Weighted Polynomial Approximations: Limits for Learning and Pseudorandomness

Mark Bun[*] and Thomas Steinke[†]

Harvard University, School of Engineering and Applied Sciences, USA
{mbun,tsteinke,}@seas.harvard.edu

Abstract

Low-degree polynomial approximations to the sign function underlie pseudorandom generators for halfspaces, as well as algorithms for agnostically learning halfspaces. We study the limits of these constructions by proving inapproximability results for the sign function. First, we investigate the derandomization of Chernoff-type concentration inequalities. Schmidt et al. (SIAM J. Discrete Math. 1995) showed that a tail bound of δ can be established for sums of Bernoulli random variables with only $O(\log(1/\delta))$-wise independence. We show that their results are tight up to constant factors. Secondly, the "polynomial regression" algorithm of Kalai et al. (SIAM J. Comput. 2008) shows that halfspaces can be efficiently learned with respect to log-concave distributions on $\mathbb{R}^n$ in the challenging agnostic learning model. The power of this algorithm relies on the fact that under log-concave distributions, halfspaces can be approximated arbitrarily well by low-degree polynomials. In contrast, we exhibit a large class of non-log-concave distributions under which polynomials of any degree cannot approximate the sign function to within arbitrarily low error.

1998 ACM Subject Classification F.2.0 Analysis of Algorithms and Problem Complexity – General

Keywords and phrases Polynomial Approximations, Pseudorandomness, Concentration, Learning Theory, Halfspaces

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2015.625

1 Introduction

Approximation theory is a classical area of mathematics that studies how well functions can be approximated by simpler ones. It has found many applications in computer science. Most of these applications of approximation theory focus on the approximation of functions by polynomials in the uniform norm (or infinity norm). For instance, *approximate degree*, which captures how well a boolean function can be approximated by low-degree polynomials in the uniform norm, underlies important lower bounds in circuit complexity [6, 7, 66], quantum query complexity [5, 1], and communication complexity [65]. It also underlies state-of-the art algorithms in learning theory [35, 41], streaming [31], and in spectral methods [63].

While it is compelling to study polynomial approximations under the uniform norm, there are scenarios where it is more natural to study *weighted polynomial approximations*, where error is measured in terms of an L_p norm under some distribution. For instance, in agnostic learning, the polynomial regression algorithm of Kalai et al. [35] has guarantees based on how well functions in a concept class of interest can be approximated by low-degree polynomials in L_1 distance.

[*] Supported by an NDSEG Fellowship and NSF grant CNS-1237235.
[†] Supported by NSF grant CCF-1116616 and the Lord Rutherford Memorial Research Fellowship.

18th Int'l Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'15) /
19th Int'l Workshop on Randomization and Computation (RANDOM'15).
Editors: Naveen Garg, Klaus Jansen, Anup Rao, and José D. P. Rolim; pp. 625–644
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In this work, we show how ideas from weighted approximation theory can yield tight lower bounds for several problems in theoretical computer science. As our first application, in the area of derandomization, we give a tight characterization of the amount of k-wise independence necessary to establish Chernoff-like concentration inequalities. Second, we establish a strong limitation on the distributions under which halfspaces can be learned using the polynomial regression algorithm of Kalai et al.

1.1 Tail Bounds for Limited Independence

The famous Hoeffding bound [32] implies that if $X \in \{\pm 1\}^n$ is a uniform random variable and $r \in \mathbb{R}^n$ is fixed, then, for all $T \geq 0$,

$$\underset{X}{\mathbb{P}}\left[|X \cdot r| \geq T\right] \leq 2e^{-\frac{T^2}{2||r||_2^2}}.$$

We ask the following question:

For what *pseudorandom* X is the Hoeffding bound true?

More precisely, given T and δ, can we construct a pseudorandom $X \in \{\pm 1\}^n$ such that $\underset{X}{\mathbb{P}}\left[|X \cdot r| \geq T\right] \leq \delta$ for all $r \in \{\pm 1\}^n$?[1] Of particular interest is the parameter regime $\delta = 1/\operatorname{poly}(n)$ and $T = \Theta(||r||_2 \sqrt{\log(1/\delta)})$, which is natural in the context of derandomizing efficient randomized algorithms.[2] The probabilistic method gives a non-constructive proof that there exists such an X which can be sampled with seed length $O(\log(n/\delta))$. The challenge is to give an explicit construction of such an X which can be *efficiently* sampled with a short seed.

This is a very natural pseudorandomness question: Concentration of measure is a fundamental property of independent random variables and one of the key objectives of pseudorandomness research is to replicate such properties for random variables with low entropy. Finding a pseudorandom X exhibiting good concentration is also a relaxation of a more general and well-studied pseudorandomness question, namely that of constructing pseudorandom generators that fool linear threshold functions [21, 48, 28, 22]. This problem can also be viewed as a special case of constructing pseudorandom generators for space-bounded computation [57, 33, 60, 14, 15, 43, 61].

For $\delta = 1/\operatorname{poly}(n)$ and $T = \Theta(\sqrt{n \log(1/\delta)})$, we can construct explicit X that can be sampled with seed length $O(\log^2 n)$ using a variety of methods (including [57, 48]). In particular, it suffices for X to be $O(\log(1/\delta))$-wise independent:

▶ **Theorem 1** (Tail Bound for Limited Independence). *Let $n \geq 1$, $\eta > 0$, and $\delta \in (0,1)$ be given. Let $X \in \{\pm 1\}^n$ be k-wise independent for $k = 2\lceil \eta \log_e(1/\delta)\rceil$. Let $r \in \mathbb{R}^n$ and set $T = e^{(\eta+1)/2\eta}\sqrt{k}\,||r||_2$. Then*

$$\mathbb{P}\left[|X \cdot r| \geq T\right] \leq \delta.$$

Limited independence is a very general and intuitively appealing technique in pseudorandomness. As a tool for derandomization, it has been studied in the contexts of hashing [46, 50], dimensionality reduction [37], random graphs [3], and circuits [4, 13]. A k-wise

[1] For simplicity we restrict our attention to $r \in \{\pm 1\}^n$, instead of arbitrary real-valued r.

[2] Smaller values of δ are also interesting and our results apply in these settings. However, our results are most stark for reasonably large values of δ.

independent $X \in \{\pm 1\}^n$ can be sampled with seed length $O(k \cdot \log n)$ [2], yielding a seed length of $O(\log^2 n)$ for the setting of parameters above.

In this work, we ask whether the tail bound of Theorem 1 for k-wise independence is tight. That is, can we prove stronger tail bounds for k-wise independent X?

▶ **Question 2.** *How much independence is needed for X to satisfy a Hoeffding-like tail bound? That is, what is the minimum $k = k(n, \delta, T)$ for which any k-wise independent $X \in \{\pm 1\}^n$ satisfies*

$$\underset{X}{\mathbb{P}}\left[|X \cdot r| \geq T\right] \leq \delta$$

for all $r \in \{-1, 1\}^n$, where $\cdot$ denotes the inner product.

We remark that limited independence is not the only technique for derandomizing concentration bounds. Another construction which achieves seed length $O(\log n \cdot \log(1/\delta))$ is to sample X from a small-bias space [52]. Very recently, Gopalan et al. [27] constructed a much more sophisticated generator with seed length $\tilde{O}(\log(n/\delta))$, which is nearly optimal.

1.1.1 Our Results

Theorem 1 shows that $k(n, \delta, T) \leq O(\log(1/\delta))$ for $T = O(\sqrt{n \log(1/\delta)})$. In this work, we show that this is essentially tight:

▶ **Theorem 3.** *Let $c > 5$ be a constant and n sufficiently large. For $2^{-n^{o(1)}} \leq \delta \leq \frac{1}{\operatorname{poly}(n)}$ and $T = c\sqrt{n \log(1/\delta)}$, we have*

$$k(n, \delta, T) \geq \Omega\left(\frac{\log(1/\delta)}{\log(c)}\right).$$

This means there exists a k-wise independent distribution $X \in \{\pm 1\}^n$ such that

$$\mathbb{P}\left[\left|\sum_{i \in [n]} X_i\right| \geq T\right] > \delta,$$

for $k = k(n, \delta, T)$, n, δ, and T as above.

