Statistical Query Lower Bounds for Smoothed Agnostic Learning

Ilias Diakonikolas
University of Wisconsin–Madison
ilias@cs.wisc.edu

Daniel M. Kane
UC San Diego
dakane@ucsd.edu

Abstract

We study the complexity of smoothed agnostic learning, in which the learner competes with the best classifier in a target class under slight Gaussian perturbations of the inputs. Specifically, we focus on the prototypical task of agnostically learning halfspaces under subgaussian distributions in this model. The best known upper bound for this problem is based on L_1-polynomial regression and has complexity $d^{\tilde{O}(1/\sigma^2)\log(1/\epsilon)}$, where σ is the smoothing parameter and ϵ is the excess error. Our main result is a Statistical Query (SQ) lower bound showing that this upper bound is close to best possible. In particular, we prove that, even for Gaussian marginals, any SQ algorithm for smoothed agnostic learning of halfspaces requires complexity $d^{\Omega(1/\sigma^2+\log(1/\epsilon))}$. This is the first non-trivial computational lower bound for this task, and it nearly matches the known upper bound. At a conceptual level, we show that the complexity of the problem is governed by the low-degree L_1 approximation of the smoothed target $T_\sigma f$, so that applying L_1-polynomial regression to the smoothed function is essentially optimal in the SQ model. Our proof proceeds by constructing a moment-matching hard distribution via linear programming duality; the dual program corresponds exactly to finding a low-degree approximating polynomial for $T_\sigma f$, which is the same approximation-theoretic condition underlying the upper bound. To instantiate this framework for halfspaces, we prove explicit lower bounds on the approximation degree of smoothed univariate threshold functions. A key ingredient is a new structural result: we construct a distribution that matches moments with a Gaussian while exhibiting periodic structure. This result underlies our $1/\sigma^2$-degree lower bound and may be of independent interest.

1 Introduction

In the agnostic model [Hau92, KSS94], the learner is given samples from a distribution D on labeled examples in $\mathbb{R}^d \times \{\pm 1\}$, and the goal is to compute a hypothesis that is competitive with the *best-fit* function in a known concept class $\mathcal{C}$. In more detail, given n i.i.d. samples $(x^{(i)}, y_i)$ drawn from D and a desired excess error $\epsilon > 0$, an agnostic learner for $\mathcal{C}$ outputs a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ such that $\mathbf{Pr}_{(x,y)\sim D}[h(x) \neq y] \leq \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} \stackrel{\mathrm{def}}{=} \inf_{f\in\mathcal{C}} \mathbf{Pr}_{(x,y)\sim D}[f(x) \neq y]$. Importantly, in the agnostic model, no assumptions are made on the observed labels. As such, agnostic learning can be viewed as a general framework to study supervised learning in the presence of arbitrary misspecification or (potentially adversarial) label noise. It is worth recalling that the agnostic model generalizes Valiant's (realizable) PAC model [Val84], where the labels are promised to be consistent with an unknown function in the class $\mathcal{C}$ (i.e., $\mathrm{OPT} = 0$).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

In the original definition of agnostic learning, no assumptions are made on the marginal distribution of D on x (henceforth denoted by D_x). Learning in this *distribution-free* agnostic model turns out to be computationally challenging. While the model is statistically tractable for all concept classes with bounded VC dimension (see, e.g., [BEHW89]), prior work established super-polynomial computational lower bounds, even for weak agnostic learning of simple concept classes such as halfspaces [Dan16, DKMR22, Tie23].

One approach to circumvent these computational limits involves making natural assumptions about the underlying distribution D_x on feature vectors. Examples include structured continuous distributions—such as the Gaussian distribution or more broadly any log-concave distribution—and discrete distributions—such as the uniform distribution on the Boolean hypercube. While certain computational limitations persist even in this distribution-specific setting, see, e.g., [DKZ20, GGK20, DKPZ21, DKR23], it turns out that non-trivial positive results are possible. Specifically, a line of work has developed efficient learning algorithms with both exact (see, e.g., [KKMS08, KOS08, DKK+21]) and approximate (see, e.g., [ABL17, DKS18, DKTZ20, DKTZ22]) error guarantees in the distribution-specific setting.

A complementary approach—recently introduced in [CKK+24] and the focus of the current paper—involves relaxing the notion of optimality, by working in an appropriate *smoothed* setting. This approach was inspired by a classical line of work in the smoothed analysis of algorithms [ST04]. For concreteness, we define the smoothed agnostic learning framework below.

Definition 1.1 (Smoothed Agnostic Learning [CKK+24])**.** *Let $\mathcal{C}$ be a Boolean concept class on $\mathbb{R}^d$ and $\mathcal{D}_x$ be a family of distributions on $\mathbb{R}^d$. We say that an algorithm $\mathcal{A}$ agnostically learns $\mathcal{C}$ in the* smoothed setting *if it satisfies the following property. Given an error parameter $\epsilon > 0$, a smoothing parameter $\sigma > 0$ and access to i.i.d. samples from a distribution D on $\mathbb{R}^d \times \{\pm 1\}$ such that $D_x \in \mathcal{D}_x$, $\mathcal{A}$ outputs a hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ such that with high probability $\mathbf{Pr}_{(x,y)\sim D}[h(x) \neq y] \leq \mathrm{OPT}_\sigma + \epsilon$, where $\mathrm{OPT}_\sigma \stackrel{\mathrm{def}}{=} \inf_{f\in\mathcal{C}} \mathbf{E}_{z\sim\mathcal{N}_d}\left[\mathbf{Pr}_{(x,y)\sim D}[f(x+\sigma z) \neq y]\right]$.*

Some comments are in order. First note that for $\sigma = 0$ one recovers the standard agnostic model. The hope is that when $\sigma > 0$ is sufficiently large, smoothed agnostic learning becomes computationally tractable in settings where vanilla agnostic learning is not. Specifically, the goal is to characterize (in terms of both upper and lower bounds) the computational complexity of smoothed agnostic learning for natural concept classes $\mathcal{C}$ and distributions $\mathcal{D}_x$, as a function of $d, 1/\epsilon$, and $1/\sigma$. The regime of primary interest is when the smoothing is non-negligible relative to the target excess error, i.e., when $\sigma \gtrsim \epsilon$ in our setting; when $\sigma \ll \epsilon$, the problem is much closer to the classical agnostic setting.

It turns out that it is possible to develop efficient smoothed agnostic learners for natural concept classes $\mathcal{C}$ with respect to a broad class of distribution families $\mathcal{D}_x$—for which vanilla agnostic learning is computationally hard. The first positive results were given in [CKK+24] (who defined the smoothed framework). The latter work focused on the natural concept class of Multi-Index Models (MIMs)[1] that satisfy an additional mild condition (namely, having bounded Gaussian surface area). The distribution class $\mathcal{D}_x$ consists of all distributions whose univariate projections have subgaussian (or, more generally, strictly sub-exponential) tails. In the following discussion, we focus on the case of subgaussian distributions for concreteness.

