The Optimality of Polynomial Regression for Agnostic Learning under Gaussian Marginals in the SQ Model

Ilias Diakonikolas ILIAS@CS.WISC.EDU
University of Wisconsin Madison

Daniel M. Kane DAKANE@CS.UCSD.EDU
University of California, San Diego

Thanasis Pittas PITTAS@WISC.EDU
University of Wisconsin Madison

Nikos Zarifis ZARIFIS@WISC.EDU
University of Wisconsin Madison

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We study the problem of agnostic learning under the Gaussian distribution in the Statistical Query (SQ) model. We develop a method for finding hard families of examples for a wide range of concept classes by using LP duality. For Boolean-valued concept classes, we show that the L^1-polynomial regression algorithm is essentially best possible among SQ algorithms, and therefore that the SQ complexity of agnostic learning is closely related to the polynomial degree required to approximate any function from the concept class in L^1-norm. Using this characterization along with additional analytic tools, we obtain explicit optimal SQ lower bounds for agnostically learning linear threshold functions and the first non-trivial explicit SQ lower bounds for polynomial threshold functions and intersections of halfspaces. We also develop an analogous theory for agnostically learning real-valued functions, and as an application prove near-optimal SQ lower bounds for agnostically learning ReLUs and sigmoids.

Keywords: agnostic PAC learning, SQ lower bounds, Gaussian distribution

1. Introduction

1.1. Background and Motivation

In Valiant's Probably Approximately Correct (PAC) learning model Valiant (1984), a learner is given access to random examples that are consistently labeled according to an unknown function in the target concept class. Here we focus on the *agnostic framework* Haussler (1992); Kearns et al. (1994), which models learning in the presence of worst-case noise. Roughly speaking, in the agnostic PAC model, we are given i.i.d. samples from a joint distribution D on labeled examples $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^n$ is the example and $y \in \mathbb{R}$ is the corresponding label, and the goal is to compute a hypothesis that is competitive with the "best-fitting" function in the target class $\mathcal{C}$. The notion of agnostic learning is meaningful both for learning Boolean-valued functions (under the 0-1 loss) and for learning real-valued functions (typically, under the L^2-loss). For concreteness, we restrict the proceeding discussion to the Boolean-valued setting.

In the distribution-independent setting, agnostic learning is known to be computationally hard, even for simple concept classes and weak learning Guruswami and Raghavendra (2006); Feldman et al. (2006); Daniely (2016). On the other hand, under distributional assumptions, efficient learning algorithms with worst-case noise are possible. A line of work Kalai et al. (2008); Klivans et al. (2009); Awasthi et al. (2017); Daniely (2015); Diakonikolas et al. (2018, 2020d) has given efficient learning algorithms in the agnostic model for natural concept classes and distributions with various time-accuracy tradeoffs. In this paper, we will focus on agnostic learning under the Gaussian distribution on examples. For Boolean-valued concept classes, we have the following definition.

Definition 1 (Agnostic Learning Boolean-valued Functions with Gaussian Marginals) *Let $\mathcal{C}$ be a class of Boolean-valued concepts on $\mathbb{R}^n$. Given i.i.d. samples $(\mathbf{x}, y)$ from a distribution D on $\mathbb{R}^n \times \{\pm 1\}$, where the marginal $D_{\mathbf{x}}$ on $\mathbb{R}^n$ is the standard Gaussian $\mathcal{N}_n$ and no assumptions are made on the labels y, the goal is to output a hypothesis $h : \mathbb{R}^n \to \{\pm 1\}$ such that with high probability we have $\mathbf{Pr}_{(\mathbf{x},y)\sim D}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \inf_{f \in \mathcal{C}} \mathbf{Pr}_{(\mathbf{x},y)\sim D}[f(\mathbf{x}) \neq y]$.*

The only known algorithmic technique for agnostic learning in the setting of Definition 1 is the L^1-polynomial regression algorithm Kalai et al. (2008). This algorithm uses linear programming to compute a low-degree polynomial that minimizes the L^1-distance to the target function. Its performance hinges on how well the underlying concept class $\mathcal{C}$ can be approximated, in L^1-norm, by low-degree polynomials. In more detail, if d is the (minimum) degree such that any $f \in \mathcal{C}$ can be ϵ-approximated (in L^1-norm) by a degree-d polynomial, the algorithm has sample complexity and running time $n^{O(d)}/\mathrm{poly}(\epsilon)$ and outputs a hypothesis with missclassification error $\mathrm{OPT} + \epsilon$.