The only previous lower bound was

$$k(n, \delta, T) \geq \Omega\left(\frac{\log(1/\delta)}{\log n}\right),$$

which holds for any $T \leq n$ and is due to [64]. This is meaningful when $\delta < n^{-\omega(1)}$, but the lower bound is trivial for $\delta = 1/\operatorname{poly}(n)$. Thus our lower bound closes a large gap for the $\delta = 1/\operatorname{poly}(n)$ regime, which is of considerable interest [29, 48, 27].

The lower bound of [64] follows immediately from the fact that a random variable X that can be sampled with seed length s cannot satisfy a nontrivial tail bound with $\delta < 2^{-s}$, and that there exist k-wise independent distributions that can be sampled with seed length $s \leq O(k \cdot \log n)$. Indeed this lower bound holds for all distributions with small seed length and is not specific to k-wise independence.

The most natural way to prove Theorem 3 would be to construct a family of k-wise independent distributions that do not satisfy the required tail bound. However, we instead study the *dual* formulation of the problem (following [4, 20, 9]) and then use lower bound techniques from approximation theory. To the best of our knowledge, this indirect approach is novel for proving impossibility results for k-wise independence. Our results imply the existence of k-wise independent distributions with poor tail bounds, but give no immediate indication as to how to construct them!

We now describe the proof idea in slightly more detail. The answer to Question 2 can be posed in terms of the value of a certain linear program. The variables represent the probability distribution of the random variable X and the constraints force X to be k-wise independent. The objective of the linear program is maximize the tail probability $\mathbb{P}[|X \cdot r| \geq T]$. Thus, the value of the program is at most δ if and only if $k \geq k(n, \delta, T)$. Taking the dual of this linear program and appealing to strong duality yields an alternative characterization of $k(n, \delta, T)$. Namely, $k(n, \delta, T)$ is the smallest k for which the threshold function $F_T(x) = \mathbb{1}(|x| \geq T)$ admits an *upper sandwiching polynomial* of degree k and expectation at most δ. Here, an upper sandwiching polynomial is simply a polynomial p for which $p(x) \geq F_T(x)$ pointwise.

We then use ideas from weighted approximation theory to give a lower bound on k for which such sandwiching polynomials exist. In order to apply these ideas, we make a few symmetrization and approximation arguments to reduce the problem to a continuous one-dimensional problem: Find a degree lower bound for a univariate polynomial that is a good upper sandwich for the function $f_T(x) = \operatorname{sgn}(|x| - T)$, with respect to a Gaussian distribution. The solution to this problem appeals to a weighted Markov-type inequality. This inequality generalizes the classical Markov inequality for uniform approximations, which gives a bound on the derivative of a low-degree polynomial that is bounded on the unit interval:

▶ **Theorem 4** ([47]). *Let p be a polynomial of degree d with $|p(x)| \leq 1$ on the interval $[-1, 1]$. Then $|p'(x)| \leq d^2$ on $[-1, 1]$.*

The idea is that an upper sandwich for f_T must have a large jump at the threshold T, which is impossible for low-degree polynomials. The formal proof of this claim is based on a variant of an "infinite-finite range" inequality, which asserts that the weighted norm of a polynomial on the real line is bounded by its norm on a finite interval.

1.2 Agnostically Learning Halfspaces

Halfspaces are a fundamental concept class in machine learning, both in theory and in practice.[3] Their study dates back to the Perceptron algorithm of the 1950s. Halfspaces serve as building blocks in many applications, including boosting and kernel methods.

Halfspaces can be learned in the PAC model [67] either by solving a linear program, or via simple iterative update algorithms (e.g. the Perceptron algorithm). However, learning halfspaces with classification noise is a much more difficult problem, and often needs to be dealt with in practice.

In this work, we study a challenging model of *adversarial noise* – the agnostic learning model of Kearns et al. [38]. In this model, a learner has access to examples drawn from a distribution $\mathcal{D}$ on $X \times \{\pm 1\}$ and must output a hypothesis $h : X \to \{\pm 1\}$ such that

$$\mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \leq \mathrm{opt} + \varepsilon,$$

where opt is the error of the best concept in the concept class – that is,

$$\mathrm{opt} = \min_{f \in \mathcal{C}} \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[f(x) \neq y].$$

[3] A halfspace is a function $f : \mathbb{R}^n \to \{\pm 1\}$ given by $f(x) = \operatorname{sgn}(w \cdot x - \theta)$ for $w \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$, where $\operatorname{sgn}(x) = 1$ if $x \geq 0$ and $\operatorname{sgn}(x) = -1$ otherwise.

The theory of agnostic learning is not well-understood, even in the case of halfspaces. Positive results for efficient agnostic learning of high-dimensional halfspaces are restricted to limited classes of distributions.[4] For instance, halfspaces can be learned under the uniform distribution over the hypercube or the unit sphere, or on any log-concave distribution [35, 39]. On the negative side, a variety of both computational and information-theoretic hardness results are known. For instance, proper agnostic learning of halfspaces (where the learner is required to output a hypothesis that is itself a halfspace) is known to be NP-hard [24, 30]. Moreover, agnostically learning halfspaces under arbitrary distributions is as hard as PAC learning DNFs [44], which is a longstanding open problem.

There is essentially only one known technique for agnostically learning high-dimensional halfspaces: the L_1 regression algorithm [35], which we discuss in more detail in Section 3.2. In its most general form, the algorithm selects a linear space of functions $\mathcal{H} \subset \{h : X \to \mathbb{R}\}$. After drawing a number of examples (x_i, y_i) from $\mathcal{D}$, it computes

$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} \sum_i |h(x_i) - y_i|.$$

The output of the algorithm is $\mathrm{sgn}(h^*(x) - t)$ for some t. We need to ensure that the minimisation can be computed efficiently (e.g. by linear programming) and that every concept $f \in \mathcal{C}$ can be approximated by some $h \in \mathcal{H}$ – that is $\mathop{\mathbb{E}}_{x \sim \mathcal{D}}[|h(x) - f(x)|] \le \varepsilon$. If this is the case, then $\mathcal{C}$ is agnostically learnable in time $\mathrm{poly}(|\mathcal{H}|)$.

Kalai et al. (and most subsequent work on learning using L_1 regression, e.g. [40, 26, 11, 36, 25]) chose $\mathcal{H}$ to be the class of low-degree polynomials. They showed that under certain classes of distributions, every halfspace can be approximated by a polynomial of degree $O_\varepsilon(1)$, and hence halfspaces are agnostically learnable in time $n^{O_\varepsilon(1)}$.

Distributional assumptions arise because the L_1 approximation measure (namely $\mathop{\mathbb{E}}_{x \sim \mathcal{D}}[|h(x) - f(x)|]$) depends on the underlying distribution. A distribution-independent approximation would require an L_∞ approximation, which is too much to hope for in many circumstances.

1.2.1 Our Results

In this work, we ask whether polynomial regression can be extended to work beyond the classes of distributions studied by Kalai et al. In particular, can polynomials provide good L_1 approximations to halfspaces under distributions with heavy tails? Such distributions, including power law distributions, arise naturally in many physical, biological, and networking contexts. Certain learning problems even require heavy tailed distributions on examples [51].

Our result addressing this question (Theorem 6) is a negative one. We show that polynomial approximations to halfspaces do not exist for a large class of distributions, namely:

▶ **Definition 5.** An absolutely continuous distribution $\mathcal{D}$ on $\mathbb{R}$ is a *log-superlinear (LSL) distribution* if there exist $C > 0$ and $\gamma \in (0, 1)$ such that the density w of $\mathcal{D}$ satisfies $w(x) \ge C \exp(-|x|^\gamma)$.[5]

[4] An efficient algorithm is one which runs in time polynomial in the dimension n for any constant $\varepsilon > 0$ – that is, time $n^{O_\varepsilon(1)}$.

[5] The name log-superlinear comes from the fact that the tails of the probability density function of a LSL distribution are heavier than that of the log-linear Laplace distribution.