In this context, [CKK+24] shows that the L_1-polynomial regression algorithm [KKMS08] is a smoothed agnostic learner for K-MIMs with Gaussian surface area at most $\Gamma > 0$ with (sample and computational) complexity $d^{\mathrm{poly}(K,\Gamma,1/\sigma,1/\epsilon)}$. This upper bound was recently improved in the same context by [KW25], who showed that the same algorithm has complexity $d^{\mathrm{poly}(K,1/\sigma,\log(1/\epsilon))}$ independent of the surface area bound. More precisely, the exponent of d in the upper bound of [KW25] is $\tilde{O}(K/\sigma^2)\log(1/\epsilon)$.

Motivated by these positive results, here we ask the question whether the known upper bounds for smoothed agnostic learning can be qualitatively improved. Interestingly, all known algorithmic works on the topic (including [CKK+24, KW25] and the work of [KM25] studying a Boolean version of smoothing) essentially rely on the L_1 polynomial regression method—the standard agnostic learner in the "non-smoothed" setting. The technical novelty of these works lies in the analysis: they establish appropriate polynomial approximation results that yield their upper bounds.

[1] A function $f : \mathbb{R}^d \to \{\pm 1\}$ is called a K-MIM if there exists a K-dimensional subspace V such that for each $x \in \mathbb{R}^d$ $f(x)$ only depends on the projection of x onto V.

Since all currently known algorithms for the problem ultimately rely on L_1-polynomial regression, a natural goal is to determine whether this dependence on low-degree approximation is inherent in the smoothed setting. More concretely, focusing on the subgaussian case for single-index models ($K = 1$) and halfspaces in particular, the best known upper bound is $d^{\tilde{O}(1/\sigma^2)\log(1/\epsilon)}$. Is there a $\mathrm{poly}(d, 1/\epsilon, 1/\sigma)$ time algorithm? Or is the exponential dependence on $1/\sigma^2$ and $\log(1/\epsilon)$ inherent? Motivated by this gap in our understanding, we study the following question:

What is the complexity of smoothed agnostic learning for halfspaces under subgaussian distributions?

As our main result, we give formal evidence—in the form of a Statistical Query lower bound—that the complexity of known algorithms for this problem is essentially optimal. Thus, for the basic benchmark problem of learning halfspaces, our results largely resolve the computational complexity of smoothed agnostic learning in the SQ model.

Before we formally state our results, we need basic background on the Statistical Query (SQ) Model and the class of halfspaces.

Statistical Query Model Statistical Query (SQ) algorithms are a class of algorithms that are allowed to query expectations of bounded functions of the underlying distribution rather than directly access samples. Formally, an SQ algorithm has access to the following oracle.

Definition 1.2 (STAT Oracle)**.** *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. A statistical query is a function $q : \mathbb{R}^d \times \{\pm 1\} \to [-1, 1]$. We define $\mathrm{STAT}(\tau)$ to be the oracle that given a query $q(\cdot, \cdot)$ outputs a value v such that $|v - \mathbf{E}_{(x,y)\sim D}[q(x, y)]| \leq \tau$, where $\tau > 0$ is the tolerance of the query.*

The SQ model was introduced in [Kea98] as a natural restriction of the PAC model and has been extensively studied in learning theory; see, e.g., [FGR$^+$13, FGV17, Fel17] and references therein. The class of SQ algorithms is fairly broad: a wide range of known algorithmic techniques in machine learning are known to be implementable using SQs (see, e.g., [CKL$^+$06, FGR$^+$13, FGV17]).

Halfspaces A *halfspace* is any Boolean-valued function $f : \mathbb{R}^d \to \{\pm 1\}$ of the form $f(x) = \mathrm{sign}(w \cdot x - t)$, for a weight vector $w \in \mathbb{R}^d$ and a threshold $t \in \mathbb{R}$. The algorithmic task of learning halfspaces from labeled examples is one of the most basic and extensively studied problems in machine learning [Ros58, Nov62, MP68, FS97, Vap98, STC00]. Halfspaces are a prototypical family of Single-Index models. It is known that even *weak* agnostic learning of halfspaces is computationally hard in the distribution-free setting [Dan16]. Moreover, in the distribution-specific setting where the distribution is Gaussian, achieving error $\mathrm{OPT} + \epsilon$ requires SQ complexity $d^{\Omega(1/\epsilon^2)}$ [DKPZ21, DKR23].

1.1 Our Results and Techniques

We are now ready to state our main result:

Theorem 1.3 (Main Result)**.** *There is a universal constant $\epsilon_0 > 0$ such that the following holds. Let $0 < \sigma \leq 1$ and $0 < \epsilon < \epsilon_0$, and let d be sufficiently large in terms of σ and ϵ. Any smoothed agnostic learner for the class $\mathcal{H}$ of halfspaces on $\mathbb{R}^d$ under the standard Gaussian distribution, with excess error ϵ and smoothing parameter σ, requires SQ complexity $d^{\Omega((\sigma+\epsilon)^{-2}+\log(1/\epsilon))}$. More precisely, any SQ algorithm for this task must either make $2^{d^{\Omega(1)}}$ queries or use a query of tolerance less than $d^{-\Omega((\sigma+\epsilon)^{-2}+\log(1/\epsilon))}$.*

Some comments are in order. First note that for $\sigma \geq \epsilon$, our SQ lower bound nearly matches the upper bound of [KW25] which holds for all subgaussian distributions. That is, the special case of the Gaussian distribution seems to capture the hardness of the subgaussian family. The same proof with $\sigma = 0$ recovers the known optimal SQ lower bound of $d^{\Omega(1/\epsilon^2)}$ [DKPZ21] for vanilla agnostic learning under Gaussian marginals. On the other hand, when $0 < \sigma = \Omega(1)$, we obtain an SQ lower bound of $d^{\Omega(\log(1/\epsilon))}$. In particular, the theorem is most informative in the genuinely smoothed regime $\sigma \gtrsim \epsilon$, where one is no longer essentially in the classical agnostic setting.