For several natural concept classes and distributions on examples, the aforementioned degree d is independent of the dimension n, and only depends on the error ϵ (and potentially other size parameters). For these settings, the L^1-regression algorithm can be viewed as a polynomial-time approximation scheme (PTAS) for agnostic learning. Examples of such concept classes include Linear Threshold Functions (LTFs) Kalai et al. (2008); Diakonikolas et al. (2010a,c), Bounded Degree Polynomial Threshold Functions (PTFs) Diakonikolas et al. (2010b); Kane (2010); Diakonikolas et al. (2014); Harsha et al. (2014), Intersections of Halfspaces Kalai et al. (2008); Klivans et al. (2008); Kane (2014), and other geometric concepts Klivans et al. (2008). Specifically, for the class of LTFs under the Gaussian distribution, the L^1-regression algorithm is known to have sample and computational complexity of $n^{O(1/\epsilon^2)}$.

For each of the above concept classes, L^1-polynomial regression is the fastest (and, essentially, the only) known agnostic learner. It is natural to ask whether there exists an agnostic learner with significantly improved sample/computational complexity.

Can we beat L^1-polynomial regression for agnostic learning under Gaussian marginals?

As our first main contribution, we answer the above question in the negative for all concept classes satisfying some mild properties (including all the geometric concept classes mentioned above). Our lower bound applies for the class of Statistical Query (SQ) algorithms. 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 2 (STAT Oracle) *Let D be a distribution on labeled examples supported on $X \times [-1, 1]$, for some domain X. A statistical query is a function $q : X \times [-1, 1] \to [-1, 1]$. We define $\mathrm{STAT}(\tau)$ to be the oracle that given any such query $q(\cdot, \cdot)$ outputs a value v such that $|v - \mathbf{E}_{(\mathbf{x},y)\sim D}[q(\mathbf{x}, y)]| \leq \tau$, where $\tau > 0$ is the tolerance parameter of the query.*

The SQ model was introduced by Kearns (1998) as a natural restriction of the PAC model Valiant (1984) and has been extensively studied in learning theory; see, e.g., Feldman et al. (2013, 2015, 2017); Feldman (2017) for some recent references. The reader is referred to Feldman (2016) for a survey. 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., Chu et al. (2006); Feldman et al. (2013, 2017)).

Returning to our agnostic learning setting, roughly speaking, *we show that a lower bound of d on the degree of any L^1 approximating polynomial can be translated to an SQ lower bound of $n^{\Omega(d)}$ for the agnostic learning problem.* This lower bound is tight, since the L^1-regression algorithm can be implemented in the SQ model with complexity $n^{O(d)}$.

We note that a similar characterization had been previously shown, under somewhat different assumptions, for agnostic learning under the uniform distribution on the hypercube Dachman-Soled et al. (2015). We explain the technical differences and similarities with our results in Section 1.4. It is worth pointing out that learning under the Gaussian distribution is generally believed to be computationally easier than learning under the uniform distribution on the hypercube in a number of settings. For example, prior work Awasthi et al. (2017); Diakonikolas et al. (2018, 2020d) has given "constant factor" agnostic learners for LTFs on $\mathbb{R}^n$ under the Gaussian distribution — i.e., algorithms with error $O(\mathrm{OPT})+\epsilon$ — that run in $\mathrm{poly}(n/\epsilon)$ time. No polynomial time algorithm with such an error guarantee is known for any discrete distribution. At a high-level, known algorithms for these problems make essential use of the *anti-concentration* of the Gaussian distribution, which fails in the discrete setting. Similar algorithmic gaps exist for robustly learning low-degree PTFs and intersections of halfspaces Diakonikolas et al. (2018).