▶ **Theorem 6.** *For any LSL distribution $\mathcal{D}$, there exists $\varepsilon > 0$ such that no polynomial (of any degree) can approximate the sign function with L_1 error less than ε with respect to $\mathcal{D}$.*

In particular, this implies that the polynomial regression algorithm is not able to agnostically learn thresholds on the real line to within arbitrarily small error. Note that this result does not rule out the possibility that halfspaces can be agnostically learned by other techniques. Indeed, the classic approach of empirical risk minimization (see [38] and the references therein) gives an efficient algorithm for learning thresholds (which are halfspaces in one dimension) under arbitrary distributions. Thus the problem of learning real thresholds under LSL distributions is an explicit example for which polynomial regression fails while other techniques can succeed.

If we were to take $\gamma \geq 1$, the probability density function $C(\gamma)e^{-|x|^\gamma}$ (where $C(\gamma)$ is a normalising constant) would give a log-concave distribution, in which case Kalai et al. [35] show that good polynomial approximations to halfspaces exist. Thus our result gives a threshold between where polynomial approximations to halfspaces exist and where they do not.

Our result for thresholds extends readily to an impossibility result for learning halfspaces over $\mathbb{R}^n$:

▶ **Theorem 7.** *For any product distribution $\mathcal{D}$ on $\mathbb{R}^n$ with a LSL marginal distribution on some coordinate, there exists $\varepsilon > 0$ and a halfspace h such that no polynomial can approximate h with L_1 error less than ε with respect to $\mathcal{D}$.*

As with Theorem 3, the proof of Theorem 6 relies on several Markov-type inequalities for weighted polynomial approximations. Early work on the approximate degree of boolean functions [56, 59] used Markov's inequality to get tight lower bounds on the degree of uniform approximations to symmetric functions. For weighted approximations under LSL distributions, we actually get a much stronger statement. It turns out that the derivative of a polynomial is bounded near the origin *independent of the degree* as long as that polynomial is absolutley bounded when integrated under a LSL distribution. With this powerful fact in hand, the proof of Theorem 6 is quite simple. Consider the threshold function $f(t) = \text{sgn}(t)$. Since f has a "jump" at zero, any good polynomial approximation to f must be bounded and have a large derivative near zero. The higher quality the approximation, the larger a derivative we need. But since the derivative of any polynomial is bounded by a constant, we cannot get arbitrarily good approximations to f using polynomials.

We give the full proof in Section 3.4, and discuss the multivariate generalization in Section 3.5.

1.2.2 Related Work

Our result echoes prior work establishing the limits of *uniform* polynomial approximations for various concept classes. For instance, the seminal work of Minsky and Papert [49] showed that there is an *intersection* of two halfpsaces over $\mathbb{R}^n$ which cannot be represented as the sign of any polynomial. Building on work of Nisan and Szegedy [56], Paturi [59] gave tight lower bounds for uniform approximations to symmetric boolean functions. This, and subsequent work on lower bounds for approximate degree, immediately imply limitations for distribution-independent agnostic learning via polynomial regression. Klivans and Sherstov [42] also showed a strong generalization of Paturi's result to disjunctions, giving limitations on how well they can be approximated by linear combinations of arbitrary features. By contrast to all of these results, our work shows a strong limitation for certain *distribution-dependent* polynomial approximations.

In the distribution-dependent setting, Feldman and Kothari [25] showed that polynomial regression cannot be used to learn disjunctions with respect to symmetric distributions on the hypercube. Recent work of Daniely et al. [19] also uses ideas from approximation theory to show limitations on a broad class of regression and kernel-based methods for learning halfspaces, even under a margin assumption. While our results only apply to polynomial regression, they hold for approximations of arbitrarily high complexity (i.e. degree), and for a large class of natural distributions.

The limitations we prove for polynomial regression do not rule out the existence of other agnostic learning algorithms, including those using L_1 regression with different feature spaces. Wimmer [68] showed how to use a different family of basis functions to learn halfspaces over symmetric distributions on the hypercube. Subsequent work of Feldman and Kothari [25] improved the running time in the special case of disjunctions. We leave it as an intriguing open question to determine whether other basis functions can be used to learn halfspaces under LSL distributions.

2 Tail Bounds for Limited Independence

Our proof consists of three steps:

§2.1 First we reformulate the question of tail bounds for k-wise independent distributions using linear programming duality and symmetrisation. This reduces the problem to proving a degree lower bound on univariate polynomials. Namely we need to give a lower bound on the degree of a polynomial $p : \{0, 1, \cdots, n\} \to \mathbb{R}$ such that $p(i) \geq 0$ for all i, $p(i) \geq 1$ if $|i - n/2| \geq T$, and $\mathbb{E}[p(i)] \leq \delta$, where i is drawn from the binomial distribution.

§2.2 We then transform the problem from one about polynomials with a discrete domain to one about polynomials with a continuous domain. This amounts to showing that, since $\mathbb{E}[p(i)] \leq \delta$ with respect to the binomial distribution, we can bound $\mathbb{E}[p(x + n/2)]$ with respect to a truncated Gaussian distribution on x.

§2.3 Finally we can apply the tools of weighted approximation theory. We know that $p(x + n/2)$ is small for x near the origin, but $p(T + n/2) \geq 1$. We show that any low-degree polynomial that is bounded near the origin cannot grow too quickly. This implies that p must have high degree.

2.1 Dual Formulation

Question 2 from the introduction is equivalent to finding the smallest k for which the value of the following linear program is at most δ.

Linear Program Formulation of Question 2

$$\begin{aligned} \max_{\psi} \quad & \sum_{x \in \{-1,1\}^n} \psi(x) F_T(x) & \\ \text{s.t.} \quad & \sum_{x \in \{-1,1\}^n} \psi(x) \chi_S(x) = 0 & \text{for all } |S| \leq k \\ & \sum_{x \in \{-1,1\}^n} \psi(x) = 1 & \\ & 0 \leq \psi(x) \leq 1 & \text{for all } x \in \{-1,1\}^n. \end{aligned}$$

Here, $F_T(x) = 1$ if $|x| \geq T$ and is 0 otherwise, and $\chi_S(x)$ is the Fourier character corresponding to $S \subseteq [n]$.

If we set $\underset{X}{\mathbb{P}}[X = x] = \psi(x)$, then the constraints impose that X is a k-wise independent distribution, while the objective function is $\underset{X}{\mathbb{P}}\left[\left|\sum_{i\in[n]} X_i\right| \geq T\right]$. Thus the above linear program finds the k-wise independent distribution with the worst tail bound. If the value of the program is at most δ, then all k-wise independent distributions satisfy the tail bound, as required.

Taking the dual of the above linear program yields the following.

Dual Formulation of Question 2

$$\begin{aligned} &\min_p \; 2^{-n} \sum_{x\in\{-1,1\}^n} p(x) \\ &\text{ s.t. } \deg(p) \leq k \\ &\qquad p(x) \geq F_T(x) \qquad\qquad \text{for all } x \in \{-1,1\}^n. \end{aligned}$$

By strong duality, the value of the dual linear program is the same as that of the primal.

The multilinear polynomial p is an "upper sandwich" of F_T – that is, $p \geq F_T$ and $\underset{X\in\{\pm1\}^n}{\mathbb{E}}[p(X)]$ is minimal. Therefore, $k(n, \delta, T)$ is the smallest k for which F_T admits an upper sandwiching polynomial of degree k with expectation δ.

Consider the shifted univariate symmetrization of F_T

$$F'_T(x) = \begin{cases} 1 \text{ if } |x - n/2| \geq T \\ 0 \text{ otherwise.} \end{cases}$$

By applying the well-known Minsky-Papert symmetrization [49] to the dual formulation above, we get the following characterization.

▶ **Theorem 8.** *The quantity $k(n, \delta, T)$ from Question 2 is the smallest k for which there exists a degree-k univariate polynomial $p : \{0, \dots, n\} \to \mathbb{R}$ such that*
1. *$p(i) \geq F'_T(i)$ for all $0 \leq i \leq n$ and*
2. *$2^{-n} \sum_{i=0}^{n} \binom{n}{i} p(i) \leq \delta$.*

The upper bound on $k(n, \delta, T)$ (Theorem 1) is proved (in the appendix) by showing that

$$p(i) = \left(\frac{i - n/2}{T}\right)^k$$

satisfies the requirements of Theorem 8 for an appropriate even k.[6] So this characterisation does in fact capture how upper bounds are proved. The fact that it is a tight characterisation allows us to prove that a barrier to the technique is in fact an impossibility result.

