c^k |b|_1 = c^k$, the L^∞ norm of $D_x^k(\hat{b})$ is at most c^k. Hence $\hat{b}$ is c-smooth. ■

For any constant $c > 0$, we let $\rho_c(x) := \prod_{i=1}^m \text{sinc}(cx_i/\sqrt{m})$. We note that $\rho_c(x)$ is the Fourier transform of the uniform probability distribution over the box of side length $c/\sqrt{m}$. This by Lemma 13, $\rho_c(x)$ is c-smooth.

We now consider the behavior of $\rho_c(x)^N$ for N a large even integer. For $|x| \leq c^{-1}\sqrt{\frac{m}{N}}$, it is not hard to see that $|\rho_c(x)^N| = \Omega(1)$. This happens if $|x_i| \leq c^{-1}N^{-1/2}$ for each i. Thus

the integral of $\rho_c(x)^N$ over $\mathbb{R}^m$ is at least $\Omega(c^{-m}N^{-m/2})$. On the other hand, it is easy to see that $\int_{x\in\mathbb{R}|x|>N^{-1/4}} \text{sinc}(x)^N dx = \exp(-\Omega(\sqrt{N}))$. Thus the integral of $\rho_c(x)^N$ over the region where $|x| > c^{-1}mN^{-1/4}$ is at most $c^{-m}m^{m/2}\exp(-\Omega(\sqrt{N}))$. These results together imply that an appropriate multiple of $\rho_c(x)$ would make a suitable approximation to the δ-function.

We have now stated all the important technical ingredients for proving Theorem 6. We show how to combine them (in a not too difficult way) to prove Theorem 6 in the Appendix, Section A.3.

Using Theorem 6, it is easy to show that the characteristic function of an intersection of half-spaces can be approximated in L^1 with respect to any log-concave measure.

Corollary 14 *Let $\mathcal{D}$ be any log-concave distribution on $\mathbb{R}^n$. Let $f : \mathbb{R}^n \to [-1, 1]$ be a function so that $f(x)$ depends only on the values of $\mathsf{sgn}(\langle x, w_i\rangle - \theta_i)$ for some fixed vectors $w_1, \ldots, w_m$ and real numbers $\theta_1, \ldots, \theta_m$ (so for example f could be the indicator function of an intersection of m half-spaces). Let $\varepsilon > 0$ be a real number. Then there exists a polynomial p of degree $d = \exp(O(m^4\varepsilon^{-4}))$ so that $\mathbf{E}_{X\sim\mathcal{D}}[|f(X) - p(X)|] \le \varepsilon$.*

Proof Let y be the vector given by $y_i = \langle x, w_i\rangle$. By Theorem 7, the induced probability distribution on y is log-concave. We note that y takes values in $\mathbb{R}^m$. If it is the case that the w_i are linearly dependent, then y will only take values along some subspace. In such a case, we project y further by removing coordinates until this is no longer the case. Replacing y be an affine transformation if necessary, we may assume that y is mean 0 and has identity covariance matrix. By construction, it is the case that the value of $f(x)$ is equal to $g(y)$ for some function g. By Theorem 6, there is a polynomial q on $\mathbb{R}^m$ of degree d so that $\mathbf{E}[|q(y) - g(y)|] \le \frac{\varepsilon}{2} + 2\mu(S_{\varepsilon/(8m)})$. Where above μ is the appropriate measure on y, and $S_{\varepsilon/(8m)}$ is the set of points within distance at most $\varepsilon/(8m)$ of one of the planes corresponding to $\langle x, w_i\rangle = \theta_i$. Let z be the signed distance from one of these planes. By assumption, the covariance matrix for y is the identity matrix. This implies that the variance of z is 1. By Theorem 7, the distribution on z is log-concave, and therefore by Lemma 8, the probability that $|z| \le \varepsilon/(8m)$ is at most $\varepsilon/(8m)$. Summing this probability over all m of these hyperplanes, yields that $\mu(S_{\varepsilon/(8m)}) \le \varepsilon/4$. Therefore we have that $\mathbf{E}[|q(y) - g(y)|] \le \varepsilon$. Letting $p(x) = q(y)$, yields our result. ■

Theorem 1 with the runtime given in Table 1.1 follows from the above result and Theorem 5.

4. Moment-Matching Proof

The second proof develops a theory of "moment-matching polynomials." Our main result is the following.

Theorem 15 *Let $\mathcal{D}$ be a log-concave distribution over $\mathbb{R}^n$. Let $h_1, \ldots, h_m : \mathbb{R}^n \to \{1, -1\}$, be halfspaces and let $g : \{1, -1\}^m \to \{1, -1\}$ be an arbitrary function. Define $f : \mathbb{R}^n \to \{1, -1\}$ by $f(x) = g((h_1(x), \ldots, h_m(x)))$. Then, there exists a real-valued polynomial P of degree at most $k = \exp((\log((\log m)/\varepsilon))^{O(m)}/\varepsilon^4)$ such that $\mathbb{E}_{X\leftarrow D}[\,|\,f(X) - P(X)\,|\,] \le \varepsilon$.*

Theorem 1 with the runtime given in Table 1.1 follows from the above result and Theorem 5. The theorem is proved in Section 4.4. We start with some preliminaries.

4.1. Preliminaries

We start with some notational conventions. For a random variable $X \in \mathbb{R}^m$, let $\varphi_X : \mathbb{R}^m \to \mathbb{R}$ be the characteristic function defined by $\varphi_X(t) = \mathbb{E}[\exp(-i\langle t, x\rangle)]$, where $i = \sqrt{-1}$. For $I =$

$(i_1, \ldots, i_n) \in \mathbb{Z}^n$, and $x \in \mathbb{R}^n$, let $x(I) = \prod_{j=1}^n x_j^{i_j}$. For $k > 0$, let $I(k,n) = \{I = (i_1, \ldots, i_n) \in \mathbb{Z}^n : \sum_{j=1}^n i_j \leq k,\ i_j \geq 0\}$.

We say that a class of functions $\mathcal{C}$ is ε-approximated in ℓ_1 by polynomials of degree d under a distribution $\mathcal{D}$ if for every $f \in \mathcal{C}$, there exists a degree d polynomial p such that $E_{x\sim\mathcal{D}}[|p(x) - f(x)|] \leq \varepsilon$.

We shall use the following measures of closeness between random variables $X, Y \in \mathbb{R}^m$.
(1) The λ-metric: $\mathsf{d}_\lambda(X,Y) = \min_{T>0}\ \max\{\max_{\|t\|\leq T}\{|\varphi_X(t) - \varphi_Y(t)|\}, 1/T\}$.
(2) The Levy distance: for $\mathbf{1}$ being the all 1's vector,

$$\mathsf{d}_{\mathsf{LV}}(X,Y) = \inf_{\varepsilon>0} \{\forall t \in \mathbb{R}^m,\ \mathsf{Pr}[X < t - \varepsilon\mathbf{1}] - \varepsilon < \mathsf{Pr}[Y < t] < \mathsf{Pr}[X < t + \varepsilon\mathbf{1}] + \varepsilon\}.$$

(3) Kolmogorov-Smirnov or cdf distance: $\mathsf{d}_{\mathsf{cdf}}(X,Y) = \sup_{t\in\mathbb{R}^m} \{|\,\mathsf{Pr}[X \geq t] - \mathsf{Pr}[Y \geq t]|\}$.

We use the following properties of log-concave distributions (equivalent formulations can be found in Lovász and Vempala (2003)).