Our generic lower bound result for the Boolean case (Theorem 4) reduces the problem of proving explicit SQ lower bounds for agnostic learning to the structural question of proving lower bounds on the L^1 polynomial approximation degree (under the Gaussian measure). As our second contribution, we provide a toolkit to prove explicit degree lower bounds. As a corollary, we prove optimal or near-optimal SQ lower bounds for various natural classes, including LTFs, PTFs, and intersections of halfspaces.

Moving away from the Boolean-valued setting, an interesting direction is to understand the complexity of agnostic learning for real-valued function classes. In recent years, this broad question has been intensely investigated in learning theory, in part due to its connections to deep learning. Here we focus on agnostic learning under the L^2-loss.

Definition 3 (Agnostic Learning Real-valued Functions with Gaussian Marginals) *Let $\mathcal{C}$ be a class of real-valued concepts on $\mathbb{R}^n$. Given i.i.d. samples $(\mathbf{x}, y)$ from a distribution D on $\mathbb{R}^n \times \mathbb{R}$, where the marginal $D_{\mathbf{x}}$ on $\mathbb{R}^n$ is the standard Gaussian $\mathcal{N}_n$ and no assumptions are made on the labels y, the goal is to output a hypothesis $h : \mathbb{R}^n \to \mathbb{R}$ such that with high probability we have $\mathbf{E}_{(\mathbf{x},y)\sim D}[(h(\mathbf{x}) - y)^2]^{1/2} \leq \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \inf_{f\in\mathcal{C}} \mathbf{E}_{(\mathbf{x},y)\sim D}[(f(\mathbf{x}) - y)^2]^{1/2}$.*

A prototypical concept class of significant recent interest are Rectified Linear Units (ReLUs). A ReLU is any real-valued function $f : \mathbb{R}^n \to \mathbb{R}_+$ of the form $f(\mathbf{x}) = \mathrm{ReLU}(\langle \mathbf{w}, \mathbf{x}\rangle + \theta)$, $\mathbf{w} \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$, where $\mathrm{ReLU} : \mathbb{R} \to \mathbb{R}_+$ is defined as $\mathrm{ReLU}(u) = \max\{0, u\}$. ReLUs are the most commonly used activation functions in modern deep neural networks. The corresponding agnostic learning problem is a fundamental primitive in the theory of neural networks that has been extensively studied in recent years Goel et al. (2017); Manurangsi and Reichman (2018); Goel et al. (2019); Diakonikolas et al. (2020a); Goel et al. (2020b); Diakonikolas et al. (2020c).

Our techniques extend to real-valued concepts leading to improved and nearly tight SQ lower bounds for natural concept classes. We describe our contributions in the following subsection.

1.2. Our Contributions

Contributions for Boolean-valued Concepts Our main general result for Boolean-valued concepts is the following:

Theorem 4 (Generic SQ Lower Bound, Boolean Case) *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Fix a function $f : \mathbb{R}^m \to \{\pm 1\}$. Let d be the smallest integer such that there exists a degree at most d polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[|p(\mathbf{x}) - f(\mathbf{x})|] < 2\epsilon$. Let $\mathcal{C}$ be a class of Boolean-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any $\mathbf{P} \in \mathbb{R}^{m\times n}$ such that $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Any SQ algorithm that agnostically learns $\mathcal{C}$ under $\mathcal{N}_n$ to error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

The L^1-polynomial regression algorithm and Theorem 4 characterize the complexity of agnostic learning under the Gaussian distribution – within the class of SQ algorithms – for a range of concept classes. If d is the (minimum) degree for which any function in $\mathcal{C}$ can be ϵ-approximated by a degree-d polynomial in L^1-norm, the complexity of agnostically learning $\mathcal{C}$ is, roughly, $n^{\Theta(d)}$.