With this characterisation of our problem, we may move on to proving inapproximability results.

2.2 A Continuous Version

To apply techniques from the theory of weighted polynomial approximations, we move to polynomials on a continuous domain. We replace the binomial distribution upon which Theorem 8 evaluates p with a Gaussian distribution.

[6] While our results show that this polynomial is *asymptotically* optimal, numerical experiments have shown that it is not exactly optimal.

Define the probability density function

$$w(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}.$$

We define the L_∞ norm with respect to the weight w:

$$\|g\|_{L_\infty(S)} = \sup_{x \in S} |g(x)| w(x).$$

Now we can give the continuous version of the problem:

▶ **Theorem 9.** *Let $T = c\sqrt{n \log(1/\delta)}$ for $c \geq 5$, and $d = k(n, \delta, T)$. Assume $n \geq (12c)^2 (3 \log(1/\delta))^3$. Then for $T' = 4cT/\sqrt{n}$, there is a degree d polynomial q such that*

1. *$q(T') = q(-T') \geq 1$ and*
2. *$\|q\|_{L_\infty[-\sqrt{d}, \sqrt{d}]} \leq \delta^{0.9}(n+1)$.*

The following lemma is key to moving from the discrete to the continous setting. It shows that if a polynomial is bounded at evenly spaced points, then it must also be bounded between those points, assuming the number of points is sufficiently large relative to the degree.

▶ **Lemma 10** (adaptation of [23, 62, 56]). *Let q be a polynomial of degree d such that $|q(i)| \leq 1$ for $i = 0, 1, \ldots, m$, where $3d^2 \leq m$. Then $|q(x)| \leq \frac{3}{2}$ for all $x \in [0, m]$.*

Proof. Let $a = \max_{x \in [0,m]} |q'(x)|$. Then by the mean value theorem, $|q(x)| \leq 1 + a/2$ for $x \in [0, m]$. By Markov's inequality ([47], see also [17]),

$$a \leq \frac{2d^2(1 + a/2)}{m}.$$

Rearranging gives

$$\frac{a}{2+a} \leq \frac{d^2}{m} \leq \frac{1}{3}.$$

Therefore, $a \leq 1$, and hence $|q(x)| \leq \frac{3}{2}$ for $x \in [0, m]$. ◀

We also require the following anti-concentration lemma.

▶ **Lemma 11.**

$$\binom{n}{n/2 + \alpha\sqrt{n}} \geq \frac{2^{n - 6\alpha^2}}{n+1}.$$

Proof. It is well known via Stirling's approximation that $\binom{n}{k} \geq 2^{nH(k/n)}/(n+1)$, where $H(\cdot)$ denotes the binary entropy function. We estimate

$$\begin{aligned} H\left(\frac{1}{2} + \frac{\alpha}{\sqrt{n}}\right) &\geq \left(\frac{1}{2} + \frac{\alpha}{\sqrt{n}}\right)\left(1 - \frac{2\alpha}{(\log 2)\sqrt{n}}\right) + \left(\frac{1}{2} - \frac{\alpha}{\sqrt{n}}\right)\left(1 + \frac{2\alpha}{(\log 2)\sqrt{n}}\right) \\ &\geq 1 - \frac{4\alpha^2}{(\log 2)n}, \end{aligned}$$

which concludes the proof. ◀

Proof of Theorem 9. Let p be the polynomial promised by Theorem 8. By Theorem 1, we know that $d \leq 3\log(1/\delta)$. Define

$$q(x) = p(x\sqrt{n}/4c + n/2).$$

Then $q(\pm T') = p(\pm T + n/2) \geq F'_T(\pm T + n/2) = 1$, dispensing with the first claim.

Now for all integers i in the interval $n/2 \pm \sqrt{nd}/4c$, we have

$$2^{-n}\binom{n}{i}|p(i)| \leq \delta$$

and hence, by Lemma 11,

$$|p(i)| \leq \frac{2^n\delta}{\binom{n}{n/2+\sqrt{nd}/4c}} \leq (n+1)\delta 2^{6d/16c^2} \leq (n+1)\delta^{1-18/16c^2} \leq (n+1)\delta^{0.9}.$$

By Lemma 10, $|p(x)| \leq \frac{3}{2}(n+1)\delta^{0.9}$ on the whole interval $n/2 \pm \sqrt{nd}/4c$. Thus $|q(x)| \leq \frac{3}{2}(n+1)\delta^{0.9}$ on $[-\sqrt{d}, \sqrt{d}]$, completing the proof. ◀

2.3 The Lower Bound

Now we state the result we need from approximation theory. The following "infinite-finite range inequality" shows that the norm of weighted polynomial on the real line is determined by its norm on a finite interval around the origin. Thus, an upper bound on the magnitude of a polynomial near the origin yields a bound on its growth away from the origin. We will apply this to the polynomial given to us in Theorem 9.

▶ **Theorem 12.** *For any polynomial p of degree d and $B > 1$,*

$$\|p\|_{L_\infty(\mathbb{R}\setminus[-B\sqrt{d},B\sqrt{d}])} \leq (2eB)^d \exp(-B^2 d)\|p\|_{L_\infty[-\sqrt{d},\sqrt{d}]}.$$

The proof follows [45, Theorem 6.1] and [54, Theorem 4.16.12].

Proof. Let $\tilde{p}$ be a polynomial of degree d. Let $T_d(x)$ denote the dth Chebyshev polynomial of the first kind [17]. By the extremal properties of T_d, we have

$$|\tilde{p}(x)| \leq |T_d(x)|\left(\max_{t\in[-1,1]} |\tilde{p}(t)|\right) \leq (2|x|)^d\left(\max_{t\in[-1,1]} |\tilde{p}(t)|\right)$$

for $|x| \geq 1$. Rescaling $p(x) = \tilde{p}(x/\sqrt{d})$ yields

$$|p(x)| \leq \left(\frac{2|x|}{\sqrt{d}}\right)^d \left(\max_{t\in[-\sqrt{d},\sqrt{d}]} |p(t)|\right) \leq \sqrt{\pi}e^d\left(\frac{2|x|}{\sqrt{d}}\right)^d \|p\|_{L_\infty[-\sqrt{d},\sqrt{d}]}$$

for $|x| \geq \sqrt{d}$. Now let $|x| = B\sqrt{d}$ for some $B > 1$. Then

$$|p(x)|w(x) \leq e^d(2B)^d\exp(-B^2 d)\|p\|_{L_\infty[-\sqrt{d},\sqrt{d}]}.$$

Since the coefficient $(2eB)^d\exp(-B^2 d)$ is decreasing in B, this proves the claim. ◀

The above approximation theory result, combined with our continuous formulation Theorem 9, enables us to complete the proof.

▶ **Theorem 13.** *Let $T = c\sqrt{n\log(1/\delta)}$ for $c \geq 5$. Assume $n \geq (12c)^2(3\log(1/\delta))^3$ and $\delta \leq 1/n^4$. Then $k(n, \delta, T) > \log(1/\delta)/9\log c$.*

Proof. Let q be the polynomial given by Theorem 9. Let $T' = 4cT/\sqrt{n}$, $d = \log(1/\delta)/9\log c$, and $B = T'/\sqrt{d} = 12c^2\sqrt{\log c}$. For the sake of contradiction, we suppose that q satisfies the conditions of Theorem 9, but $\deg(q) \leq d$. Then

$$\|q\|_{L_\infty(\mathbb{R}\setminus[-B\sqrt{d},B\sqrt{d}])} = \|q\|_{L_\infty(\mathbb{R}\setminus[-T',T'])} \geq \frac{\exp(-T'^2)}{\sqrt{\pi}}.$$

On the other hand, applying Theorem 12, gives

$$\|q\|_{L_\infty(\mathbb{R}\setminus[-B\sqrt{d},B\sqrt{d}])} \leq (2eB)^d \exp(-T'^2)\delta^{0.9}(n+1).$$

Combining the two inequalities gives

$$\frac{1}{\sqrt{\pi}} \leq (2eB)^d\delta^{0.9}(n+1) \leq \left(24ec^2\sqrt{\log(c)}\right)^{\log(1/\delta)/9\log(c)} \delta^{0.9}(n+1) \leq \delta^{1/3}(n+1),$$

which is a contradiction. ◀

Theorem 13 yields Theorem 3.