Theorem 16 (Carbery and Wright (2001)) *Let random-variable $X \in \mathbb{R}^n$ be drawn from a log-concave distribution. Then, for every $w \in \mathbb{R}^n$, and $r > 0$, $\mathbb{E}[|\langle w, X\rangle|^r] \leq r^r \cdot \mathbb{E}[\langle w, X\rangle^2]^{r/2}$.*

Theorem 17 (Carbery and Wright (2001)) *There exists a universal constant C such that the following holds. For any real-valued log-concave random variable X with $\mathbb{E}[X^2] = 1$ and all $t \in \mathbb{R}$, $\varepsilon > 0$, $\mathsf{Pr}[X \in [t, t+\varepsilon]] < C\varepsilon$.*

We also use the following simple lemmas. The first helps us convert closeness in Levy distance to closeness in cdf distance, while the second helps us go from fooling intersections of halfspaces to fooling arbitrary functions of halfspaces.

Fact 18 *Let $X = (X_1, \ldots, X_m) \in \mathbb{R}^m$ be a random variable such that for every $r \in [m]$, $t \in \mathbb{R}, \varepsilon > 0$, $\mathsf{Pr}[X_r \in [t, t+\varepsilon]] < \beta \cdot \varepsilon$ for a fixed $\beta > 0$. Then, for any random variable Y, $\mathsf{d}_{\mathsf{cdf}}(X,Y) \leq m \cdot \beta \cdot \mathsf{d}_{\mathsf{LV}}(X,Y)$.*

Lemma 19 *Let $X, Y \in \mathbb{R}^m$ be real-valued random variables such that for every $a_1, \ldots, a_m \in \{1,-1\}$, $\mathsf{d}_{\mathsf{cdf}}(\,(a_1X_1, a_2X_2, \ldots, a_mX_m),\ (a_1Y_1, a_2Y_2, \ldots, a_mY_m)\,) \leq \varepsilon$. Then, for any function $g : \{1,-1\}^m \to \{1,-1\}$ and thresholds $\theta_1, \ldots, \theta_m$, $|\,\mathbb{E}[g(\,\mathsf{sign}(X_1 - \theta_1), \ldots, \mathsf{sign}(X_m - \theta_m)\,)] - \mathbb{E}[g(\,\mathsf{sign}(Y_1 - \theta_1), \ldots, \mathsf{sign}(Y_m - \theta_m)\,)]\,| \leq 2^m\varepsilon$.*

Proof deferred to Appendix, Section B.1.

4.2. LP Duality

It is now well known in the pseudorandomness literature that with respect to the uniform distribution over $\{-1,1\}^n$, a concept class $\mathcal{C}$ has degree k sandwiching polynomials if and only if $\mathcal{C}$ is fooled by k-wise independent distributions Bazzi (2009). The proof of this fact follows from LP duality where feasible solutions to the primal are k-wise independent distributions and feasible dual solutions are approximating polynomials.

In our setting, we consider continuous distributions over $\mathbb{R}^n$ that are not necessarily product. As such, this equivalence is more subtle. In fact, it is not even clear how to define k-wise independence for non-product distributions (such as log-concave densities). Still, given a distribution $\mathcal{D}$ we can write a semi-infinite linear program (a program with infinitely many variables but finitely many

constraints) whose feasible solutions are distributions that match all of $\mathcal{D}$'s moments up to degree k (in the case where $\mathcal{D}$ is uniform over $\{-1,1\}^n$, matching all moments is equivalent to being k-wise independent).

For $I \in I(k,n)$, let $\sigma_I = \mathbb{E}_{X \leftarrow \mathcal{D}}[X(I)]$. Let $f \in \mathcal{C}$. We write the primal program as follows:

$$\sup_{\mu} \int_{\mathbb{R}^k} f(x)\mu(x)dx$$
$$\int_{\mathbb{R}^k} x(I)\mu(x)dx = \sigma_I, \quad \forall I \in I(k,n), \qquad \int_{\mathbb{R}^k} \mu(x) = 1. \tag{1}$$

The supremum is over all probability measures μ on $\mathbb{R}^k$. As in the finite dimensional case, feasible solutions to the dual program correspond to degree k approximating polynomials. The dual can be written as

$$\inf_{a \in \mathbb{R}^{I(k,n)}} \sum_{I \in I(k,n)} a_I \sigma_I$$
$$\sum_I a_I x(I) \geq f(x), \quad \forall x \in \mathbb{R}^k. \tag{2}$$

The issue here is that in general, strong duality does not hold for semi-infinite linear programs. In our case, however, where the σ_i's are obtained as moments from a distribution $\mathcal{D}$ (as opposed to just arbitrary reals), it turns out that strong duality does hold. To see this, we note that the above primal LP is a special case of the so-called *generalized moment problem* LP, a classical problem from probability and analysis that asks if there exists a multivariate distribution with moments specified by the σ_i's. In our case, feasibility is immediate, as the σ_i's are obtained from $\mathcal{D}$.

As for strong duality, it is known that if the σ_i's are in the interior of a particular set (the details are not relevant here), then the optimal value of the primal equals the optimal value of the dual. In the case that the σ_i's do not satisfy this condition, strong duality holds assuming we relax the dual program constraints to some subset $\Omega \subseteq \mathbb{R}^n$. One concern is that we will now obtain an optimal approximating polynomial with respect to some distribution $\mathcal{D}'$ defined on Ω (as opposed to the original $\mathcal{D}$). But it is also known that in this case, *all* feasible distributions are supported on Ω. As such, approximation with respect to $\mathcal{D}'$ is equivalent to approximation with respect to $\mathcal{D}$. We refer the reader to Bertsimas and Popescu (2005) (Section 2) for more details and references. We next given an important definition.

Definition 20 *Given two distributions $\mathcal{D}, \mathcal{D}'$ on $\mathbb{R}^n$, $k \geq 0$, we say $\mathcal{D}'$ k moment-matches $\mathcal{D}$ if for all $I \in I(k,n)$, $\mathbb{E}_{X \leftarrow \mathcal{D}}[X(I)] = \mathbb{E}_{X \leftarrow \mathcal{D}'}[X(I)]$.*

We can now prove the main lemma of this section:

Lemma 21 *Let $f : \mathbb{R}^n \to \{0,1\}$ and let $\mathcal{D}$ be a distribution over $\mathbb{R}^n$ with all moments finite such that the following holds: For every distribution $\mathcal{D}'$ that k moment-matches $\mathcal{D}$, $|\mathbb{E}_{X \leftarrow \mathcal{D}}[f(X)] - \mathbb{E}_{X \leftarrow \mathcal{D}'}[f(X)]| < \varepsilon$. Then, there exist degree at most k polynomials $P_\ell, P_u : \mathbb{R}^n \to \mathbb{R}$ such that*

- *For every $x \in Support(\mathcal{D})$, $P_\ell(x) \leq f(x) \leq P_u(x)$.*
- *For $X \leftarrow \mathcal{D}$, $\mathbb{E}[P_u(X)] - \mathbb{E}[f(X)] \leq \varepsilon$ and $\mathbb{E}[f(X)] - \mathbb{E}[P_\ell(X)] \leq \varepsilon$.*

Proof Let opt^* be the value of the primal program Equation 1. Then, by hypothesis $opt^* < \gamma + \varepsilon$, where $\gamma = \mathbb{E}_{X \leftarrow \mathcal{D}}[f(X)]$.

Now, from the above discussion, strong duality (almost) holds for the programs in Equations (1) and (2), and we conclude that there exists a dual solution $a \in \mathbb{R}^{I(k,n)}$ with value exactly opt^* that satisfies the inequality constraints for all $x \in Support(\mathcal{D})$. Define, $P_u(x_1, \ldots, x_n) = \sum_{I \in I(k,n)} a_I x(I)$.