Applications of Theorem 4. Note that the above result does not tell us what the optimal degree d is for any given concept class $\mathcal{C}$. Using analytic techniques, we establish explicit lower bounds on the L^1 polynomial approximation degree for three fundamental concept classes: Linear Threshold Functions (LTFs), Polynomial Threshold Functions, and Intersections of Halfspaces. As a corollary, we obtain explicit SQ lower bounds for these classes. Our applications are summarized in Table 1.

Concept Class	Lower Bound	Upper Bound
LTFs	$\Omega\left(1/\epsilon^2\right)$ (Ganzburg, 2002)	$O\left(1/\epsilon^2\right)$ (Ganzburg, 2002; Diakonikolas et al., 2010c)
Degree-k PTFs	$\Omega\left(k^2/\epsilon^2\right)$ (Thm 20)	$O\left(k^2/\epsilon^4\right)$ (Kane, 2010)
Intersections of k Halfspaces	$\tilde{\Omega}\left(\sqrt{\log k}/\epsilon\right)$ (Thm 23)	$O\left(\log k/\epsilon^4\right)$ (Klivans et al., 2008)

Table 1: Bounds on the degree d of ϵ-approximating polynomials in L^1-error under the Gaussian measure. For each concept class, we obtain an SQ lower bound of $n^{\Omega(d)}$.

For the class of LTFs, using a known degree lower bound for the sign function Ganzburg (2002), we immediately obtain an SQ lower bound of $n^{\Omega(1/\epsilon^2)}$. This bound is optimal (within polynomial factors), improving on the previous SQ lower bound of $n^{\Omega(1/\epsilon)}$ Goel et al. (2020b); Diakonikolas et al. (2020c). Our approach is simpler and more general compared to these prior works, immediately extending to other families. For the broader class of degree-k PTFs, we establish a degree lower bound of $\Omega(k^2/\epsilon^2)$ (Proposition 21), which yields an SQ lower bound of $n^{\Omega(k^2/\epsilon^2)}$ for the agnostic learning problem.

Our third explicit degree lower bound is for intersections of k halfspaces. For this concept class, we prove a degree lower bound of $d = \tilde{\Omega}(\sqrt{\log k}/\epsilon)$ which implies a corresponding SQ lower bound of $n^{\tilde{\Omega}(\sqrt{\log k}/\epsilon)}$. In the process, we establish a new structural result translating lower bounds on the

Gaussian Noise Sensitivity (GNS) of *any* Boolean function to the L^1-polynomial approximation degree of the same function.

Recall that the Gaussian Noise Sensitivity (GNS) of a function $f : \mathbb{R}^n \to \{\pm 1\}$ is defined as $\mathrm{GNS}_\rho(f) := \mathbf{Pr}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^\rho}[f(\mathbf{x}) \neq f(\mathbf{y})]$, where $\mathcal{N}_n^\rho$ is the distribution of a $(1-\rho)$-correlated Gaussian pair (i.e., $\mathbf{x}$ and $\mathbf{y}$ are standard Gaussians with correlation $(1-\rho)$). We show the following:

Theorem 5 (Structural Result) *Let $f : \mathbb{R}^n \to \{\pm 1\}$ and $p : \mathbb{R}^n \to \mathbb{R}$ be a degree at most d polynomial. Then, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x})-p(\mathbf{x})|] \geq \Omega(1/\log(d))\mathrm{GNS}_{(\log(d)/d)^2}(f)$. Furthermore, for any $\epsilon > 0$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - p(\mathbf{x})|] \geq \mathrm{GNS}_\epsilon(f)/4 - O(d\sqrt{\epsilon})$.*

Contributions for Real-valued Concepts For agnostically learning real-valued concepts, we provide two generic lower bound results, analogous to Theorem 4, for Correlational SQ (CSQ) algorithms and general SQ algorithms respectively. A conceptual message of our results is that L^2 regression is essentially optimal against CSQ algorithms, but not necessarily optimal against general SQ algorithms.