3 Agnostically Learning Halfspaces

The class of log-concave distributions over $\mathbb{R}^n$ (defined below) is essentially the broadest under which we know how to agnostically learn halfspaces. While many distributions used in machine learning are log-concave, such as the normal, Laplace, beta, and Dirichlet distributions, log-concave distributions do not capture everything. For instance, the log-normal distribution and heavier-tailed exponential power law distributions are not log-concave. The main motivating question for this section is whether we can relax the assumption of log-concavity for agnostically learning halfspaces. To this end, we show a negative result: for LSL distributions, agnostic learning of halfspaces will require new techniques.

3.1 Background

Our starting point is the work of Kalai et al. [35]. Among their results is the following.

▶ **Theorem 14** ([35]). *The concept class of halfspaces over $\mathbb{R}^n$ is agnostically learnable in time* $\mathrm{poly}(n^{O_\varepsilon(1)})$ *under log-concave distributions.*

A log-concave distribution is an absolutely continuous probability distribution such that the logarithm of the probability density function is concave. For example, the standard multivariate Gaussian distribution on $\mathbb{R}^n$ has the probability density function $x \mapsto e^{-||x||_2^2/2}/(2\pi)^{n/2}$. The natural logarithm of this is $-||x||_2^2/2 - n/2 \cdot \log(2\pi)$, which is concave. The class of log-concave distributions also includes the Laplace distribution and other natural distributions. However, it does not contain heavy-tailed distributions (such as power laws) nor non-smooth distributions (such as discrete probability distributions).

Kalai et al. also show that we can agnostically learn halfspaces under the uniform distribution over the hypercube $\{\pm 1\}^n$ or over the unit sphere $\{x \in \mathbb{R}^n : ||x||_2 = 1\}$.

3.2 The L_1 Regression Algorithm

The results of Kalai et al. are based on the so-called L_1 regression algorithm, which relies on being able to approximate the concept class in question by a low-degree polynomial:

▶ **Theorem 15** ([35]). *Fix a distribution $\mathcal{D}$ on $X \times \{\pm 1\}$ and a concept class $\mathcal{C} \subset \{f : X \to \{\pm 1\}\}$.*[7] *Suppose that, for all $f \in \mathcal{C}$, there exists a polynomial $p : X \to \mathbb{R}$ of degree at most d such that $\underset{x \sim \mathcal{D}_X}{\mathbb{E}}[|p(x) - f(x)|] \le \varepsilon$, where $\mathcal{D}_X$ is the marginal distribution of $\mathcal{D}$ on X. Then, with probability $1 - \delta$ the L_1 regression algorithm outputs a hypothesis h such that*

$$\underset{(x,y) \sim \mathcal{D}}{\mathbb{P}}[h(x) \neq y] \le \min_{f \in \mathcal{C}} \underset{(x,y) \sim \mathcal{D}}{\mathbb{P}}[f(x) \neq y] + \varepsilon$$

in time $\mathrm{poly}(n^d, 1/\varepsilon, \log(1/\delta))$ with access only to examples drawn from $\mathcal{D}$.

The L_1 regression algorithm solves a linear program to find a polynomial p of degree at most d that minimises $\sum_i |p(x_i) - y_i|$, where (x_i, y_i) are the examples sampled from $\mathcal{D}$. The hypothesis is then $h(x) = \mathrm{sgn}(p(x) - t)$, where $t \in [-1, 1]$ is chosen to minimise the error of h on the examples.

Given Theorem 15, proving Theorem 14 reduces to showing that halfspaces can be approximated by low-degree polynomials under the distributions we are interested in. It is important to note that making assumptions on the distribution is necessary (barring a major breakthrough): Agnostically learning halfspaces under arbitrary distributions is at least as hard as PAC learning DNF formulas [44]. Moreover, proper learning of halfspaces under arbitrary distributions is known to be NP-hard [24].

In fact, we can reduce the task of approximating a halfspace to a one-dimensional problem. A halfspace is given by $f(x) = \mathrm{sgn}(w \cdot x - \theta)$ for some $w \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$. It suffices to find a univariate polynomial p of degree at most d such that $\underset{x \sim \mathcal{D}_{w,\theta}}{\mathbb{E}}[|p(x) - \mathrm{sgn}(x)|] \le \varepsilon$, where $\mathcal{D}_{w,\theta}$ is the distribution of $w \cdot x - \theta$ when x is drawn from $\mathcal{D}_X$. If $\mathcal{D}_X$ is log-concave, then so is $\mathcal{D}_{w,\theta}$.

3.3 On the Density of Polynomials

In this section, we give some intuition for why one might expect that polynomial approximations do not suffice for learning under LSL distributions. It turns out that under a LSL distribution w, polynomials actually fail to be dense in the space $C_0[w]$ of continuous functions vanishing at infinity when weighted by w. This is in stark contrast to the classical Weierstrass approximation theorem, which asserts that the polynomials are dense in C_0 under the uniform weight. These kinds of results address *Bernstein's approximation problem* [10], a precise statement of which is as follows.

▶ **Question 16.** *Let $w : \mathbb{R} \to [0, 1]$ be a measurable function. Let $C_0[w]$ denote the space of continuous functions f for which $\lim_{|x| \to \infty} f(x)w(x) = 0$. Under what conditions on w is it true that for every $f \in C_0[w]$, there is a sequence of polynomials $\{p_n\}_{n=1}^{\infty}$ for which*

$$\lim_{n \to \infty} \|(p_n - f)w\|_\infty = 0?$$

(The choice of the L_∞ norm here appears to make very little difference). If Bernstein's problem admits a positive resolution, we say that the polynomials are *dense* in $C_0[w]$. The excellent survey of Lubinsky [45] presents a number of criteria for when polynomials are dense. The one that is most readily applied was proved by Carleson [16] (but appears to be implicit in [34]):

[7] Here $X = \mathbb{R}^n$.

▶ **Theorem 17.** *Let w be even and positive with $\log(w(e^x))$ concave. Then the polynomials are dense in $C_0[w]$ iff*

$$\int_0^\infty \frac{\log w(x)}{1+x^2}\mathrm{d}x = -\infty.$$

This immediately yields the following dichotomy result for exponential power distributions:

▶ **Corollary 18.** *For $\gamma > 0$ and $w_\gamma(x) = \exp(-|x|^\gamma)$, the polynomials are dense in $C_0[w_\gamma]$ iff $\gamma \geq 1$.*

In particular, this justifies our assertion that the polynomials fail to be dense in the continuous functions under LSL distributions.

So what does this have to do with agnostically learning halfspaces? Recall that the analysis of the L_1-regression algorithm of Kalai et al. [35] reduces approximating a halfspace under a distribution $\mathcal{D}$ to the problem of approximating each threshold function $\mathrm{sgn}(x-\theta)$ under each marginal distribution of $\mathcal{D}$. So for the algorithm to work, we require $\mathcal{D}$ to have marginals w under which $\mathrm{sgn}(x-\theta)$ can be approximated arbitrarily well by polynomials. Now if the polynomials are dense in $C_0[w]$, then threshold functions can also be approximated arbitrarily well (since $C_0[w]$ is in turn dense in $L_1[w]$). Such an appeal to density actually underlies Kalai et al.'s proof of approximability under log-concave distributions. On the other hand, if the polynomials fail to be dense, then one might conjecture that thresholds cannot be arbitrarily well approximated.