Then, $P_u(\ )$ is a degree at most k polynomial, and $P_u(x) \geq f(x)$ for every $x \in Support(\mathcal{D})$. Further, the assumption in the lemma implies $\mathbb{E}_{X \leftarrow \mathcal{D}}[P_u(X)] = \sum_{I \in I(k,n)} a_I \sigma_I = opt^* < \gamma + \varepsilon$. We have the existence of the lower sandwiching polynomial P_ℓ similarly. ■

4.3. The Classical Moment Problem

In the previous section, we reduced the problem of constructing low-degree sandwiching polynomial approximators with respect to $\mathcal{D}$ to understanding the optimal value of a semi-infinite linear program. The feasible solutions of the linear program correspond to all distributions that are k moment-matching to $\mathcal{D}$. As such, for any k moment-matching distribution $\mathcal{D}'$ we need to bound $|E_{\mathcal{D}}[f] - E_{\mathcal{D}'}[f]|$. We need the following result showing that multivariate distributions whose marginals have matching lower order moments have close characteristic functions (as quantified by λ-metric) provided the moments are well behaved.

Theorem 22 (Theorem 2, Page 171, Klebanov and Rachev (1996)) *Let $X, Y \in \mathbb{R}^m$ be two random variables such that for any $t \in \mathbb{R}^m$, the real-valued random variables $\langle t, X\rangle, \langle t, Y\rangle$ have identical first $2k$ moments. Then, for a universal constant C,*

$$\mathsf{d}_\lambda(X, Y) \leq C\beta_k^{-1/4}\left(1 + \mu_2(X)^{1/2}\right),$$

where $\mu_j(X) = \sup\{\mathbb{E}[|\langle t, x\rangle|^j] : t \in \mathbb{R}^m, \|t\| \leq 1\}$, and $\beta_k = \beta_k(X) = \sum_{j=1}^k 1/\mu_{2j}(X)^{1/2j}$.

We now need to convert the above bound on closeness of characteristic functions to more direct measures of closeness like Levy or Kolmogorov-Smirnov metrics. Such inequalities play an important role in Fourier theoretic proofs of limit theorems (eg., Esseen's inequality; cf. Chapter XVI Feller (1971)). Here we use the following relation between d_λ and d_{LV} (proof deferred to Appendix B) which follows from a related inequality due to Gabovich (1981).

Lemma 23 *Let X, Y be two vector-valued random variables with $\mathsf{d}_\lambda(X, Y) \leq \delta$. Let $N(\varepsilon) \in \mathbb{R}$ be such that $\mathsf{Pr}[X \notin [-N(\varepsilon), N(\varepsilon)]^m], \mathsf{Pr}[Y \notin [-N(\varepsilon), N(\varepsilon)]^m] \leq \delta$. Then,*

$$\mathsf{d}_{\mathsf{LV}}(X, Y) \leq O\left((\log N(\delta) + 2\log(1/\delta))^m \cdot \delta\right).$$

4.4. Low-Order Moments, Functions of Halfspaces, and Log-Concave Densities

We are now ready to complete our second proof of the main theorem for learning functions of halfspaces with respect to log-concave distributions - Theorem 1. We do so by using the tools from the previous section on moment bounds to analyze the optimum value of the primal LP from Section 4. This will imply low-degree ℓ_1 approximators with low error for any $f \in \mathcal{C}$. We can then apply known results due to Kalai et al. (2008) (Theorem 5) relating approximability by low-degree polynomials and agnostic learning.

Proof [Proof of Theorem 15] Without loss of generality suppose that $\mathcal{D}$ is in isotropic position. We can do so, as any distribution can be brought to isotropic position by an affine transformation and the class of intersections of halfspaces is invariant under affine transformations.

Let halfspace $h_i : \mathbb{R}^n \to \{1, -1\}$ be given as $h_i(x) = \mathsf{sign}(\langle w_i, x\rangle - \theta_i)$ for $w_i \in \mathbb{R}^n$ with $\|w_i\| = 1$ and $\theta_i \in \mathbb{R}$.

Let $X \leftarrow \mathcal{D}$ and let $X' \leftarrow \mathcal{D}'$, where $\mathcal{D}'$ is any distribution that is $2k$-moment matching to $\mathcal{D}$ for $k = 2^{O((m/\varepsilon)^4)}$ to be chosen later. Let $Y = (\langle w_1, X\rangle, \langle w_2, X\rangle, \cdots, \langle w_m, X\rangle)$ and $Y' = (\langle w_1, X'\rangle, \cdots, \langle w_m, X'\rangle)$. Observe that for every $t \in \mathbb{R}^m$, the first $2k$ moments of $\langle t, Y\rangle, \langle t, Y'\rangle$ are identical. Thus, we can apply Theorem 22 to the random variables Y, Y'. For $t \in \mathbb{R}^m$, $\|t\| = 1$, $j > 0$,

$$\mathbb{E}[|\langle t, Y\rangle|^j] = \mathbb{E}[\,|\langle \sum_{r=1}^m t_r w_r, X\rangle|^j\,] \le j^j \cdot \mathbb{E}[\langle \sum_{r=1}^m t_r w_r, X\rangle^2]^{j/2} = j^j \cdot \|\sum_{r=1}^m t_r w_r\|^j \le j^j m^j,$$

where the first inequality follows from Theorem 16 and the second equality from X being isotropic. Therefore, for $\mu_j(Y)$ and β_k as defined in Theorem 22,

$$\beta_k = \sum_{j=1}^k \frac{1}{\mu_j(Y)^{1/2j}} \ge \sum_{j=1}^k \frac{1}{2mj} = \Omega((\log k)/m). \tag{3}$$

We now wish to get a good estimate on $N(\delta)$ as defined in Lemma 23. From Theorem 16 and Markov's inequality, for every $\alpha > 0$, and $r \in [m]$, $j \le 2k$ even

$$\mathsf{Pr}[|\langle w_r, X\rangle| > \alpha] \le \frac{\mathbb{E}[\langle w_r, X\rangle^j]}{\alpha^j} \le \frac{j^j}{\alpha^j}.$$

Therefore, for $j = \log(m/\delta)$, and $\alpha = 2j$, $\mathsf{Pr}[|\langle w_r, X\rangle| > 2j] < \delta/m$. Thus, by using a union bound over all the components of Y, for $N = 2j = 2\log(m/\delta)$,

$$\mathsf{Pr}[Y \notin [-N, N]^m] < \delta. \tag{4}$$

As the above calculation only involved the first $2k$ moments of X, the same property should hold for Y'. From Equations (3), (4) and Theorem 22,

$$\mathsf{d}_\lambda(Y, Y') \le O\left(\frac{m^{1/4}}{\log^{1/4} k}\right). \tag{5}$$

Let $k = 2^{O(m/\delta^4)}$ be large enough so that the above error bound is $\mathsf{d}_\lambda(Y, Y') \le \delta$. Therefore, from Lemma 23, $\mathsf{d}_{\mathsf{LV}}(Y, Y') \le (\log((\log m)/\delta))^{O(m)} \cdot \delta$. Now observe that by Theorem 17, for every $r \in [m]$, $t \in \mathbb{R}$, $\alpha > 0$, $\mathsf{Pr}[Y_r \in [t, t+\alpha]] = O(\alpha)$. Thus, from the above equation and Fact 18,

$$\mathsf{d}_{\mathsf{cdf}}(Y, Y') \le O(m \cdot \mathsf{d}_{\mathsf{LV}}(Y, Y')) = (\log((\log m)/\delta))^{O(m)} \cdot \delta \equiv \varepsilon. \tag{6}$$

Since the above argument worked for any weight vectors $w_1, \ldots, w_m \in \mathbb{R}^m$, a similar argument applied to weight vectors $a_1 w_1, a_2 w_2, \ldots, a_m w_m$ for $a \in \{1, -1\}^m$, gives

$$\mathsf{d}_{\mathsf{cdf}}((a_1 Y_1, \ldots, a_m Y_m), (a_1 Y'_1, \ldots, a_m Y'_m)) \le \varepsilon.$$

Therefore, by Lemma 19 applied to Y, Y' and g, $|\,\mathbb{E}[f(X)] - \mathbb{E}[f(X')]| \le 2^m \varepsilon$.