The fact that our SQ lower bound for smoothed agnostic learning of halfspaces qualitatively matches the upper bound obtained via L_1 polynomial regression is not a coincidence. Roughly speaking, we show the following more general result (Theorem 3.1): for any concept class $\mathcal{C}$ with natural closure properties (including halfspaces and more general MIMs), the degree of L_1 polynomial

approximation to the smoothed function $T_\sigma f(x) := \mathbf{E}_{z\sim\mathcal{N}_d}[f(x+\sigma z)]$ (directly related to the definition of the smoothed optimum) over $f \in \mathcal{C}$ determines the SQ complexity of smoothed agnostic learning. Roughly speaking, if the optimal polynomial approximation degree for $\mathcal{C}$ is m^*, then the SQ complexity of smoothed agnostic learning $\mathcal{C}$ is $d^{\Omega(m^*)}$—matching the complexity of L_1-polynomial regression. In this sense, our generic result shows that the smoothed model is again governed by polynomial approximation, but now for the smoothed target $T_\sigma f$ rather than for f itself.

It is natural to ask whether one can show stronger SQ lower bounds for agnostically learning more general multi-index models in the smoothed setting. Interestingly, we show that this is not possible when the marginal distribution is promised to be Gaussian (as in Theorem 1.3). Specifically, we show in Proposition A.1 (given in Section A) that L_1-polynomial regression is a smoothed agnostic learner for any distribution (X, y) with X Gaussian, with complexity $d^{O(\log(1/\epsilon)/\sigma^2)}$.

Technical Overview The general framework that we leverage to obtain our results follows a line of work on SQ lower bounds for Non-Gaussian Component Analysis (NGCA) and its generalizations, developed in a series of papers over the past decade; see, e.g., [DKS17, DKPZ21, DKRS23, DIKR25]. At a high level, the outer structure of the proof is a generalization of prior SQ lower bounds for (non-smoothed) agnostic learning [DKPZ21]. The main new challenge is to handle the smoothed setting, where the relevant approximation object is the smoothed target $T_\sigma f$.

Specifically, our proof proceeds in two steps: First, we show a generic SQ lower bound (Theorem 3.1) that is applicable to any concept class $\mathcal{C}$ satisfying natural closure properties. In the context of smoothed agnostic learning of halfspaces under the Gaussian distribution, this result shows the following: If f is a univariate threshold function, then any SQ algorithm for the problem has complexity $d^{\Omega(m)}$, where d is the dimension and m is the minimum degree of L_1-polynomial approximation with error $O(\epsilon)$ of the smoothed function $T_\sigma f(x) = \mathbf{E}_{z\sim\mathcal{N}}[f(x+\sigma z)]$. The proof of our generic lower bound follows the overall strategy of [DKPZ21] (which considers the special case of $\sigma = 0$). The main technical novelty is showing that this framework can be instantiated sharply in the smoothed setting, and in particular proving the explicit lower bounds for smoothed univariate threshold functions needed to obtain Theorem 1.3. The basic idea is to construct a moment-matching distribution on y so that $\mathrm{OPT}_\sigma + \epsilon$ is still somewhat less than $1/2$. Thus, any algorithm that produces a hypothesis with error $\mathrm{OPT}_\sigma + \epsilon$ would allow one to distinguish between this y and a y that is independent of x. We then show that this distinguishing (testing) problem is SQ-hard, as it corresponds to a hard instance of conditional NGCA [DIKR25]. To construct the desired moment-matching distribution on y, we use LP duality and show that the relevant dual program corresponds exactly to finding a low-degree L_1 polynomial approximation of $T_\sigma f$.

The second step of our proof involves proving explicit degree lower bounds for $T_\sigma f$, which is our main novel technical contribution. Here we establish a degree lower bound of $c(\log(1/\epsilon)+1/(\sigma+\epsilon)^2)$ for some universal constant $c > 0$[2].

To establish a degree lower bound of $\log(1/\epsilon)$ (Lemma 3.8) we proceed as follows. We start by noting the identity $T_\sigma f = U_a f$, for the function $f(x) = \mathrm{sign}(x)$, where U_a is the Ornstein-Uhlenbeck (OU) operator and $a = 1/\sqrt{1+\sigma^2}$. An explicit analysis, using properties of Hermite polynomials and the OU operator, allows us to show that the L_2 error of any degree-k polynomial approximation to $T_\sigma f$ must be at least $2^{-O(k)}$. In fact, we can even certify this error by noting that the inner product of $T_\sigma f - p$ with an appropriate Hermite polynomial must be large. The next step is to compare the L_1 and L_2 polynomial approximations, which we carry out by using results about the concentration of p and this Hermite polynomial. As a result, we show that the L_1 error of any degree-k polynomial approximation to $T_\sigma f$ must be $2^{-O(k)}$ as well (for a different universal constant hidden in the $O()$.)

To establish the L_1 degree lower bound of $(\sigma+\epsilon)^{-2}$ (Lemma 3.5), we prove the following new structural result (Proposition 3.6) that may be of broader interest: there exists a k-wise independent family of Ck univariate standard Gaussians, for a universal constant C, whose sum modulo 1 lies in any prescribed measurable set $S \subseteq [0,1]$ of positive measure with probability at least $1 - 2^{-k}$. Their sum divided by $\sqrt{Ck}$ matches the first k Gaussian moments while exhibiting periodic structure at scale $1/\sqrt{k}$. By letting $X \mid y = 1$ be such a distribution missing a large fraction of the possible values modulo $1/\sqrt{k}$, it is easy to see that for some threshold f that the $|T_\sigma f(X) - T_\sigma f(G)| \gg \epsilon$

[2]Note that the second term is roughly $1/\sigma^2$, as long as $\sigma \geq \epsilon$ (this is the interesting regime in the smoothed setting). For $\sigma = 0$, degree $O(1/\epsilon^2)$ is known to suffice.

so long as $\epsilon, \sigma \ll 1/\sqrt{k}$. This gives us an example of a moment-matching distribution (X, y) with OPT_σ substantially less than $1/2$, which can be plugged into our generic SQ lower bound.

Regarding our upper bound for the smoothed Gaussian case (Proposition A.1), it suffices to prove an upper bound on the degree of L_2 polynomial approximation of $T_\sigma f$. To do this, we note that $T_\sigma f = Ug$ for $g(x) = f(\sqrt{1+\sigma^2}x)$ and U is the relevant element of the Ornstein-Uhlenbeck semigroup. But if g has Hermite transform $g = \sum g^{[k]}$, then Ug is $\sum g^{[k]}(1+\sigma^2)^{-k/2}$. Since $\|g^{[k]}\|_2 \le \|g\|_2 \le 1$, taking the first $O(\log(1/\epsilon)/\sigma^2)$ terms suffices to give error ϵ.

2 Preliminaries

Notation For $n \in \mathbb{Z}_+$, we denote $[n] \stackrel{\mathrm{def}}{=} \{1, \ldots, n\}$. We typically use small letters to denote random variables when the underlying distribution is clear from the context. We use $\mathbf{E}[x]$ for the expectation of the random variable x and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. Let $\mathcal{N}$ denote the standard univariate Gaussian distribution and $\mathcal{N}_d$ denote the standard d-dimensional Gaussian distribution. We use ϕ_d to denote the pdf of $\mathcal{N}_d$.