Our result, presented in the next section, confirms the conjecture that even the *sign* function cannot be approximated arbitrarily well by polynomials under LSL distributions

3.4 Lower Bound for One Variable

Consider the LSL density function

$$w_\gamma(x) := C(\gamma)\exp(-|x|^\gamma)$$

on the reals for $\gamma \in (0,1)$, where $C(\gamma)$ is a normalizing constant. Define the sign function $\mathrm{sgn}(x) = 1$ if $x \geq 0$ and $\mathrm{sgn}(x) = -1$ otherwise. In this section, we show that for sufficiently small ε, the sign function does not have an L_1 approximation under the distribution w_γ. More formally,

▶ **Proposition 19.** *For any $\gamma \in (0,1)$, there exists an $\varepsilon = \varepsilon(\gamma)$ such that for any polynomial p,*

$$\int_{\mathbb{R}} |p(x) - \mathrm{sgn}(x)| w_\gamma(x)\, dx > \varepsilon.$$

The proof is based on the following Markov-type inequality, which roughly says that a bounded polynomial cannot have a large derivative (under the weight w_γ). This implies the claim, since the sign function we are trying to approximate has a large "jump" at the origin.

▶ **Lemma 20.** *For $\gamma \in (0,1)$ there is a constant $M(\gamma)$ such that*

$$\sup_{x\in\mathbb{R}}(|p'(x)|w_\gamma(x)) \leq M(\gamma)\int_{\mathbb{R}} |p(x)|w_\gamma(x)\, dx.$$

Proof. The lemma is a combination of a Markov-type inequality and a Nikolskii-type, available in a survey of Nevai [54]:

▶ **Theorem 21** ([55], [54, Theorem 4.17.4]). *There exists a constant $C_1(\gamma)$ such that for any polynomial p,*

$$\int_{\mathbb{R}} |p'(x)|w_\gamma(x)\ dx \leq C_1(\gamma) \int_{\mathbb{R}} |p(x)|w_\gamma(x)\ dx.$$

▶ **Theorem 22** ([53], [54, Theorem 4.17.5]). *There exists a constant $C_2(\gamma)$ such that for any polynomial p,*

$$\sup_x(|p(x)|w_\gamma(x)) \leq C_2(\gamma) \int_{\mathbb{R}} |p(x)|w_\gamma(x)\ dx.$$

◀

Proof of Proposition 19. Fix $\varepsilon \in (0,1)$ and suppose p is a polynomial satisfying

$$\int_{\mathbb{R}} |p(x) - \operatorname{sgn}(x)|w_\gamma(x)\ dx \leq \varepsilon.$$

Since the absolute value of the sign function integrates to 1, this forces

$$\int_{\mathbb{R}} |p(x)|w_\gamma(x)\ dx \leq 1 + \varepsilon \leq 2.$$

Therefore, we have by Lemma 20 that $|p'(x)|w_\gamma(x) \leq 2M(\gamma)$ for every x.

The idea is now to show that there is some x_0 for which $|p'(x_0)|w_\gamma(x_0) \geq \Omega(1/\varepsilon)$. To see this, let $\delta = 4\varepsilon/C(\gamma)$ and observe that there must exist some $x_+ \in [0,\delta]$ such that $p(x_+) \geq 1/2$. If this were not the case, then we would have

$$\int_{\mathbb{R}} |p(x) - \operatorname{sgn}(x)|w_\gamma(x)\ dx \geq \int_0^\delta \left(1 - \frac{1}{2}\right) C(\gamma)\exp(-x^\gamma)\ dx \geq \frac{\delta}{2} C(\gamma)\exp(-\delta^\gamma) \geq \varepsilon$$

for ε small enough, and hence δ small enough, to make $\exp(-\delta^\gamma) \geq 1/2$, yielding a contradiction. A similar argument shows that there is some $x_- \in [-\delta, 0]$ with $p(x_-) \leq -1/2$. Therefore, by the mean value theorem, there is some $x_0 \in [x_-, x_+]$ with $p'(x_0) \geq 1/2\delta = C(\gamma)/8\varepsilon$. Moreover, because we took δ small enough, we also have $p'(x_0)w(x_0) \geq C(\gamma)/16\varepsilon$. This shows that no polynomial ε-approximates sgn as long as $\varepsilon < C/32M$. ◀

Moreover, the proposition shows that it is impossible to get arbitrarily close polynomial approximations to halfspaces under densities w for which there are constants C and $\gamma \in (0,1)$ with $w(x) \geq C\exp(-|x|^\gamma)$ for all $x \in \mathbb{R}$. This shows that LSL distributions on $\mathbb{R}$ do not enable arbitrarily close polynomial approximations to halfspaces.

3.5 Extending the Lower Bound to Multivariate Distributions

It is straightforward to extend the lower bound from the previous section to product distributions with LSL marginals.

▶ **Theorem 23.** *Let $X = (X_1, \ldots, X_n)$ be a random variable over $\mathbb{R}^n$ with density $f_X(x) = w(x_1)f(x_2, \ldots, x_n)$. Suppose the density w specifies a univariate γ-LSL distribution. Then there exists an $\varepsilon = \varepsilon(\gamma)$ such that for any polynomial p,*

$$\int_{\mathbb{R}^n} |p(x_1, \ldots, x_n) - \operatorname{sgn}(x_1)|f_X(x_1, \ldots, x_n)\ dx_1 dx_2 \ldots dx_n > \varepsilon.$$

That is, the linear threshold function $\operatorname{sgn}(x_1)$ *cannot be approximated arbitrarily well by polynomials.*

Proof. Let $p(x_1, \dots, x_n)$ be a polynomial, and define a univariate polynomial q by "averaging out" the variables $x_2, \dots, x_n$:

$$q(x_1) := \int_{\mathbb{R}^{n-1}} p(x_1, \dots, x_n) f(x_2, \dots, x_n) \; dx_2 \dots \; dx_n.$$

Then we have

$$\begin{aligned}
&\int_{\mathbb{R}} |q(x_1) - \operatorname{sgn}(x_1)| w(x_1) \; dx_1 \\
&\quad = \int_{\mathbb{R}} \left| \int_{\mathbb{R}^{n-1}} (p(x_1, \dots, x_n) - \operatorname{sgn}(x_1)) f(x_2, \dots, x_n) \; dx_2 \dots dx_n \right| w(x_1) \; dx_1 \\
&\quad \le \int_{\mathbb{R}} \left(\int_{\mathbb{R}^{n-1}} |p(x_1, \dots, x_n) - \operatorname{sgn}(x_1)| f(x_2, \dots, x_n) \; dx_2 \dots dx_n \right) w(x_1) \; dx_1 \\
&\quad = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} |p(x_1, \dots, x_n) - \operatorname{sgn}(x_1)| f_X(x_1, \dots, x_n) \; dx_1 dx_2 \dots dx_n.
\end{aligned}$$

By Proposition 19, the latter quantity must be at least $\varepsilon(\gamma)$. ◀

Let $w_\gamma^n(x) \propto \exp(-(|x_1|^\gamma + \dots + |x_n|^\gamma))$ denote the density of the "prototypical" multivariate LSL distribution, with each marginal having the same exponential power law distribution. Our impossibility result holds uniformly for every distribution in the sequence $\{w_\gamma^n\}$. That is, for every $\gamma \in (0, 1)$, there exists $\varepsilon = \varepsilon(\gamma)$ for which halfspaces cannot be learned by polynomials under any of the distributions specified by $\{w_\gamma^n\}$.

As a consequence, we get inapproximability results for several natural classes of distributions that dominate $\{w_\gamma^n\}$ by constant factors (i.e. not growing with n).

1. Any power-law distribution, i.e. a distribution with density $\propto \|x\|^{-M}$ for some constant M, since such a distribution dominates every w_γ^n.
2. Multivariate generalizations of the log-normal distribution, i.e. any distribution with density $\propto \exp(-\operatorname{polylog}(\|x\|))$.
3. Multivariate exponential power distributions, which have densities $\propto \exp(-\|x\|^\gamma)$ for $\gamma \in (0, 1)$. These distributions dominate the prototypical w_γ^n by the inequality of ℓ_p-norms:

   $$\|x\|^\gamma \le |x_1|^\gamma + \dots + |x_n|^\gamma$$

   for every $0 \le \gamma \le 2$.

4 Further Work

Our negative results naturally suggest a number of directions for future work.

Are there other suitable derandomizations of concentration inequalities? In this work, we focused on understanding the limits of k-wise independent distributions. Gopalan et al. [27] gave a much more sophisticated generator with nearly optimal seed length. But could simple, natural pseudorandom distributions, such as small-bias spaces, give strong tail bounds themselves?