Hence, by Lemma 21, for $P \equiv P_u$ a degree at most k polynomial as in Lemma 21,

$$\mathbb{E}[|P(X) - f(X)|] = \mathbb{E}[P(X)] - \mathbb{E}[f(X)] \le 2^m \varepsilon.$$

The theorem now follows from setting $\varepsilon = \varepsilon'/2^m$ as $k = 2^{O(m/\delta^4)} = 2^{(\log((\log m)/\varepsilon)^{O(m)})/\varepsilon^4}$. ∎

5. Smoothed Complexity of Learning Functions of Halfspaces

We defer this section to the Appendix (Section C) due to lack of space.

Acknowledgments

Adam Klivans acknowledges support from an NSF CAREER award and NSF grant AF 1018829.

References

N.I Akhiezer. *The Classical Moment Problem and Some Related Questions in Analysis*. Hanfer Publishing Co, 1 edition, 1965.

Louay M. J. Bazzi. Polylogarithmic independence can fool DNF formulas. *SIAM J. Comput*, 38(6): 2220–2272, 2009. URL http://dx.doi.org/10.1137/070691954.

Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

Avrim Blum and John Dunagan. Smoothed analysis of the perceptron algorithm for linear programming. In *SODA*, pages 905–914, 2002.

Avrim Blum and Ravi Kannan. Learning an intersection of k halfspaces over a uniform distribution. In *FOCS*, pages 312–320, 1993.

V.V. Buldygin and I.U.V. Kozachenko. *Metric Characterization of Random Variables and Random Processes*. Translations of Mathematical Monographs. American Mathematical Society, 2000. ISBN 9780821805336. URL http://books.google.com/books?id=ePDXvIhdEjoC.

Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001. URL http://www.mrlonline.org/mrl/2001-008-003/2001-008-003-001.html.

Ilias Diakonikolas, Daniel M. Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *FOCS*, pages 11–20, 2010a.

Ilias Diakonikolas, Rocco A. Servedio, Li-Yang Tan, and Andrew Wan. A regularity lemma, and low-weight approximators, for low-degree polynomial threshold functions. In *IEEE Conference on Computational Complexity*, pages 211–222, 2010b.

William Feller. *An Introduction to Probability Theory and Its Applications, Vol. 2 (Volume 2)*. Wiley, 2 edition, January 1971. ISBN 0471257095. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471257095.

Yu. R. Gabovich. Stability of the characterization of the multivariate normal distribution in the Skitovich-Darmois theorem. *Journal of Mathematical Sciences*, 16:1341–1349, 1981. ISSN 1072-3374. URL http://dx.doi.org/10.1007/BF01091625.

Parikshit Gopalan, Adam Tauman Kalai, and Adam R. Klivans. Agnostically learning decision trees. In *STOC*, pages 527–536, 2008.

Prahladh Harsha, Adam Klivans, and Raghu Meka. Bounding the sensitivity of polynomial threshold functions, 2009. arXiv: 0909.5175.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM J. Computing*, 37(6):1777–1805, 2008. doi: 10.1137/060649057.

Daniel Kane. A structure theorem for poorly anticoncentrated gaussian chaoses and applications to the study of polynomial threshold functions. In *FOCS*, 2012.

Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

L. B. Klebanov and S. T. Mkrtchyan. Estimation of the closeness of distributions in terms of identical moments. *Journal of Mathematical Sciences*, 32:54–60, 1986. ISSN 1072-3374. URL http://dx.doi.org/10.1007/BF01084500.

L. B. Klebanov and S. T. Rachev. Proximity of probability measures with common marginals in a finite number of directions. In *Distributions with fixed marginals and related topics*, volume 28 of *Lecture Notes – Monograph Series*, pages 162–174. Institute of Mathematical Statistics, 1996.

Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. *J. Computer and System Sciences*, 68(4):808–840, 2004. doi: 10.1016/j.jcss.2003.11.002.

Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via Gaussian surface area. In *FOCS*, pages 541–550, 2008. doi: 10.1109/FOCS.2008.64.

Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, 1993.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer Verlag, 1991.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993. doi: 10.1145/174130.174138.

L. Lovász and S. Vempala. Logconcave functions: Geometry and efficient sampling algorithms. In *FOCS*, pages 650–659, 2003.

Doron Lubinsky. A survey of weighted polynomial approximation with exponential weights. *Surveys in Approximation Theory*, 2007.

Raghu Meka and David Zuckerman. Pseudorandom generators for polynomial threshold functions. In *STOC*, pages 427–436, 2010.

Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low in.uences invariance and optimality. In *FOCS*, pages 21–30, 2005. doi: 10.1109/SFCS.2005.53.

Iosif Pinelis. Extremal probabilistic problems and hotellings T^2 test under a symmetry condition. *Ann. Statist.*, 22(1):357–368, 1994. doi: 10.1214/aos/1176325373.

Santosh Vempala. Learning convex concepts from gaussian distributions with PCA. In *FOCS*, pages 124–130, 2010a. ISBN 978-0-7695-4244-7. URL http://dx.doi.org/10.1109/FOCS.2010.19.

Santosh Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *J. ACM*, 57(6):32, 2010b. URL http://doi.acm.org/10.1145/1857914.1857916.

Karl Wimmer. Agnostically learning under permutation invariant distributions. In *FOCS*, pages 113–122, 2010.

Appendix A. Deferred Proofs from The Convolution Proof

A.1. Proof of Lemma 11

Proof By Lemma 10, for each i there exists a polynomial p_i of degree less than $2^i k$ so that

$$|f(x) - p_i(x)| \leq \frac{(c|x|)^{2^i k}}{(2^i k)!}.$$

We let $p(x) = p_0(x) \prod_{i=0}^{n-2} p_i(x)$. This is clearly a polynomial of appropriate degree, we have left to show that it is an appropriate approximation of f^N. Letting $y = c|x|$, we have that

$$p(x) = \left(f(x) \pm \frac{y^k}{k!}\right) \prod_{i=0}^{N-2} \left(f(x) \pm \frac{y^{2^i k}}{(2^i k)!}\right).$$

Expanding out the right hand side above we find that $p(x)$ is equal to $(f(x))^N$ plus a sum of $2^N - 1$ other terms. If $y \leq (k!)^{1/k}$, then each of these terms is at most $\frac{y^k}{k!}$ (since $|f(x)|, \frac{y^{2^i k}}{(2^i k)!} \leq 1$). Thus in this range of y, we have that $|(f(x))^N - p(x)| \leq 2^N \frac{y^k}{k!}$. Otherwise, let j be the largest integer with $j \leq n-2$ so that $\frac{y^{2^j k}}{(2^j k)!} > 1$. It should be noted that for $i < j$ that $\frac{y^{2^i k}}{(2^i k)!} > 1$. Therefore, in this case, we have that

$$\begin{aligned}
|p(x)| &= \left|f(x) \pm \frac{y^k}{k!}\right| \cdot \prod_{i=0}^{N-2} \left|f(x) \pm \frac{y^{2^i k}}{(2^i k)!}\right| \\
&\leq \left(\frac{2y^k}{k!}\right) \prod_{i=0}^{j} \left(\frac{2y^{2^i k}}{(2^i k)!}\right) \prod_{i=j+1}^{N-2} 2 \\
&= 2^N \frac{y^{2^{j+1} k}}{(2^{j+1} k)!} \prod_{i=0}^{j} \binom{2^{i+1} k}{2^i k} \\
&\leq 2^N \frac{y^{2^{j+1} k}}{(2^{j+1} k)!} \prod_{i=0}^{j} 2^{2^{i+1} k} \leq 2^N \frac{(2y)^{2^{j+1} k}}{(2^{j+1} k)!}.
\end{aligned}$$