Recall that Correlational SQ (CSQ) algorithms are a subclass of SQ algorithms, where the algorithm is allowed to choose any bounded query function on the examples and obtain estimates of its correlation with the labels. (See Appendix A.1 for a detailed description.) This class of algorithms is fairly broad, capturing many learning algorithms used in practice (including gradient-descent). For CSQ algorithms, we prove.

Theorem 6 (Generic CSQ Lower Bound, Real-valued Case) *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Let $f : \mathbb{R}^m \to \mathbb{R}$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f^2(\mathbf{x})] = 1$ and d be the smallest integer such that there exists a degree at most d polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\|f-p\|_2 < \epsilon$. Let $\mathcal{C}$ be a class of real-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any matrix $\mathbf{P} \in \mathbb{R}^{m\times n}$ satisfying $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Then, any CSQ algorithm that agnostically learns $\mathcal{C}$ over $\mathcal{N}_n$ to L^2-error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

Our lower bound for the general SQ model is presented below. The difference between the two is that the latter uses the L^1-norm to measure the approximation of f by polynomials.

Theorem 7 (Generic SQ Lower Bound, Real-valued Case) *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Let $f : \mathbb{R}^m \to \mathbb{R}$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f^2(\mathbf{x})] = 1$ and d be the smallest integer such that there exists a degree at most d polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\|f-p\|_1 < \epsilon$. Let $\mathcal{C}$ be a class of real-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any matrix $\mathbf{P} \in \mathbb{R}^{m\times n}$ satisfying $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Then, any SQ algorithm that agnostically learns $\mathcal{C}$ over $\mathcal{N}_n$ to L^2-error $\mathrm{OPT}+\epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

Applications of Theorems 6 and 7. As in the Boolean-valued setting, obtaining explicit (C)SQ lower bounds for agnostically learning real-valued concepts requires analytic tools to establish lower bounds on the degree of polynomial approximations. In this paper, we give such lower bounds for two fundamental concept classes: ReLUs and sigmoids. Establishing degree lower bounds for other non-linear activations is left as a question for future work. Our degree lower bounds applications for both L^1 and L^2 polynomial approximations are summarized in Table 2. Combining these degree

	$p = 1$		$p = 2$	
Concept Class	Lower Bound	Upper Bound	Lower Bound	Upper Bound
ReLUs	$\Omega\left(1/\epsilon\right)$	$O\left(1/\epsilon\right)$	$\Omega\left(1/\epsilon^{4/3}\right)$	$O\left(1/\epsilon^{4/3}\right)$
Sigmoids	$\Omega(\log(1/\epsilon))$	$O(\log^2(1/\epsilon))$	$\Omega\left(\log^2(1/\epsilon)\right)$	$O\left(\log^2(1/\epsilon)\right)$

Table 2: Bounds on the degree d of ϵ-approximating polynomials in L^1 and L^2-error under the Gaussian measure. For each concept class, we obtain a CSQ (resp. SQ) lower bound of $n^{\Omega(d)}$, where d is the L^2 degree (resp. L^1 degree).

lower bounds Theorems 6 and 7 implies explicit SQ lower bounds for ReLUs and sigmoids. Concretely, for agnostically learning ReLUs, we establish a CSQ lower bound of $n^{\Omega(1/\epsilon^{4/3})}$ (matching the $n^{O(1/\epsilon^{4/3})}$ upper bound obtained via L^2-regression); and an SQ lower bound of $n^{\Omega(1/\epsilon)}$, improving on the previous best bound of $n^{\Omega((1/\epsilon)^{1/36})}$ Goel et al. (2020b); Diakonikolas et al. (2020c).