Small letters are used for vectors and capital letters are used for matrices. Let $\|x\|_2$ denote the ℓ_2-norm of the vector $x \in \mathbb{R}^d$. We use $u \cdot v$ for the inner product of vectors $u, v \in \mathbb{R}^d$. For a matrix $P \in \mathbb{R}^{m \times n}$, let $\|P\|_2$ denote its spectral norm and $\|P\|_F$ denote its Frobenius norm. We use I_d to denote the $d \times d$ identity matrix.

Hermite Analysis Consider $L_2(\mathbb{R}^d, \mathcal{N}_d)$, the vector space of all functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbf{E}_{x \sim \mathcal{N}_d}[f(x)^2] < \infty$. This is an inner product space under the inner product $\langle f, g \rangle = \mathbf{E}_{x \sim \mathcal{N}_d}[f(x)g(x)]$. This inner product space has a complete orthogonal basis given by the *Hermite polynomials*. In the univariate case, we will work with normalized Hermite polynomials defined below.

Definition 2.1 (Normalized Probabilist's Hermite Polynomial)**.** *For $k \in \mathbb{N}$, the k-th* probabilist's *Hermite polynomial $He_k : \mathbb{R} \to \mathbb{R}$ is defined as $He_k(t) = (-1)^k e^{t^2/2} \cdot \frac{\mathrm{d}^k}{\mathrm{d}t^k} e^{-t^2/2}$. We define the k-th* normalized *probabilist's Hermite polynomial $h_k : \mathbb{R} \to \mathbb{R}$ as $h_k(t) = He_k(t)/\sqrt{k!}$.*

Note that for $G \sim \mathcal{N}$ we have $\mathbf{E}[h_n(G)h_m(G)] = \delta_{n,m}$, and $\sqrt{m+1}h_{m+1}(t) = th_m(t) - h'_m(t)$.

We will use various well-known properties of these polynomials in our proofs.

Smoothing Operators We will also need the following smoothing operators:

- For a function $f : \mathbb{R}^k \to \mathbb{R}$ and $\sigma > 0$, we define the operator $T_\sigma f(x) = \mathbf{E}_{z \sim \mathcal{N}_k}[f(x + \sigma z)]$. This operator is motivated by the definition of our smoothing model.
- For a function $f : \mathbb{R}^k \to \mathbb{R}$ and $\rho \in [0, 1]$, the Ornstein-Uhlenbeck (OU) operator is defined as follows $U_\rho f(x) = \mathbf{E}_{z \sim \mathcal{N}_k}\left[f(\rho x + \sqrt{1-\rho^2}z)\right]$. This operator arises in some of our analysis.

We note that for homogeneous functions these two operators are essentially equivalent up to scaling. Specifically, we have that $T_\sigma f = U_a g$, where $a = 1/\sqrt{1+\sigma^2}$ and $g(x) = f(\sqrt{1+\sigma^2}x)$.

We will use the following standard properties of the OU operator:

Fact 2.2 (See, e.g., [Bog98] and [O'D14] (Chapter 11))**.** *We have the following:*

1. *(Self-adjointness) For $f, g \in L_2(\mathcal{N})$ and $\rho \in (0, 1)$, $\mathbf{E}_{x \sim \mathcal{N}}[(U_\rho f(x))g(x)] = \mathbf{E}_{x \sim \mathcal{N}}[(U_\rho g(x))f(x)]$.*
2. *(Hermite eigenfunctions) For $i \in \mathbb{Z}_+$, $U_\rho h_i(z) = \rho^i h_i(z)$.*

Analytic Facts For a polynomial $p : \mathbb{R}^k \to \mathbb{R}$, we consider the random variable $p(x)$, where $x \sim \mathcal{N}_k$. We will use $\|p\|_r$, for $r \ge 1$, to denote the L_r-norm of the random variable $p(x)$, i.e., $\|p\|_r \stackrel{\mathrm{def}}{=} \mathbf{E}_{x \sim \mathcal{N}_k}[|p(x)|^r]^{1/r}$.

We will need the following analytic and probabilistic facts. We first recall the following moment bound for low-degree polynomials, which is equivalent to the hypercontractive inequality [Bon70, Gro75]:

Fact 2.3 (Hypercontractive Inequality)**.** *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a degree-d polynomial and $q > 2$. Then $\|p\|_q \le (q-1)^{d/2}\|p\|_2$.*

As a simple corollary, we can relate the L_1-norm of a polynomial with the L_2-norm. Specifically, we obtain the following folklore fact:

Fact 2.4. *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a degree-d polynomial. Then $\|p\|_2 \leq 2^{3d/2}\|p\|_1$.*

Proof. Using the hypercontractive inequality for $q = 3$ gives $\|p\|_3 \leq 2^{d/2}\|p\|_2$. By the Cauchy–Schwarz inequality, $\|p\|_2 \leq \|p\|_1^{1/4}\|p\|_3^{3/4}$. Combining the bounds gives the proof. □

3 Main Results

In Section 3.1, we prove our generic SQ lower bound (Theorem 3.1) for smoothed agnostic learning of any concept class $\mathcal{C}$ satisfying appropriate closure properties. Roughly speaking, this result shows that the SQ complexity of the task is $d^{\Omega(m)}$, where m is the optimal degree of L_1-polynomial approximation of a smoothed version of the target functions. In Section 3.2, we prove explicit degree lower bounds for the class of halfspaces, thereby establishing Theorem 1.3.

3.1 Generic SQ Lower Bound

Theorem 3.1 (Generic SQ Lower Bound for Smoothed Agnostic Learning)**.** *Let $\mathcal{C}$ be a class of Boolean functions on $\mathbb{R}^d$ that, for some $f : \mathbb{R}^k \to \{\pm 1\}$, contains all functions $x \mapsto f(Px)$ for linear maps $P : \mathbb{R}^d \to \mathbb{R}^k$ with $PP^\top = I_k$. Suppose that m is a positive even integer and that d is sufficiently large in terms of k and m. Suppose further that, for some $\epsilon, \sigma > 0$, for any degree at most m polynomial $p : \mathbb{R}^k \to \mathbb{R}$*

$$\mathbf{E}_{x\sim\mathcal{N}_k}[|T_\sigma f(x) - p(x)|] > \epsilon + 2\tau \ , \tag{1}$$

where $\tau = c_{k,m}d^{-m/5}$ for a sufficiently small constant $c_{k,m} > 0$. Then any SQ algorithm that σ-smoothed-agnostically learns $\mathcal{C}$ under the standard Gaussian distribution to excess error ϵ must either make $2^{d^{\Omega(1)}}$ queries or a query of accuracy better than τ.