In any of the above cases, we have that $|(f(x))^N - p(x)|$ is bounded by the desired bound. ■

A.2. Proof of Lemma 12

Proof We use the polynomial p given to us in Lemma 11. We have that

$$|(f(x))^N - p(x)| \leq 2^N \sum_{j=0}^{N-1} \int_{\mathbb{R}^m} \left(\frac{(2c|x|)^{2^j k}}{(2^j k)!} \right) d\mu_{\mathcal{D}}(x).$$

This is equal to

$$2^N \sum_{j=0}^{n-1} \int_0^\infty \frac{\partial}{\partial R} \left(\frac{(2cR)^{2^j k}}{(2^j k)!} \right) \Pr_{X \sim \mathcal{D}}(|X| \geq R) dR.$$

By Corollary 9 this is at most

$$2^N 2^{O(m)} \sum_{j=0}^{N-1} \int_0^\infty 2^j k \frac{(2cR)^{2^j k}}{(2^j k)!} e^{-R/32} \frac{dR}{R}.$$

Substituting $S = R/32$, we get

$$2^N 2^{O(m)} \sum_{j=0}^{N-1} 2^j k \int_0^\infty \frac{(128cS)^{2^j k}}{(2^j k)!} e^{-S} \frac{dS}{S}.$$

Recalling the standard Γ-integral, this is at most

$$2^N 2^{O(m)} \sum_{j=0}^{N-1} 2^j k (128c)^{2^j k}.$$

If $c < 1/256$, this is clearly at most $2^N 2^{O(m)} 2^{-k}$, as desired. ■

A.3. Proof of Theorem 6

Here we give the Proof of Theorem 6. The basic idea is as follows. Let ρ be the normalized N^{th} power of a c-smooth function. By Corollary 12, any translation of ρ can be approximated in $L^1_{\mathcal{D}}$ by bounded degree polynomials. Therefore, the convolution of ρ with any function of bounded L^1 norm can be approximated by a polynomial. The polynomial that we use will be such an approximation to the convolution of ρ with a truncation of f to a moderately-sized ball about the origin. Since ρ is a large power of another function, it forms a reasonable approximation of the δ-function. Thus, the convolution in question will be a good approximation to f except for points outside the truncation (a region with small total probability) and points near the locus of discontinuity (i.e. those in S_δ). An appropriate setting of parameters in this construction will produce the desired bound.

Proof [Proof of Theorem 6] We assume throughout that ε and δ are sufficiently small. Let C be a sufficiently large constant.

Let $N \geq \max(m^4 256^{-4}\delta^{-4}, \log(C256^{-2m}m^{3m}\varepsilon^{-1})^2)$, and $k \geq 2N + 4Cm + m\log_2(N) + 2m\log_2(\log(\varepsilon^{-1}))$ be integers. Let

$$\rho(x) := \frac{\rho_{1/256}^N(x)}{\int_{\mathbb{R}^m} \rho_{1/256}^N(y)dy}.$$

Clearly,

$$\int_{\mathbb{R}^m} \rho(y)dy = 1.$$

Furthermore, by the above we have that

$$\int_{|x|>\delta} \rho(y)dy \leq \varepsilon/16.$$

We have that $\int_{\mathbb{R}^m} \rho_{1/256}^N(y)dy = \Omega(256^{-m}N^{-m/2})$. By Corollary 12, for any $y \in \mathbb{R}^m$, there exists a polynomial $p_y(x)$ of degree at most $k2^N$ so that

$$|p_y(x) - \rho(x+y)|_1 \leq 2^{-k}2^{O(m)}N^{m/2} \leq \varepsilon 2^{-Cm}\log(\varepsilon^{-1})^{-m}$$

Next let $P(x)$ be the polynomial given by

$$P(x) := \int_{|y|\leq 64\log(2^{Cm}/\varepsilon)+1} p_{-y}(x)f(y)dy.$$

We note that the L^1 error between $P(x)$ and

$$q(x) := \int_{|y|\leq 64\log(2^{Cm}/\varepsilon)+1} \rho(x-y)f(y)dy$$

is at most $\varepsilon/4$. Now $q(x)$ is the convolution of ρ with the restriction of f to a ball of radius $64\log(2^{Cm}/\varepsilon)+1$. This means that inside the ball of radius $64\log(2^{Cm}/\varepsilon)$ that $q(x)$ agrees with $\rho * f$ to within $\varepsilon/8$. Off of this ball, $|q(x) - (\rho * f)(x)| \leq 2$, but by Corollary 9, the probability that x lies in this range is at most $(\varepsilon/16)$ if C is sufficiently large. Thus, we have that

$$|P(x) - (\rho * f)(x)|_1 \leq 3\varepsilon/8.$$

Now if $x \notin S_\delta$, then $\rho * f$ is

$$\begin{aligned}
&\int_{|x-y|<\delta} \rho(x-y)f(y)dy + \int_{|x-y|>\delta} \rho(x-y)f(y)dy \\
=&f(x)\int_{|x-y|<\delta} \rho(x-y)dy \pm \int_{|x-y|>\delta} \rho(x-y)dy \\
=&f(x) \pm \varepsilon/4.
\end{aligned}$$

Thus $|P(x) - f(x)|_1$ is at most

$$\begin{aligned}
&|P(x) - (\rho * f)(x)|_1 + |(\rho * f)(x) - f(x)|_1 \\
\leq&3\varepsilon/8 + 2\int_{\mathbb{R}^k \setminus S_\delta} (\varepsilon/4)d\mu_{\mathcal{D}}(x) + \int_{S_\delta} 2d\mu_{\mathcal{D}}(x) \\
\leq&\varepsilon + 2\mu_{\mathcal{D}}(S_\delta).
\end{aligned}$$

■

Appendix B. Deferred Proofs from Moment-Matching Proof

We now prove Lemma 23. We shall use the following result from Gabovich (1981).