Are halfspaces agnostically learnable under LSL distributions? Our negative result does not even necessarily rule out the use of L_1 regression for this task: The polynomial regression algorithm of Kalai et al. [35] is in fact quite flexible. Nothing is really special about the basis of low-degree monomials, and the algorithm works equally well over any small, efficiently evaluable "feature space". That is, if we can show that halfspaces are well-approximated by linear combinations of features from a feature space $\mathcal{F}$ under a distribution $\mathcal{D}$, then we can agnostically learn halfspaces with respect to $\mathcal{D}$ in time proportional to $|\mathcal{F}|$. Could one hope for such approximations? Wimmer [68] and Feldman and Kothari [25] have shown how to use non-polynomial basis functions to obtain faster learning algorithms on the boolean hypercube. On the other hand, recent work of Dachman-Soled et al. [18] shows that, at least for product distributions on the hypercube, polynomials yield the best basis for L_1 regression.

Acknowledgements. We thank Varun Kanade, Scott Linderman, Raghu Meka, Jelani Nelson, Justin Thaler, Salil Vadhan, Les Valiant, and several anonymous reviewers for helpful discussions and comments.

References

1 Scott Aaronson and Yaoyun Shi. Quantum lower bounds for the collision and the element distinctness problems. *J. ACM*, 51(4):595–605, 2004.

2 Noga Alon, László Babai, and Alon Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *Journal of Algorithms*, 7(4):567–583, 1986.

3 Noga Alon and Asaf Nussboim. K-wise independent random graphs. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 813–822. IEEE, 2008.

4 Louay M. J. Bazzi. Polylogarithmic independence can fool DNF formulas. *SIAM J. Comput.*, 38(6):2220–2272, March 2009.

5 Robert Beals, Harry Buhrman, Richard Cleve, Michele Mosca, and Ronald de Wolf. Quantum lower bounds by polynomials. *J. ACM*, 48(4):778–797, 2001.

6 Richard Beigel. The polynomial method in circuit complexity. In *Structure in Complexity Theory Conference*, pages 82–95, 1993.

7 Richard Beigel. Perceptrons, PP, and the polynomial hierarchy. *Computational Complexity*, 4:339–349, 1994.

8 M. Bellare and J. Rompel. Randomness-efficient oblivious sampling. In *FOCS*, pages 276–287, Nov 1994.

9 Itai Benjamini, Ori Gurel-Gurevich, and Ron Peled. On k-wise independent distributions and boolean functions. *arXiv preprint arXiv:1201.3261*, 2012.

10 S. N. Bernstein. Le problème de l'approximation des fonctions continues sur tout l'axe réel et l'une de ses applications. *Bull. Math. Soc. France*, 52:399–410, 1924.

11 Eric Blais, Ryan O'Donnell, and Karl Wimmer. Polynomial regression under arbitrary product distributions. *Machine Learning*, 80(2-3):273–294, 2010.

12 Aline Bonami. Étude des coefficients de fourier des fonctions de $l^p(g)$. *Annales de l'institut Fourier*, 20(2):335–402, 1970.

13 Mark Braverman. Polylogarithmic independence fools AC0 circuits. *J. ACM*, 57(5):28:1–28:10, June 2008.

14 Mark Braverman, Anup Rao, Ran Raz, and Amir Yehudayoff. Pseudorandom generators for regular branching programs. *FOCS*, pages 40–47, 2010.

15 Joshua Brody and Elad Verbin. The coin problem and pseudorandomness for branching programs. In *FOCS*, pages 30–39, 2010.

16 Lennart Carleson. Bernstein's approximation problem. *Proc. Amer. Math. Soc.*, 2:953–961, 1951.

17 E.W. Cheney. *Introduction to Approximation Theory.* AMS Chelsea Publishing Series. AMS Chelsea Pub., 1982.

18 Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. *CoRR*, abs/1405.5268, 2014. To appear in SODA 2015.

19 Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. The complexity of learning halfspaces using generalized linear methods. *CoRR*, abs/1211.0616, 2014.

20 Anindya De, Omid Etesami, Luca Trevisan, and Madhur Tulsiani. Improved pseudorandom generators for depth 2 circuits. In Maria Serna, Ronen Shaltiel, Klaus Jansen, and Jose Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 6302 of *Lecture Notes in Computer Science*, pages 504–517, 2010.

21 Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. In *In Proc. 50th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 171–180, 2009.

22 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 *Proceedings of the 2010 IEEE 25th Annual Conference on Computational Complexity*, CCC'10, pages 211–222, Washington, DC, USA, 2010. IEEE Computer Society.

23 H. Ehlich and K. Zeller. Schwankung von polynomen zwischen gitterpunkten. *Mathematische Zeitschrift*, 86:41–44, 1964.

24 Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS'06, pages 563–574, Washington, DC, USA, 2006. IEEE Computer Society.

25 Vitaly Feldman and Pravesh Kothari. Agnostic learning of disjunctions on symmetric distributions. *CoRR*, abs/1405.6791, 2014.

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

27 Parikshit Gopalan, Daniel Kane, and Raghu Meka. Pseudorandomness for concentration bounds and signed majorities. *CoRR*, abs/1411.4584, 2014.

28 Parikshit Gopalan, Ryan O'Donnell, Yi Wu, and David Zuckerman. Fooling functions of halfspaces under product distributions. In *Proceedings of the 2010 IEEE 25th Annual Conference on Computational Complexity*, CCC'10, pages 223–234, Washington, DC, USA, 2010. IEEE Computer Society.

29 Parikshit Gopalan and Jaikumar Radhakrishnan. Finding duplicates in a data stream. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 402–411. Society for Industrial and Applied Mathematics, 2009.

30 V. Guruswami and P. Raghavendra. Hardness of learning halfspaces with noise. In *Proceedings of FOCS'06*, pages 543–552, 2006.

31 Nicholas J. A. Harvey, Jelani Nelson, and Krzysztof Onak. Sketching and streaming entropy via approximation theory. In *FOCS*, pages 489–498, 2008.

32 Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):pp. 13–30, 1963.

33 Russell Impagliazzo, Noam Nisan, and Avi Wigderson. Pseudorandomness for network algorithms. In *STOC*, pages 356–364, 1994.

34 S. Izumi and T. Kawata. Quasi-analytic class and closure of $\{t^n\}$ in the interval $(-\infty, \infty)$. *Tohoku Math. J.*, 43:267–273, 1937.

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

36 Daniel M. Kane, Adam Klivans, and Raghu Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In *COLT*, pages 522–545, 2013.

37 Daniel M Kane and Jelani Nelson. A derandomized sparse Johnson-Lindenstrauss transform. *arXiv preprint arXiv:1006.3585*, 2010.

38 Michael Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. In *Machine Learning*, pages 341–352. ACM Press, 1994.

39 Adam R. Klivans, Philip M. Long, and Alex K. Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In Irit Dinur, Klaus Jansen, Joseph Naor, and José Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 5687 of *Lecture Notes in Computer Science*, pages 588–600. Springer Berlin Heidelberg, 2009.

40 Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via gaussian surface area. In *FOCS*, pages 541–550, 2008.

41 Adam R. Klivans and Rocco A. Servedio. Learning DNF in time $2^{\tilde{o}(n^{1/3})}$. *J. Comput. Syst. Sci.*, 68(2):303–318, 2004.

42 Adam R. Klivans and Alexander A. Sherstov. Lower bounds for agnostic learning via approximate rank. *Computational Complexity*, 19(4):581–604, 2010.

43 Michal Koucký, Prajakta Nimbhorkar, and Pavel Pudlák. Pseudorandom generators for group products. In *STOC*, pages 263–272, 2011.

44 Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. On efficient agnostic learning of linear combinations of basis functions. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, COLT'95, pages 369–376, New York, NY, USA, 1995. ACM.

45 Doron Lubinsky. A survey of weighted polynomial approximation with exponential weights. *Surveys in Approximation Theory*, 3:1–105, 2007.