To prove Theorem 3.1, we need the following proposition:

Proposition 3.2. *In the context of Theorem 3.1, there exists a distribution (X, Y) on $\mathbb{R}^k \times \{\pm 1\}$ such that:*

- *(i) The X-marginal is a standard Gaussian $\mathcal{N}_k$.*
- *(ii) The expectation of Y is 0.*
- *(iii) The conditional distributions $X \mid Y = 1$ and $X \mid Y = -1$ each match their first m moments with the standard Gaussian $\mathcal{N}_k$.*
- *(iv) If Z is an independent standard Gaussian on $\mathbb{R}^k$, then $\mathbf{Pr}[f(X + \sigma Z) \neq Y] < 1/2 - \epsilon - 2\tau$.*

We start by showing how Proposition 3.2 can be used to prove Theorem 3.1.

Proof of Theorem 3.1. We will use the distribution (X, Y), whose existence is established in Proposition 3.2, to produce a conditional non-Gaussian component analysis construction [DIKR25]. We begin by defining a hard ensemble. In particular, for a random projection $P : \mathbb{R}^d \to \mathbb{R}^k$, we define the distribution on $\mathbb{R}^d \times \{\pm 1\}$ as $(P^\top(X) + X', Y)$, where (X, Y) is the distribution from Proposition 3.2 and X' is an independent standard Gaussian supported on the subspace $V^\perp$ where V is the image of $P^\top$. Notice that this gives an instance of a Relativized Hidden Subspace Distribution for $A = (X, Y)$ and $U = P^\top$ from Definition 3.1 of [DIKR25].

Furthermore, as $X|Y = 1$ and $X|Y = -1$ both match m moments with a standard Gaussian, applying Theorem 3.3 of [DIKR25] with $\lambda = 1/20$, $c = 1/80$, and sufficiently large d, implies that any Statistical Query algorithm distinguishing (x, Y) (with randomly chosen P) from (x, Y') (where Y' is an independent, uniform $\{\pm 1\}$ random variable) must either use a query of accuracy better than τ or use $2^{d^{\Omega(1)}}$ queries.

On the other hand, we claim that given a smoothed-agnostic learner for $\mathcal{C}$ with excess error ϵ, an additional statistical query of tolerance τ would allow us to correctly solve this distinguishing problem.

This is because the learned hypothesis $h : \mathbb{R}^d \to \{\pm 1\}$ would have the property that

$$\begin{aligned}\mathbf{Pr}[h(x) \neq Y] &\leq \mathrm{OPT}_\sigma + \epsilon \\ &\leq \mathbf{Pr}[f(P(x+\sigma z)) \neq Y] + \epsilon \\ &= \mathbf{Pr}[f(X+\sigma Z) \neq Y] + \epsilon < 1/2 - 2\tau \ .\end{aligned}$$

On the other hand, since Y' is independent of x, we have $\mathbf{Pr}[h(x) \neq Y'] = 1/2$. As a single statistical query can distinguish between $\mathbf{Pr}[h(x) \neq Y]$ being $1/2$ or less than $1/2 - 2\tau$, this completes the reduction and the proof of Theorem 3.1. □

To prove Proposition 3.2, we use linear programming duality techniques, similar to those from [DKPZ21] (which corresponds to the $\sigma = 0$ special case of our proof below).

Proof of Proposition 3.2. To begin with, we note that finding a distribution (X, Y) is equivalent to specifying the function $g(x) = \mathbf{E}[Y \mid X = x]$. In particular, note that g must be a function from $\mathbb{R}^k$ to $[-1, 1]$. We now translate the desired properties of (X, Y) to corresponding properties of the function g. We start with the following simple claim:

Claim 3.3. *Conditions (ii) and (iii) on (X, Y) together are equivalent to the condition that g satisfies $\mathbf{E}[g(X)p(X)] = 0$ for any degree at most m polynomial $p : \mathbb{R}^k \to \mathbb{R}$.*

Proof. One direction of the equivalence is straightforward. Suppose that conditions (ii) and (iii) on (X, Y) are satisfied. Since $g(x) = \mathbf{E}[Y \mid X = x]$, it follows that for any degree at most m polynomial p we have that

$$\begin{aligned}\mathbf{E}[g(X)p(X)] &= \mathbf{E}[Yp(X)] \\ &= \mathbf{Pr}(Y=1)\mathbf{E}[p(X)|Y=1] - \mathbf{Pr}(Y=-1)\mathbf{E}[p(X)|Y=-1] \\ &= (\mathbf{E}[p(X)|Y=1] - \mathbf{E}[p(X)|Y=-1])/2 = 0.\end{aligned}$$

For the opposite direction, suppose that $\mathbf{E}[g(X)p(X)] = 0$ for any degree at most m polynomial $p : \mathbb{R}^k \to \mathbb{R}$. Setting $p \equiv 1$ gives condition (ii). Note that the latter implies that $\mathbf{Pr}[Y = 1] = \mathbf{Pr}[Y = -1] = 1/2$. Given this, we have that for any degree at most m polynomial p, it holds

$$\mathbf{E}[p(X) \mid Y = 1] = \frac{\mathbf{E}[p(X)\mathbf{1}\{Y=1\}]}{\mathbf{Pr}[Y=1]} = 2\,\mathbf{E}[p(X)\mathbf{1}\{Y=1\}] = \mathbf{E}[p(X)(g(X)+1)] = \mathbf{E}[p(X)] \ .$$

An identical argument can be used for $\mathbf{E}[p(X) \mid Y = -1]$, giving (iii). This completes the proof of Claim 3.3. □

We finally translate condition (iv). We can write

$$\begin{aligned}\mathbf{Pr}[f(X+\sigma Z) \neq Y] &= (1 - \mathbf{E}[f(X+\sigma Z)Y])/2 \\ &= (1 - \mathbf{E}[(T_\sigma f)(X)Y])/2 \\ &= (1 - \mathbf{E}[(T_\sigma f)(X)g(X)])/2 \ .\end{aligned}$$

Thus, condition (iv) is equivalent to the condition

$$\mathbf{E}[T_\sigma f(X)g(X)] > \epsilon + 2\tau \ . \tag{2}$$

In summary, to prove the proposition, it suffices to find a function $g : \mathbb{R}^k \to \mathbb{R}$ such that

(A) $\|g\|_\infty \leq 1$.

(B) $\mathbf{E}_{X \sim \mathcal{N}_k}[g(X)p(X)] = 0$ for all degree at most m polynomials p.

(C) $\mathbf{E}_{X \sim \mathcal{N}_k}[(T_\sigma f)(X)g(X)] > \epsilon + 2\tau$.