Theorem 24 (Gabovich (1981) Equation (8)) *Let $X, Y \in \mathbb{R}^m$ be two vector-valued random variables. Then, for a universal constant C and all sufficiently large $N, T > 0$,*

$$\mathsf{d}_{\mathsf{LV}}(X,Y) \leq \int_{\frac{1}{NT} \leq t_1,\ldots,t_m \leq T} \frac{C|\varphi_X((t_1,\ldots,t_m)) - \varphi_Y(t_1,\ldots,t_m)|}{t_1 t_2 \cdots t_m} dt_1 \cdots dt_m + \frac{C(\log T)(\log(NT))}{T} + \mathsf{Pr}[X \notin [-N,N]^m] + \mathsf{Pr}[Y \notin [-N,N]^m].$$

Proof [Proof of Lemma 23] Without loss of generality suppose that $\delta < 1/m^2$, as else the statement is trivial. Let T^* be the value of T that attains the minimum in the definition of d_λ:

$$\mathsf{d}_\lambda(X,Y) = \max\{\max_{\|t\| \leq T^*} \{|\varphi_X(t) - \varphi_Y(t)|\}, 1/T^*\}.$$

As $\mathsf{d}_\lambda(X,Y) \leq \delta$, $T^* \geq 1/\delta$. Therefore, for every $t \in \mathbb{R}^m$ with $\|t\| \leq 1/\delta$, $|\varphi_X(t) - \varphi_Y(t)| \leq \delta$. Thus, applying Theorem 24 with $N = N(\delta)$ and $T = 1/\delta\sqrt{m}$, we get

$$\begin{aligned}
\mathsf{d}_{\mathsf{LV}}(X,Y) &\leq C \int_{\frac{1}{NT} \leq t_1,\ldots,t_m \leq T} \frac{|\varphi_X(t) - \varphi_Y(t)|}{t_1 \cdots t_m} dt + O(\log^2(NT) \cdot \delta\sqrt{m}) + O(\delta) \\
&\leq C \int_{\frac{1}{NT} \leq t_1,\ldots,t_m \leq T} \frac{\delta}{t_1 \cdots t_m} dt + O(\log^2(NT) \cdot \delta\sqrt{m}) \\
&\leq C\delta \cdot (\log N + 2\log T)^m + O(\log^2(NT) \cdot \delta\sqrt{m}) \\
&= O((\log N(\delta) + 2\log(1/\delta))^m \cdot \delta).
\end{aligned}$$

■

B.1. Proof of Lemma 19

Proof Fix $\theta_1, \ldots, \theta_m$ and let $X' = (\mathsf{sign}(X_1 - \theta_1), \ldots, \mathsf{sign}(X_m - \theta_m))$ and define Y' similarly. Then, from the assumptions of the lemma, for every $a \in \{1,-1\}^m$,

$$|\,\mathsf{Pr}[X' = a] - \mathsf{Pr}[Y' = a]| < \mathsf{d}_{\mathsf{cdf}}((a_1X_1, a_2X_2, \ldots, a_mX_m), (a_1Y_1, a_2Y_2, \ldots, a_mY_m)) < \varepsilon.$$

Therefore, $\mathsf{d}_{\mathsf{TV}}(X', Y') < 2^{m-1}\varepsilon$. The lemma now follows. ■

Appendix C. Smoothed Complexity of Learning Functions of Halfspaces

We now consider the smoothed complexity of learning convex sets defined by intersections of halfspaces and extend our learning results to handle any distribution whose marginals obey a

subexponential tail bound. We feel this is a mild restriction to place on the distribution. It is well known that any (isotropic) log-concave distribution obeys such a tail bound.

Our high level approach will be similar to that for log-concave densities: we use moment bounds and results from Section 4.3 to show that functions of halfspaces cannot distinguish (smoothed) distributions with strong moment bounds. Adding a Gaussian perturbation plays an important role in our setting, by essentially allowing us to impose certain *probabilistic* margin constraints in the form of anti-concentration bounds. One interpretation of our results is that in the setting of smoothed-analysis, learning geometric classes becomes easier in many cases because the underlying Gaussian perturbation makes the distribution anti-concentrated (i.e., no sharp peaks) "for free."

We state our results below and defer the proofs (and definitions of sub-exponential, sub-gaussian densities) to the full version.

Theorem 25 *Let $\mathcal{D}$ be a sub-exponential distribution over $\mathbb{R}^n$. Let $h_1, \ldots, h_m : \mathbb{R}^n \to \{1, -1\}$, be halfspaces and let $g : \{1, -1\}^m \to \{1, -1\}$ be an arbitrary function. Define $f : \mathbb{R}^n \to \{1, -1\}$ by $f(x) = g((h_1(x), \ldots, h_m(x)))$. Then, for every $\sigma > 0$, there exists a polynomial P of degree at most*

$$k = \exp((\log((\log m)/\sigma\varepsilon))^{O(m)}/(\sigma\varepsilon)^4)$$

such that $\mathbb{E}_{X \leftarrow \mathcal{D}(\sigma)}[\,|\, f(X) - P(X)\,|\,] \leq \varepsilon$.

Theorem 26 *Let $\mathcal{D}$ be a sub-Gaussian distribution over $\mathbb{R}^n$. Let $h_1, \ldots, h_m : \mathbb{R}^n \to \{1, -1\}$, be halfspaces and let $g : \{1, -1\}^m \to \{1, -1\}$ be an arbitrary function. Define $f : \mathbb{R}^n \to \{1, -1\}$ by $f(x) = g((h_1(x), \ldots, h_m(x)))$. Then, for every $\sigma > 0$, there exists a real-valued polynomial P of degree at most $k = (\log((\log m)/\sigma\varepsilon))^{O(m)}/(\sigma\varepsilon)^4$ such that $\mathbb{E}_{X \leftarrow \mathcal{D}(\sigma)}[\,|\, f(X) - P(X)\,|\,] \leq \varepsilon$.*

Theorem 3 and the precise runtimes as given in Table 1.1 follow from the above results and Theorem 5.

Appendix D. Sub-Exponential Densities

In this section we study sub-exponential densities and prove Theorem 25.

Definition 27 *We say an isotropic distribution $\mathcal{D}$ on $\mathbb{R}^n$ is sub-exponential if there exist constants $C, \alpha > 0$, such that for every $w \in \mathbb{R}^n$, $\|w\| = 1$, and $t > 0$,*

$$\Pr_{X \leftarrow \mathcal{D}}[|\langle w, X\rangle| > t] < C\exp(-\alpha t).$$

More generally, we say a distribution $\mathcal{D}$ on $\mathbb{R}^n$ is sub-exponential if the isotropic distribution obtained by putting $\mathcal{D}$ in an isotropic position by an affine transformation is sub-exponential.

We shall use the following standard fact giving strong moment bounds for random variables with sub-exponential tails.

Fact 28 *Let X be unit variance random variable such that $\Pr[|X| > t] < C\exp(-\alpha t)$. Then, for all $k > 0$, $\mathbb{E}[|X|^k] < C(k/\alpha)^k$.*

Finally, we need the following fact showing that convolving any distribution with a Gaussian distribution leads to *anti-concentration*.

Fact 29 *For any real-valued random variable X, $Z \leftarrow \mathcal{N}(0,\sigma)$ and $t \in \mathbb{R}$, $\alpha > 0$, $\mathsf{Pr}[X+Z \in [t,t+\alpha)] < C\alpha/\sigma$, where C is a universal constant.*

Proof Fix $t \in \mathbb{R}$ and $\alpha > 0$. Then,

$$\mathsf{Pr}[Z \in [t,t+\alpha)] = \int_t^{t+\alpha} \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dx < \int_t^{t+\alpha} \frac{1}{\sqrt{2\pi\sigma^2}} dx = C\alpha/\sigma,$$

where $C = 1/\sqrt{2\pi}$. The claim now follows from:

$$\mathsf{Pr}[X+Z \in [t,t+\alpha)] = \mathop{\mathbb{E}}_X[\mathop{\mathsf{Pr}}_Z[Z \in [(t-X),(t-X)+\alpha)]\,] < \mathop{\mathbb{E}}_X[C\alpha/\sigma] < C\alpha/\sigma.$$

■

Proof [Proof of Theorem 25] The proof follows the same approach as that of Theorem 15. Without loss of generality, we can suppose that $\mathcal{D}$ is in isotropic position as functions of halfspaces are closed under affine transformations.