46 Michael Luby and Avi Wigderson. *Pairwise independence and derandomization.* Citeseer, 1995.

47 A. A. Markov. On a question of D. I. Mendeleev. *Zapiski Imperatorskoi Akademii Nauk,*, 62:1–24, 1890.

48 Raghu Meka and David Zuckerman. Pseudorandom generators for polynomial threshold functions. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC'10, pages 427–436, New York, NY, USA, 2010. ACM.

49 Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry.* MIT Press, Cambridge MA, 1972.

50 Michael Mitzenmacher and Salil Vadhan. Why simple hash functions work: Exploiting the entropy in a data stream. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA'08, pages 746–755, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

51 Elchanan Mossel and Mesrob I Ohannessian. On the impossibility of learning the missing mass. *arXiv preprint arXiv:1503.03613*, 2015.

52 Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM J. Computing*, 22:838–856, 1993.

53 P. Nevai and V. Totik. Sharp Nikolskii inequalities with exponential weights. *Analysis Mathematica*, 13(4):261–267, 1987.

54 Paul Nevai. Géza Freud, orthogonal polynomials and Christoffel functions. A case study. *Journal of Approximation Theory*, 48(1):3–167, 1986.

55 Paul Nevai and Vilmos Totik. Weighted polynomial inequalities. *Constructive Approximation*, 2(1):113–127, 1986.

56 N. Nisan and M. Szegedy. On the degree of boolean functions as real polynomials. *Computational Complexity*, 4:301–313, 1994.

57 Noam Nisan. $\mathcal{RL} \subset \mathcal{SC}$. In *STOC*, pages 619–623, 1992.

58 Ryan O'Donnell. *Analysis of Boolean Functions.* Cambridge University Press, New York, NY, USA, 2014.

59 Ramamohan Paturi. On the degree of polynomials that approximate symmetric boolean functions (preliminary version). In *STOC*, pages 468–474, 1992.

60 Omer Reingold. Undirected connectivity in log-space. *J. ACM*, 55(4):17:1–17:24, September 2008.

61 Omer Reingold, Thomas Steinke, and Salil Vadhan. Pseudorandomness for regular branching programs via fourier analysis. In *APPROX-RANDOM*, pages 655–670, 2013.

62 T. J. Rivlin and E. W. Cheney. A comparison of uniform approximations on an interval and a finite subset thereof. *SIAM J. Numer. Anal.*, 3(2):311–320, 1966.

63 Sushant Sachdeva and Nisheeth K. Vishnoi. Faster algorithms via approximation theory. *Foundations and Trends in Theoretical Computer Science*, 9(2):125–210, 2014.

64 J. Schmidt, A. Siegel, and A. Srinivasan. Chernoff–Hoeffding bounds for applications with limited independence. *SIAM J. Discrete Mathematics*, 8(2):223–250, 1995.

65 Alexander A. Sherstov. Communication lower bounds using dual polynomials. *Bulletin of the EATCS*, 95:59–93, 2008.

66 Alexander A. Sherstov. Separating AC^0 from depth-2 majority circuits. *SIAM J. Comput.*, 38(6):2113–2129, 2009.

67 Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

68 Karl Wimmer. Agnostically learning under permutation invariant distributions. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS'10, pages 113–122, Washington, DC, USA, 2010. IEEE Computer Society.

A Upper Bound for Limited Independence

Theorem 1 follows from the following well-known [64, 8] lemma, which we prove for completeness.

▶ **Lemma 24.** *Let $X \in \{\pm 1\}^n$ be uniform and $r \in \mathbb{R}^n$. For all even $k \geq 2$,*

$$\mathbb{E}\left[(X \cdot r)^k\right] \leq \left(e\,||r||_2^2\,k\right)^{k/2}.$$

An even stronger form of Lemma 24 follows immediately from the hypercontractivity theorem [12] [58, §9]: Letting $f(x) = x \cdot r$, we have

$$\mathbb{E}\left[(X \cdot r)^k\right] = ||f||_k^k \leq \left((k-1)^{\deg(f)/2}\,||f||_2\right)^k = \left(\sqrt{k-1}\,||r||_2\right)^k,$$

as required. A self-contained proof follows.

Proof. We start by bounding the moment generating function of $X \cdot r$: Let $t \in \mathbb{R}$ be fixed later. For any $i \in [n]$, we have

$$\mathbb{E}\left[e^{tr_iX_i}\right] = \frac{1}{2}\left(e^{tr_i} + e^{-tr_i}\right) = \sum_{k=0}^{\infty} \frac{(tr_i)^k + (-tr_i)^k}{2k!} = \sum_{k=0}^{\infty} \frac{(tr_i)^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{(t^2r_i^2)^k}{2^k k!} = e^{t^2r_i^2/2}.$$

By independence,

$$\mathbb{E}\left[e^{t(X\cdot r)}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{tr_iX_i}\right] \leq \prod_{i=1}^{n} e^{t^2r_i^2/2} = e^{t^2||r||_2^2/2}.$$

We wish to bound a single moment, namely $\mathbb{E}\left[(X\cdot r)^{k_*}\right]$ for an even k_*. We do this by picking one term out of the taylor series of $\mathbb{E}\left[e^{t(X\cdot r)}\right]$. First we remove the odd terms:

$$\sum_{k \text{ even}} \frac{t^k}{k!}\mathbb{E}\left[(X\cdot r)^k\right] = \frac{1}{2}\left(\mathbb{E}\left[e^{t(X\cdot r)}\right] + \mathbb{E}\left[e^{-t(X\cdot r)}\right]\right) \le e^{t^2||r||_2^2/2}$$

We have $\mathbb{E}\left[(X\cdot r)^k\right] \ge 0$ for even k, so we can remove terms from the above infinite sum without increasing it. Thus

$$\frac{t^{k_*}}{k_*!}\mathbb{E}\left[(X\cdot r)^{k_*}\right] \le \sum_{k \text{ even}} \frac{t^k}{k!}\mathbb{E}\left[(X\cdot r)^k\right] \le e^{t^2||r||_2^2/2}.$$

Rearranging and setting $t = \sqrt{k_*}/\left|\left|r\right|\right|_2$, we obtain

$$\mathbb{E}\left[(X\cdot r)^{k_*}\right] \le \frac{k_*!}{t^{k_*}}e^{t^2||r||_2^2/2} = \frac{k_*!\left|\left|r\right|\right|_2^{k_*}e^{k_*/2}}{\sqrt{k_*}^{k_*}} \le \left(\frac{k_*^2\left|\left|r\right|\right|_2^2 e}{k_*}\right)^{k_*/2} = (e\left|\left|r\right|\right|_2^2 k_*)^{k_*/2},$$

as required. ◀

Now we can prove the upper bound for k-wise independence using the connection between moment bounds and tail bounds [64].

Proof of Theorem 1. Note that, if $X \in \{\pm 1\}^n$ is k-wise independent, then

$$\mathbb{E}\left[(X\cdot r)^k\right] = \sum_{i_1\cdots i_k\in[n]}\left(\prod_{j=1}^{k} r_{i_j}\right)\cdot\mathbb{E}\left[\prod_{j=1}^{k} X_{i_j}\right]$$

is the same as for uniform X, as this is the expectation of a degree-k polynomial. By Lemma 24 and Markov's inequality, we have (assuming k is even),

$$\mathbb{P}\left[|X\cdot r| \ge T\right] = \mathbb{P}\left[(X\cdot r)^k \ge T^k\right] \le \frac{\mathbb{E}\left[(X\cdot r)^k\right]}{T^k} \le \left(\frac{e\left|\left|r\right|\right|_2^2 k}{T^2}\right)^{k/2}.$$

Substituting $k = 2\lceil \eta \log_e(1/\delta)\rceil$ and $T = e^{(\eta+1)/2\eta}\sqrt{k}\left|\left|r\right|\right|_2$, we have

$$\mathbb{P}\left[|X\cdot r| \ge T\right] \le \left(\frac{e\left|\left|r\right|\right|_2^2 k}{(e^{(\eta+1)/2\eta}\sqrt{k}\left|\left|r\right|\right|_2)^2}\right)^{\lceil \eta \log_e(1/\delta)\rceil} = e^{-\lceil \eta \log_e(1/\delta)\rceil/\eta} \le \delta.$$

◀