We note that constraints (A)-(C) define an infinite linear program. By a generalization of standard LP duality to this setting, this linear program is feasible if and only if the corresponding dual program is infeasible. This would follow immediately from the standard LP duality result were it a finite linear program. In our case, we need to use some somewhat more complicated analytic tools, and leverage Corollary 55 of [DKPZ21]. This asks if there is a linear combination of constraints that

reach a contradiction. In particular, the constraints in part (A) combine to form constraints of the form $\mathbf{E}[g(X)a(X)] \leq \|a\|_1$ for all a. The constraints in part (B) in general are of the form $\mathbf{E}[g(X)p(X)] = 0$. The constraints in part (C) are just multiples of the single constraint that $\mathbf{E}[T_\sigma f(X)g(X)] > \epsilon + 2\tau$.

Thus, a solution to the dual program must consist of a function a, a polynomial p and a negative real number c such that

$$a(X) + p(X) + c\,T_\sigma f(X)$$

is identically 0 and

$$\|a\|_1 + c(\epsilon + 2\tau) < 0\ .$$

By homogeneity, we can take $c = -1$, so $a(X) = T_\sigma f(X) - p(X)$ and $\|a\|_1 < \epsilon + 2\tau$. In particular, there is a solution to the dual program if and only if there is a degree at most m polynomial p such that for $\|T_\sigma f(X) - p(X)\|_1 < \epsilon + 2\tau$. However, this is ruled out by assumption. Therefore, our primal LP must have a solution, proving Proposition 3.2. □

3.2 Smoothed Agnostic Learning of Halfspaces: Proof of Theorem 1.3

Our main application is for smoothed agnostic learning of halfspaces, when the distribution D_x is the standard Gaussian. This corresponds to the case $k = 1$ and a suitable univariate threshold $f(x) = \mathrm{sign}(x - t)$ in Theorem 3.1.

Theorem 1.3 follows directly from Theorem 3.1 and the following result, which is established in this section.

Theorem 3.4 (Smoothed L_1-Degree Lower Bound for Threshold Functions)**.** *There are universal constants $c, c_0 > 0$ such that, for every $0 < \sigma \leq 1$ and $0 < \epsilon < c_0$, there exists $t \in [0, 1]$ such that, for $f(x) = \mathrm{sign}(x - t)$, every polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree at most cm, where $m = (\sigma + \epsilon)^{-2} + \log(1/\epsilon)$, satisfies $\mathbf{E}_{x\sim\mathcal{N}}[|T_\sigma f(x) - p(x)|] > \epsilon$.*

The theorem follows from the two lemmas below, since the maximum of their two degree lower bounds is at least half their sum.

We start by proving a degree lower bound of $(\sigma + \epsilon)^{-2}$.

Lemma 3.5. *Theorem 3.4 holds instead taking $m = (\sigma + \epsilon)^{-2}$.*

The key new technical ingredient that we require is in the following proposition. To the best of our knowledge, this moment-matching construction is new and may find other applications.

Proposition 3.6. *For $k > 0$ and any subset S of $[0, 1]$, there exists a distribution X on $\mathbb{R}$ so that:*

- *(A) X matches its first k moments with the standard Gaussian.*
- *(B) $X(x)/\phi(x) \leq (1/|S|)$, where $|S|$ denotes the measure of S.*
- *(C) For some constant $C > 0$, all but 2^{-k} of the probability mass of X is supported on points x so that $\mathrm{FractionalPart}(x(\sqrt{Ck})) \in S$.*

Proof. Let C be a large enough integer. We will let $X = 1/\sqrt{Ck}\sum_{i=1}^{Ck} X_i$, where the X_i's are non-independent Gaussians. In particular, you can write the standard Gaussian G as a mixture of $U([0, 1])$ with some other distribution E. In particular, $G = \alpha U([0, 1]) + (1 - \alpha)E$. We sample from X in the following way:

(1) Sample y_i i.i.d. from $\{0, 1\}$ with $y_i = 1$ with probability α.

(2) For each $y_i = 0$, set X_i to be an independent sample from E.

(3) If $\sum y_i < k + 1$, set each X_i with $y_i = 1$ to be an independent sample from $U([0, 1])$.

(4) Otherwise, set the X_i's with $y_i = 1$ to be independent samples from $U([0, 1])$ conditioned on the sum of all X_i's mod 1 lying in S.

(5) Let $X = 1/\sqrt{Ck}\sum_{i=1}^{Ck} X_i$.

To analyze this, first note that if the conditioning in step (4) is omitted, the X_i's are independent standard Gaussians, and hence X is standard Gaussian.

To prove (A), condition on the y_i's and on the samples in step (2). In step (4), any collection of at most k of the remaining X_i's is still distributed as independent $U([0,1])$ variables: at least one unobserved uniform variable makes the conditioning event have probability $|S|$, independently of that collection. Consequently, all joint moments of total degree at most k agree with those of the unconditioned construction, and X matches the first k Gaussian moments. For (B), we note that we obtain X by taking a sampling process that generates a standard Gaussian and in step (4) conditioning on an event of probability $|S|$. Finally to prove (C), we note that if $\sum_i y_i \geq k+1$, then $\mathrm{FractionalPart}(X\sqrt{Ck}) \in S$. Since $\sum_i y_i$ has distribution $\mathrm{Bin}(Ck, \alpha)$, this happens with probability at least $1-2^{-k}$, when C is large enough. This completes the proof of Proposition 3.6. $\square$

Remark 3.7. Conditions (B) and (C) in the statement of Proposition 3.6 can be replaced by the statement that X is $2^{-\Omega(k)}$ close in total variation distance to the standard Gaussian conditioned on the fact that $\mathrm{FractionalPart}(x\sqrt{Ck}) \in S$. This statement may be useful for other applications.

Proof of Lemma 3.5. We may assume that $\sigma+\epsilon$ is sufficiently small. Otherwise, after reducing the universal degree constant, only constant polynomials need be considered, and the best constant approximation to $T_\sigma \mathrm{sign}$ has L_1 error at least $1/2$ for $0 < \sigma \leq 1$.

Given Proposition 3.6, let $k = \lfloor c_1(\sigma+\epsilon)^{-2} \rfloor$ for a sufficiently small universal constant $c_1 > 0$, let $S = [1/2, 1]$ and consider the corresponding distribution X. Let $t = 0$ and let $t' = 1/(2\sqrt{Ck})$ with C as in Proposition 3.6. Let $f(x) = \mathrm{sign}(x-t)$ and $f'(x) = \mathrm{sign}(x-t')$. Let $g = T_\sigma f$ and $g' = T_\sigma f'$. We note that t and t' differ by $\Omega(k^{-1/2})$ which is a large constant times $(\sigma+\epsilon)$. Note that $g(x) - g'(x)$ is always non-negative and it is at least $1/2$ when $x \in [t'/3, 2t'/3]$. Thus, $\mathbf{E}_{G\sim\mathcal{N}}[g(G)] - \mathbf{E}_{G\sim\mathcal{N}}[g'(G)]$ is at least a constant times the probability that G lies in $[t'/3, 2t'/3]$, which is at least a large constant times $(\sigma+\epsilon)$.