1.3. Overview of Techniques

SQ Lower Bounds for Boolean-valued Functions The starting point for our lower bounds is the work of Diakonikolas et al. (2017), which shows that if D is a univariate distribution whose low-degree moments match those of a standard Gaussian (and which satisfies some other mild niceness conditions), then it is SQ-hard to distinguish between a standard multivariate Gaussian and a distribution that is a copy of D in a random direction and a standard Gaussian in the orthogonal directions. (This is shown in Diakonikolas et al. (2017) for D a 1-dimensional distribution, but it is not hard to generalize to higher dimensional distributions.)

Note that the above setting is unsupervised. To go from distributions to functions, we will try to produce a Boolean function f of a few variables such that the distributions of X conditioned on $f(X) = 1$ and on $f(X) = -1$ match moments with a Gaussian. We generalize the techniques of Diakonikolas et al. (2020b) to show that such a function f embedded in a hidden *low-dimensional subspace* is SQ hard to distinguish from a random function. Our goal then is to find such a function f that is $(1/2 - \epsilon)$-close to a function in our family. Given this construction, learning the function to error $\mathrm{OPT} + \epsilon/2$ requires being able to distinguish f from a random function.

The aforementioned approach was recently used by Diakonikolas et al. (2020c). However, while that work constructs the function f somewhat directly, here we take a more general approach. In more detail, it is not hard to phrase the conditions that (1) f is bounded in $[-1, 1]$, (2) it matches moments with low-degree polynomials, and (3) is not too far from the function we are trying to learn, as an infinite-dimensional linear program (LP). We can then non-constructively attempt to find the optimal value of such an LP by duality. We note that "LP duality" in this setting is non-trivial – we require some (basic) functional analysis tools to show that duality applies for the LPs we are considering on function spaces. Given this, we find that the dual program is equivalent to finding a low-degree polynomial that approximates the function we are trying to learn in L^1-norm. The degree of such a polynomial conveniently matches the parameter that determines the runtime of the L^1- polynomial regression algorithm. We can thus show that, for reasonable function families, the L^1-regression algorithm is in fact optimal, among SQ algorithms, up to polynomial factors.

The above characterization allows us to determine the complexity of agnostically learning LTFs, by leverage tight degree lower bounds for the sign function. For the cases of degree-k PTFs and intersections of k halfspaces, we do not know what the correct answer is, but we are able to prove non-trivial, and qualitatively close to optimal, lower bounds.

We note that the L^1 approximation theory for these functions is more challenging than the L^2 approximation theory (which is entirely determined by the Fourier decay). To that end, we develop new techniques relating L^1 approximability to the Gaussian Noise sensitivity (Theorem 5), which allows us to prove the first non-trivial lower bounds. The proof of Theorem 5 works via a symmetrization technique. In particular, let $\theta = \arccos(1-\epsilon)$ and let X and Y be standard Gaussians. Let $F_{X,Y}(\phi) := f(\sin(\phi)X + \cos(\phi)Y)$. Then we can write $\mathrm{GNS}_\epsilon(f) = \mathbf{Pr}[F_{X,Y}(\phi) \neq F_{X,Y}(\phi+\theta)]$. On the other hand, $\|f-p\|_1 = \mathbf{E}[|F_{X,Y}(\phi) - p(\sin(\phi)X + \cos(\phi)Y)|]$. Thus, it suffices to show that if F is *any* Boolean function on the circle that the L^1 approximation error of F by low degree polynomials can be bounded below by $\mathbf{Pr}[F(\phi) \neq F(\phi+\theta)]$. To show this, we use basic Fourier analysis to show that any low-degree polynomial with small L^1 norm cannot have any large higher derivatives. This implies that if F transitions from being 0 to being 1 over some small interval, that any low-degree polynomial will not be able to match it very well in this interval.