Let the Gaussian perturbation be $Z \leftarrow \mathcal{N}(0,\Sigma)^m$, where $\Sigma \succeq \sigma \mathsf{I}_m$. We next renormalize the distribution $\mathcal{D}$ so that $\mathcal{D}(\sigma)$ is in isotropic position. Note that $\mathcal{D}(\sigma)$ is also sub-exponential. This follows from a simple union bound. For any direction $w \in \mathbb{R}^m$, $\|w\| = 1$, and $X \leftarrow \mathcal{D}$ and $Z \leftarrow \mathcal{N}(0,\Sigma)^m$,

$$\mathsf{Pr}[|\langle X+Z, w\rangle| > t] \le \mathsf{Pr}[|\langle X, w\rangle| > t/2] + \mathsf{Pr}[|\langle Z, w\rangle| > t/2] = O(\exp(-\Omega(t))),$$

where the last inequality follows from the fact that X is sub-exponential by definition and that the uni-variate Gaussian distribution is sub-exponential.

Fix halfspaces $h_i : \mathbb{R}^n \to \{1,-1\}$ and let random variables $X \leftarrow \mathcal{D}(\sigma)$ and let $X' \leftarrow \mathcal{D}'$, where $\mathcal{D}'$ k-moment matches $\mathcal{D}$ for k to be chosen later. Let Y, Y' be as in Theorem 15. Then, by Fact 28, for any $w \in \mathbb{R}^n$, $\|w\| = 1$, $\mathbb{E}[|\langle w, X\rangle|^j] < C(j/\alpha)^j$.

Observe that the proofs of Equations (3) and (4) in Theorem 15 only used moment bounds for log-concave distributions, and sub-exponential distributions have similar bounds on moments. Thus, by similar arguments, for $k = 2^{O(m/\delta^4)}$ sufficiently large, we get

$$\mu_j(Y) \le C(jm/\alpha)^j, \quad \beta_k(Y) = \Omega((\log k)/m), \tag{7}$$

and for $N = O(\log(m/\delta)/\alpha)$ sufficiently large,

$$\mathsf{Pr}[Y \notin [-N,N]^m] + \mathsf{Pr}[Y' \notin [-N,N]^m] < 2\delta.$$

Combining the above two equations and Lemma 23, we have

$$\mathsf{d}_\lambda(Y,Y') = (\log(\log(m/\delta)))^{O(m)} \cdot \delta. \tag{8}$$

Now, note that for any $r \in [m]$, Y_r can be written as $Y'_r + Z_r$, where $Z_r \leftarrow \mathcal{N}(0,\sigma)$ is independent of Y'_r. Therefore, by Fact 29, $\mathsf{Pr}[Y_r \in [t,t+\gamma]] = O(\gamma/\sigma)$ for $t \in \mathbb{R}, \alpha > 0$. Thus, by the above equation and Fact 18,

$$\mathsf{d}_{\mathsf{cdf}}(Y,Y') = O(m\mathsf{d}_\lambda(Y,Y')/\sigma) = (\log(\log(m/\delta)))^{O(m)} \cdot \delta/\sigma = \varepsilon.$$

The theorem now follows from an argument similar to that of Theorem 15 following Equation 6. ■

Appendix E. Sub-Gaussian Densities

We now study sub-Gaussian densities and show an analogue of Theorem 25 with much better parameters. The improvement in parameters comes from the fact that sub-Gaussian have much more tightly controlled moments.

Definition 30 *We say an isotropic distribution $\mathcal{D}$ is sub-Gaussian if there exist constants $C, \alpha > 0$, such that for every $w \in \mathbb{R}^n$, $\|w\| = 1$, and $t \in \mathbb{R}$,*

$$\Pr_{X \leftarrow \mathcal{D}}[|\langle w, X\rangle| > t] < C\exp(-\alpha t^2).$$

More generally, we say a distribution $\mathcal{D}$ on $\mathbb{R}^n$ is sub-exponential if the isotropic distribution obtained by putting $\mathcal{D}$ in an isotropic position by an affine transformation is sub-exponential.

Analogous to Fact 28, we have the following statement for sub-gaussian densities.

Fact 31 *Let X be unit variance random variable such that $\Pr[|X| > t] < C\exp(-\alpha t)$. Then, $\mathbb{E}[|X|^k] < C(k/\alpha^2)^{k/2}$.*

Proof [Proof of Theorem 26] The proof follows the same approach as that of Theorem 25. We only highlight the important differences. Fix halfspaces $h_i : \mathbb{R}^n \to \{1, -1\}$, and random variables X, X', Y, Y' as in the proof of Theorem 25. Now, observe that for $k = \Omega(m/\delta^4)$, sufficiently large, for any $t \in \mathbb{R}^m$, $\|t\| = 1$, and $j > 0$,

$$\begin{aligned}
\mathbb{E}[|\langle t, Y\rangle|^j] &= \mathbb{E}[\,|\langle \sum_{r=1}^{m} t_r w_r, X\rangle|^j\,] \\
&\le C j^{j/2} \cdot \mathbb{E}[\langle \sum_{r=1}^{m} t_r w_r, X\rangle^2]^{j/2}/\alpha^j \qquad \text{(Fact 31)} \\
&= O((m/\alpha)^j \cdot j^{j/2}).
\end{aligned}$$

Therefore,

$$\beta_k(Y) = \sum_{j=1}^{k} \frac{1}{\mu_{2j}(Y)^{1/2j}} \ge \sum_{j=1}^{k} \frac{\alpha}{m\sqrt{2j}} = \Omega(\sqrt{k}/m). \tag{9}$$

Note that the above bound on β_k is exponentially better than the $\Omega(\log k)$ bound we had for log-concave and sub-exponential densities and this leads to the quantitative improvements for sub-Gaussian densities.

Now, by using Markov's inequality it follows that for $k > \log(m/\delta)$, and $N = O(\sqrt{\log(m/\delta)/\alpha})$ sufficiently large,

$$\Pr[Y \notin [-N, N]^m] + \Pr[Y' \notin [-N, N]^m] < 2\delta.$$

Combining the above two equations and Lemma 23, we get

$$\mathsf{d}_\lambda(Y, Y') = (\log(\log(m/\delta)))^{O(m)} \cdot \delta.$$

The theorem now follows from the above inequality and an argument similar to that of Theorem 25 following Equation 8. ■

Appendix F. Non-Product Distributions on Hypercube

Learning intersections of halfspaces with respect to distributions on the hypercube is a long-standing and fundamental open problem in learning theory. To date, most non-trivial results pertain to product distributons on the hypercube, with the exception of the work of Wimmer Wimmer (2010) who can handle symmetric distributions on the hypercube.

Our results imply algorithms for agnostically learning functions of halfspces in the smoothed complexity setting for distributions on the hypercube that are locally-independent. Specifically, call a distribution $\mathcal{D}$ on $\{1,-1\}^n$ k-wise independent if for any $I \subseteq [n], |I| \leq k$, $X \leftarrow \mathcal{D}$, the variables $(X_i : i \in I)$ are independent. (This is the same as saying $\mathcal{D}$ k-moment matches the uniform distribution on $\{1,-1\}^n$). Our learning algorithms for sub-Gaussian densities, Theorem 26, immediately imply the following for learning with respect to k-wise independent distributions.

Theorem 32 *For all m, ε, σ there exists $k = O_{m,\varepsilon,\sigma}(1)$ such that the following holds. Functions of m halfspaces are agnostically learnable with respect to any k-wise independent distribution on $\{1,-1\}^n$ in time $n^{O_{m,\varepsilon,\sigma}(1)}$ where ε is the accuracy parameter and σ is the perturbation parameter.*

In contrast, it is not clear if any of the previous techniques can give algorithms for learning intersections of halfspaces that are even $\Omega(n)$-wise independent.