On the other hand, we claim that $\mathbf{E}[g(X)] - \mathbf{E}[g'(X)] = O(\sigma + 2^{-k})$. Indeed, the contribution from $X \in [t, t']$ is at most $2\mathbf{Pr}[X \in [t,t']] \leq 2^{1-k}$. Outside this interval, property (B) bounds the contribution by $2\mathbf{E}_{G\sim\mathcal{N}}[|g(G) - g'(G)|\mathbf{1}\{G \notin [t,t']\}]$, which is $O(\sigma)$.

Since X is bounded by $2G$, where $G \sim \mathcal{N}$, (by property (B)) and since it matches its first k moments with a Gaussian (by property (A)), if p is a degree at most k polynomial with $\|g-p\|_1 < \delta$, then

$$\begin{aligned}
\mathbf{E}[g(X)] &= \mathbf{E}[p(X)] + \mathbf{E}[(g-p)(X)] \\
&= \mathbf{E}[p(G)] + O(\delta) \\
&= \mathbf{E}[g(G)] - \mathbf{E}[(g-p)(G)] + O(\delta) \\
&= \mathbf{E}[g(G)] + O(\delta) \ .
\end{aligned}$$

A similar bound holds for g'.

For sufficiently small $\sigma+\epsilon$, our choice of k gives $2^{-k} = o(\sigma+\epsilon)$. However, if such approximations exist for both g and g', we have that

$$\mathbf{E}_{G\sim\mathcal{N}}[g(G) - g'(G)] = \mathbf{E}[g(X) - g'(X)] + O(\delta) \ ,$$

which by the above is a contradiction unless δ is at least a large constant times $\sigma+\epsilon$. This shows that there is a T_σ LTF that is not well approximable by degree k polynomials. This completes the proof of Lemma 3.5. $\square$

Lemma 3.8. *Theorem 3.4 holds instead taking $m = \log(1/\epsilon)$, uniformly for $0 < \sigma \leq 1$.*

Proof. Note that for $f(x) = \mathrm{sign}(x)$, $T_\sigma f = U_a f$, where U is the Ornstein-Uhlenbeck operator and $a = 1/\sqrt{1+\sigma^2}$, which is $\Omega(1)$ if $\sigma = O(1)$. For p a polynomial of degree less than k, with k odd, and h_k the degree-k normalized probabilist's Hermite polynomial, we have that

$$\begin{aligned}
\mathbf{E}_{G\sim\mathcal{N}}[h_k(G)(U_a f - p)(G)] &= \mathbf{E}_{G\sim\mathcal{N}}[h_k(G) U_a f(G)] \\
&= \mathbf{E}_{G\sim\mathcal{N}}[U_a h_k(G) f(G)] \\
&= a^k \mathbf{E}_{G\sim\mathcal{N}}[h_k(G) f(G)] \\
&= 2^{-O(k)} \mathbf{E}_{G\sim\mathcal{N}}[h_k(G) f(G)] \ ,
\end{aligned}$$

where the first equality uses the orthogonality of Hermite polynomials, the second equality is by self-adjointness of the OU operator, the third equality uses Mehler's formula (namely that

$U_a h_k(x) = a^k h_k(x)$), and the fourth follows from the fact that $1 > a = \Omega(1)$. Since k is odd, $|\mathbf{E}_{G\sim\mathcal{N}}[h_k(G)f(G)]| = \Omega(1/\mathrm{poly}(k))$, and we get that

$$|\mathbf{E}_{G\sim\mathcal{N}}[h_k(G)(U_a f - p)(G)]| = 2^{-O(k)} \ . \tag{3}$$

Suppose that the polynomial p satisfies $\|U_a f - p\|_1 < \epsilon < 1$. Then we have that $\|p\|_1 < 2$. By Fact 2.4, this implies that $\|p\|_2 \leq 2^{3k/2}\|p\|_1 < 2^{3k/2+1}$. Thus, by hypercontractivity (Fact 2.3), we get that

$$\|U_a f - p\|_4 \leq \|U_a f\|_4 + \|p\|_4 = 2^{O(k)} \ .$$

On the other hand, since $|\mathbf{E}[h_k(G)(U_a f(G) - p(G))]| > 2^{-O(k)}$, this implies that

$$\|U_a f - p\|_2 > 2^{-O(k)} \ .$$

Combined with the above, the Paley-Zygmund inequality implies that $|U_a f(x) - p(x)| > 2^{-O(k)}$ with probability at least $1/2^{O(k)}$ over $x \sim \mathcal{N}$, which implies $\|U_a f - p\|_1 > 2^{-O(k)}$.

Thus, if $\|U_a f - p\|_1 < \epsilon$, we must have $k \gg \log(1/\epsilon)$, completing the proof. □

4 Conclusions

In this work, we obtained the first SQ lower bounds for smoothed agnostic learning. For the benchmark problem of learning halfspaces, these results largely settle the SQ complexity of smoothed agnostic learning under subgaussian distributions. Our lower bounds can be naturally extended to various other families of single-index models. A number of open questions suggest themselves. Since a fully-polynomial time algorithm with optimal error may be unachievable in the smoothed setting, it would be of interest to develop $\mathrm{poly}(d, 1/\sigma, 1/\epsilon)$-time smoothed agnostic learners with *approximate* error guarantees, e.g., error $O(\mathrm{OPT}_\sigma) + \epsilon$ (or give evidence of hardness even for such a weaker guarantee). In a different direction, we note that previous upper bounds require that the distribution on examples has subexponential tails (or better). Is this inherent? Namely, can we prove strong computational lower bounds when the input distribution is heavy-tailed?

References

[ABL17] P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6):50:1–50:27, 2017.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[Bog98] V. Bogachev. *Gaussian measures*. Mathematical surveys and monographs, vol. 62, 1998.

[Bon70] A. Bonami. Etude des coefficients fourier des fonctiones de $l^p(g)$. *Ann. Inst. Fourier (Grenoble)*, 20(2):335–402, 1970.

[CKK+24] G. Chandrasekaran, A. R. Klivans, V. Kontonis, R. Meka, and K. Stavropoulos. Smoothed analysis for learning concepts with low intrinsic dimension. In *The Thirty Seventh Annual Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 876–922. PMLR, 2024.

[CKL+06] C.-T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. Bradski, A. Y. Ng, and K. Olukotun. Map-reduce for machine learning on multicore. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 281–288, Cambridge, MA, USA, 2006. MIT Press.