(C)SQ Lower Bounds for Real-valued Functions We now move to real-valued functions and sketch our CSQ and SQ lower bounds. For CSQ lower bounds, we obtain a similar characterization. The difference is that, in the real-valued setting, we need to find a real-valued function f whose low-degree moments vanish, and which is close to the function we are trying to learn *in L^2 norm*. This can be phrased as a similar LP and, applying duality, we find that the complexity is determined by the degree needed to approximate the function we are trying to learn in L^2 norm. For this particular setting, the LP can actually be solved explicitly and the best possible approximation function is obtained by taking the high-degree Hermite component of f. This lower bound matches (up to polynomial factors in the final error) the upper bound coming from the L^2 polynomial regression algorithm. This means that we can qualitatively characterize the complexity of agnostic learning using CSQ algorithms. In particular, we use this characterization to obtain new CSQ lower bounds on agnostically learning ReLUs and sigmoids.

Our SQ lower bounds against learning real-valued functions are somewhat more challenging, since the approximating function f must have more than just vanishing moments. It must have all its level-sets match low-degree moments with a standard Gaussian (which is equivalent only for Boolean-valued functions). Because of this additional requirement, we restrict our "imitating functions" to Boolean-valued functions. We can still find an LP defining f, however the dual gives us the relevant parameter of the degree needed to approximate the function we are trying to learn in L^1-norm (rather than L^2-norm) for which a matching upper bound is not known. So, in this case, while we can still obtain significantly improved SQ lower bounds for agnostically learning a number of concept classes, we do not obtain optimal results.

1.4. Comparison to Prior Work

At the level of results, the most relevant prior works are the two independent works Diakonikolas et al. (2020c); Goel et al. (2020b), which established the previously best SQ lower bounds for LTFs, ReLUs, and sigmoids under the Gaussian distribution. We have already provided a technical comparison to Diakonikolas et al. (2020c) in the previous subsection. The work Goel et al. (2020b) relies on a boosting procedure that translates recent SQ lower bounds for (non-agnostic) learning one-hidden-layer neural networks Diakonikolas et al. (2020b) to agnostically learning simple concept classes.

A useful point of technical comparison is the work Dachman-Soled et al. (2015), which gave an analogue of our results on agnostically learning Boolean functions on the Boolean hypercube. The basic statement is the same — that the complexity of agnostic learning Boolean functions under a discrete product distribution is characterized by the L^1-approximation degree — and the duality-based proof techniques are similar. In particular, Dachman-Soled et al. (2015) sets up a *finite* LP to find a function f that has vanishing Fourier coefficients but is close in L^1-norm to the target function. Due to the discrete nature of the setting they consider, Dachman-Soled et al. (2015) avoids the functional analysis based arguments required to establish duality in our setting.

A more significant difference with our framework is that the hard family of Dachman-Soled et al. (2015) embeds a copy of f as a junta on a random subset of coordinates, while ours embeds it in a random low-dimensional subspace. This is a critical distinction and is necessary in the Gaussian setting to obtain our tight characterization and the associated applications to LTFs/PTFs and intersections of halfspaces. Finally, we remark that the appendix of Dachman-Soled et al. (2015) sketches a generalization of their results to arbitrary product distributions (including the Gaussian distribution). We emphasize, however, that the lower bound obtained from their construction does *not* match the guarantee of the L^1-regression algorithm Kalai et al. (2008) for the following reason: The exponent for their lower bounds for the continuous setting have to do with the degree necessary to ϵ-approximate the hard function as *a linear combination of d-juntas*. On the other hand, the upper bound of Kalai et al. (2008) is related to the approximation by degree-d polynomials. Note that degree-d polynomials are always linear combinations of d-juntas, and thus the approximation degree by linear combinations of juntas is lower than the approximation degree by polynomials. In summary, while the lower bound of Dachman-Soled et al. (2015) is tight for discrete product distributions, this is not true in general.