Proof The uniform distribution on $\{1,-1\}^n$ is known to be sub-Gaussian Pinelis (1994). Further, observe that in the proof of Theorem 26 we only used properties of the first k-moments for $k = (\log((\log m)/\sigma\varepsilon))^{O(m)}/(\sigma\varepsilon)^4$. Thus, the same arguments should work for any distribution $\mathcal{D}$ which is k-wise independent. The thoerem then follows from combining the direct analogue of Theorem 26 for k-wise independent distributions $\mathcal{D}$ with Theorem 5. ■

Appendix G. Bounded Independence Fools Degree Two Threshold Functions

Here we show that the methods of Section 4 can also be used with respect to the uniform distribution over $\{1,-1\}^n$. We use the moment-matching techniques to give a new proof for the recent result of Diakonikolas, Kane, and Nelson Diakonikolas et al. (2010a) that bounded independence fools degree-2 polynomial threshold functions. Our proof gives worse parameters, but is considerably different and is perhaps simpler. We also establish a connection between the pseudorandomness problem and the well studied classical moment problem in probability (see Akhiezer (1965) for instance).

Theorem 33 *There exist constants C, C' such that the following holds. Let $\mathcal{D}$ be a m-wise independent distribution over $\{1,-1\}^n$ for $m = 2^{C/\delta^9}$. Then, for every degree 2 polynomial $P : \mathbb{R}^n \to \mathbb{R}$, and $x \leftarrow \mathcal{D}$, $y \in_u \{1,-1\}^n$, $\mathsf{d}_{\mathsf{cdf}}(P(x), P(y)) < C'\delta$. In other words, $(2^{O(1/\delta^9)})$-wise independence δ-fools degree two threshold functions.*

In comparison, Diakonikolas et al. show that $\tilde{O}(\delta^{-9})$-wise independence suffices. This bound was later improved to $O(\delta^{-8})$ in Kane (2012).

We shall use the following quantitative estimate due to Klebanov and Mkrtchyan which can be seen as a one dimensional version of Theorem 22, albeit with better parameters.

Theorem 34 (Theorem 1, Klebanov and Mkrtchyan (1986)) *Let X, Y be real-valued random variables with $\mathbb{E}[X^i] = \mathbb{E}[Y^i]$ for $1 \leq i \leq 2m$ and $\mathbb{E}[X^2] = 1$. Then, for a universal constant $C > 0$,*

$$\mathsf{d}_{\mathsf{LV}}(X, Y) \leq \frac{C_\sigma \cdot \ln(1 + \beta_m(X))}{\beta_m(X)^{1/4}}.$$

We only detail the case of *regular* polynomials here, the reduction from the general case to the regular case works via the regularity lemma of Harsha et al., Harsha et al. (2009) and Diakonikolas et al., Diakonikolas et al. (2010b).

Definition 35 *A multi-linear polynomial $P : \mathbb{R}^n \to \mathbb{R}$, $P(x) = \sum_{I \subseteq [n]} a_I \prod_{i \in I} x_i$ is δ-regular if for every $i \in [n]$,*

$$\sum_{i=1}^{n} \left(\sum_{I \subseteq [n], I \ni i} a_I^2 \right)^2 \leq \delta^2 \|P\|_2^4,$$

where $\|P\|_2^2 = \sum_I a_I^2$.

Theorem 36 *There exist constants C, C' such that the following holds. Let $\mathcal{D}$ be a m-wise independent distribution over $\{1, -1\}^n$ for $m = 2^{C/\delta^2}$. Then, for every δ-regular degree 2 polynomial $P : \mathbb{R}^n \to \mathbb{R}$, and $x \leftarrow \mathcal{D}$, $y \in_u \{1, -1\}^n$, $\mathsf{d}_{\mathsf{cdf}}(P(x), P(y)) < C'\delta^{2/9}$.*

Theorem 33 follows from the above theorem and the regularity lemma of Harsha et al., Diakonikolas et al. We refer the reader to the work of Meka and Zuckerman Meka and Zuckerman (2010) for a similar reduction of the general case to the regular case in the pseudorandomness context and omit it here.

To prove Theorem 36 we use the following results about low-degree polynomials. The lemma gives us control on how fast the moments of low-degree polynomials grow.

Theorem 37 (Hypercontractivity, Ledoux and Talagrand (1991)) *For $1 < p < q < \infty$, and $P : \mathbb{R}^n \to \mathbb{R}$ a degree d polynomial, the following holds:*

$$\mathop{\mathbb{E}}_{X \in_u \{1,-1\}^n} [|P(X)|^q]^{1/q} \leq \left(\frac{q-1}{p-1} \right)^{d/2} \mathop{\mathbb{E}}_{X \in_u \{1,-1\}^n} [|P(X)|^p]^{1/p}.$$

The next two theorems helps us get anti-concentration bounds for regular polynomials over the hypercube.

Theorem 38 (Mossel et al. Mossel et al. (2005)) *There exists a universal constant C such that the following holds. Let $P : \mathbb{R}^n \to \mathbb{R}$ be a degree d δ-regular (multi-linear) polynomial. Then, for $x \in_u \{1, -1\}^n$ and $y \leftarrow \mathcal{N}(0, 1)^n$,*

$$\mathsf{d}_{\mathsf{cdf}}(P(x), P(y)) \leq C\, d\, \delta^{2/(4d+1)}.$$

Theorem 39 (Carbery and Wright Carbery and Wright (2001)) *There exists an absolute constant C such that for any polynomial Q of degree at most d with $\|Q\| = 1$ and any interval $I \subseteq \mathbb{R}$ of length α, $\mathsf{Pr}_{X \leftarrow \mathcal{N}(0,1)^n}[Q(X) \in I] \leq Cd\, \alpha^{1/d}$.*

Proof [Proof of Corollary 36] It suffices to show the statement when $\mathcal{D}$ is $4m$-wise independent for $m = 2^{C/\delta^2}$ for C to be chosen later. Without loss of generality suppose that $\|P\| = 1$. Let random variables $X = P(x)$, for $x \leftarrow \mathcal{D}$ and $Y = P(y)$, for $y \in_u \{1, -1\}^n$. Then, $\mathbb{E}[X^i] = \mathbb{E}[Y^i]$ for $i \leq 2m$ as x is $4m$-wise independent and P is a degree 2 polynomial. Now, for $i \leq m$, by hypercontractivity, Theorem 37, applied to $q = i$, $d = 2$,

$$\mathbb{E}[X^{2i}] = \mathbb{E}[Y^{2i}] < (2i)^{2i}.$$

Therefore,

$$\beta_m = \sum_{i=1}^{m} \frac{1}{\mathbb{E}[X^{2i}]^{1/2i}} > \sum_{i=1}^{m} \frac{1}{2i} = \Omega(\log m).$$

By Theorem 34,

$$\mathsf{d}_{\mathsf{LV}}(X, Y) = O\left(\frac{\log\log m}{(\log m)^{1/4}}\right).$$

Now, by Theorem 38 and Theorem 39 applied to degree $d = 2$, $\sup_t \mathsf{Pr}[Y \in [t, t + \alpha]] = O(\delta^{2/9} + \sqrt{\alpha})$. Therefore, by Fact 18

$$\mathsf{d}_{\mathsf{cdf}}(X, Y) = O\left(\delta^{2/9} + \frac{\sqrt{\log\log m}}{(\log m)^{1/8}}\right).$$

The statement now follows by choosing C to be sufficiently large. ■

Acknowledgments

Adam Klivans acknowledges support from an NSF CAREER award and NSF grant AF 1018829.