[Dan16] A. Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the 48th Annual Symposium on Theory of Computing, STOC 2016*, pages 105–117, 2016.

[DIKR25] I. Diakonikolas, G. Iakovidis, D. M. Kane, and L. Ren. Algorithms and SQ lower bounds for robustly learning real-valued multi-index models. *CoRR*, abs/2505.21475, 2025. Conference version in NeurIPS'25.

[DKK+21] I. Diakonikolas, D. M. Kane, V. Kontonis, C. Tzamos, and N. Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1522–1551. PMLR, 2021.

[DKMR22] I. Diakonikolas, D. Kane, P. Manurangsi, and L. Ren. Cryptographic hardness of learning halfspaces with Massart noise. *Advances in Neural Information Processing Systems*, 35:3624–3636, 2022.

[DKPZ21] I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. The optimality of polynomial regression for agnostic learning under gaussian marginals in the SQ model. In *Conference on Learning Theory, COLT 2021*, volume 134 of *Proceedings of Machine Learning Research*, pages 1552–1584. PMLR, 2021.

[DKR23] I. Diakonikolas, D. Kane, and L. Ren. Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 7922–7938. PMLR, 2023.

[DKRS23] I. Diakonikolas, D. Kane, L. Ren, and Y. Sun. SQ lower bounds for non-gaussian component analysis with weaker assumptions. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023.

[DKS17] I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 73–84, 2017. Full version at http://arxiv.org/abs/1611.03473.

[DKS18] I. Diakonikolas, D. M. Kane, and A. Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 1061–1073, 2018.

[DKTZ20] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Non-convex SGD learns halfspaces with adversarial label noise. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020.

[DKTZ22] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. Learning general halfspaces with adversarial label noise via online gradient descent. In *International Conference on Machine Learning, ICML 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 5118–5141. PMLR, 2022.

[DKZ20] I. Diakonikolas, D. Kane, and N. Zarifis. Near-optimal SQ lower bounds for agnostically learning halfspaces and relus under gaussian marginals. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[Fel17] V. Feldman. A general characterization of the statistical query complexity. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 785–830. PMLR, 2017.

[FGR+13] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of STOC'13*, pages 655–664, 2013. Full version in Journal of the ACM, 2017.

[FGV17] V. Feldman, C. Guzman, and S. S. Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In Philip N. Klein, editor, *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017*, pages 1265–1277. SIAM, 2017.

[FS97] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[GGK20] S. Goel, A. Gollakota, and A. R. Klivans. Statistical-query lower bounds via functional gradients. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

[Gro75] L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

[Hau92] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[Kea98] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[KKMS08] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KM25] Y. Kou and R. Meka. Smoothed agnostic learning of halfspaces over the hypercube. *arXiv*, November 2025. Conference version in NeurIPS'25.

[KOS08] A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via Gaussian surface area. In *FOCS 2008*, pages 541–550, 2008.

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

[KW25] F. Koehler and B. Wu. Constructive approximation under carleman's condition, with applications to smoothed analysis. *arXiv*, December 2025.

[MP68] M. Minsky and S. Papert. *Perceptrons: an introduction to computational geometry*. MIT Press, Cambridge, MA, 1968.

[Nov62] A. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

[O'D14] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[Ros58] F. Rosenblatt. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958.

[ST04] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM*, 51(3):385–463, 2004.

[STC00] J. Shawe-Taylor and N. Cristianini. *An introduction to support vector machines*. Cambridge University Press, 2000.

[Tie23] S. Tiegel. Hardness of agnostically learning halfspaces from worst-case lattice problems. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 3029–3064. PMLR, 2023.

[Val84] L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, 1998.

Supplementary Material

A Upper Bound for Smoothed Agnostic Learning under the Gaussian Distribution

In this section, we prove the following simple proposition, ruling out stronger SQ lower bounds for smoothed agnostic learning of any concept class under Gaussian marginals.

Proposition A.1. *Let (X, y) be any distribution on $\mathbb{R}^d \times \{\pm 1\}$ so that X is distributed as a standard Gaussian, and let $1/2 > \sigma, \epsilon > 0$. Then there is an algorithm that learns y to 0-1 error $\mathrm{OPT}_\sigma + \epsilon$ in $d^{O(\log(1/\epsilon)/\sigma^2)}$ samples and time.*

Proof. We will simply use the L_1-polynomial regression algorithm. For this, it suffices to prove that for some $m = O(\log(1/\epsilon)/\sigma^2)$ that there is a degree-m polynomial p so that $\|p - T_\sigma f\|_1 < \epsilon/2$ where $f(x) = \mathbf{E}[y|X = x]$. In fact, we will show that if p is the degree-m part of $T_\sigma f$ then $\|p - T_\sigma f\|_2 < \epsilon/2$.

We prove this using Hermite analysis. In particular, we note that $T_\sigma f = U_a g$ where $a = 1/\sqrt{1+\sigma^2} = 1 - \Omega(\sigma^2)$ and $g(x) = f(x/a)$. If we write g in terms of its Hermite expansion, we find that $g(x) = \sum_{k=0}^\infty g^{[k]}(x)$, where $g^{[k]}$ is the degree-k part of the Hermite expansion. We have that

$$\sum_{k=0}^\infty \|g^{[k]}(x)\|_2^2 = \|g\|_2^2 \leq 1.$$

We also have that

$$T_\sigma f = U_a g = \sum_{k=0}^\infty a^k g^{[k]}.$$

Letting $p = \sum_{k=0}^m a^k g^{[k]}(x)$, we have that

$$T_\sigma f - p = \sum_{k=m+1}^\infty a^k g^{[k]}.$$

So

$$\|T_\sigma f - p\|_2^2 = \sum_{k=m+1}^\infty a^{2k} \|g^{[k]}\|_2^2 \leq a^{2m}.$$

Since $a = 1 - \Omega(\sigma^2)$, if m is at least a sufficiently large constant multiple of $\log(1/\epsilon)/\sigma^2$ this is less than $\epsilon/2$. This completes the proof. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main result of the paper is a near-optimal Statistical Query lower bound for smooth agnostic learning of halfspaces under the Gaussian distribution, as stated in the abstract and established in the body of the paper. The introduction describes how this contribution resolves an open problem in the literature by summarizing the motivation for the smoothed agnostic model and describing prior work's contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are clearly stated in the statement of each theorem and are discussed in the introduction of the paper. Moreover, in the conclusions paragraph, we outline the most closely related open questions not answered by the techniques of this paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theorem statement provides the precise assumptions under which the statement holds. We provide a complete proof of all statements either in the main body of the paper or in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work is theoretical and we do not see any major or immediate implications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work is theoretical in nature and poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This work does not use any assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This work does not use any assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve any crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This work does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.