1.5. Preliminaries

Notation For $n \in \mathbb{Z}_+$, we denote $[n] := \{1, \ldots, n\}$. 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 $\mathcal{E}$. We will use $\mathcal{U}(S)$ for the uniform distribution on the set S. Let $\mathcal{N}$ denote the standard univariate Gaussian distribution and $\mathcal{N}_n$ denote the standard n-dimensional Gaussian distribution. We use ϕ_n to denote the pdf of $\mathcal{N}_n$. Sometimes we may use the same symbol for a distribution and its pdf, i.e., denote by $D(\mathbf{x})$ the density that the distribution D gives to the point $\mathbf{x}$.

Small boldface letters are used for vectors and capital boldface letters are used for matrices. Let $\|\mathbf{x}\|_2$ denote the L^2-norm of the vector $\mathbf{x} \in \mathbb{R}^n$. We use $\langle \mathbf{u}, \mathbf{v} \rangle$ for the inner product of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. For a matrix $\mathbf{P} \in \mathbb{R}^{m \times n}$, let $\|\mathbf{P}\|_2$ denote its spectral norm and $\|\mathbf{P}\|_F$ denote its Frobenius norm. We use $\mathbf{I}_n$ to denote the $n \times n$ identity matrix. We denote by $\mathcal{P}_d^n$ the class of all polynomials from $\mathbb{R}^n$ to $\mathbb{R}$ with degree at most d. We sometimes use the notation $\tilde{O}(\cdot)$ (resp. $\tilde{\Omega}(\cdot)$), this is the same with $O(\cdot)$ (resp. $\Omega(\cdot)$), ignoring logarithmic factors, i.e., $O(d \log^k d) = \tilde{O}(d)$.

Statistical Query Dimension To bound the complexity of SQ learning a concept class $\mathcal{C}$, we use the SQ framework for problems over distributions Feldman et al. (2013).

Definition 8 (Decision Problem over Distributions) *Let D be a fixed distribution and $\mathcal{D}$ be a distribution family. We denote by $\mathcal{B}(\mathcal{D}, D)$ the decision (or hypothesis testing) problem in which the input distribution D' is promised to satisfy either (a) $D' = D$ or (b) $D' \in \mathcal{D}$, and the goal is to distinguish between the two cases.*

Definition 9 (Pairwise Correlation) *The pairwise correlation of two distributions with probability density functions $D_1, D_2 : \mathbb{R}^n \to \mathbb{R}_+$ with respect to a distribution with density $D : \mathbb{R}^n \to \mathbb{R}_+$, where the support of D contains the supports of D_1 and D_2, is defined as $\chi_D(D_1, D_2) := \int_{\mathbb{R}^n} D_1(\mathbf{x}) D_2(\mathbf{x})/D(\mathbf{x})\,\mathrm{d}\mathbf{x} - 1$.*

Definition 10 *We say that a set of s distributions $\mathcal{D} = \{D_1, \ldots, D_s\}$ over $\mathbb{R}^n$ is (γ, β)-correlated relative to a distribution D if $|\chi_D(D_i, D_j)| \le \gamma$ for all $i \ne j$, and $|\chi_D(D_i, D_j)| \le \beta$ for $i = j$.*

Definition 11 (Statistical Query Dimension) *For $\beta, \gamma > 0$, a decision problem $\mathcal{B}(\mathcal{D}, D)$, where D is a fixed distribution and $\mathcal{D}$ is a family of distributions, let s be the maximum integer such that there exists a finite set of distributions $\mathcal{D}_D \subseteq \mathcal{D}$ such that $\mathcal{D}_D$ is (γ, β)-correlated relative to D and $|\mathcal{D}_D| \ge s$. The* Statistical Query dimension *with pairwise correlations (γ, β) of $\mathcal{B}$ is defined to be s, and denoted by $\mathrm{SD}(\mathcal{B}, \gamma, \beta)$.*

Lemma 12 *Let $\mathcal{B}(\mathcal{D}, D)$ be a decision problem, where D is the reference distribution and $\mathcal{D}$ is a class of distributions. For $\gamma, \beta > 0$, let $s = \mathrm{SD}(\mathcal{B}, \gamma, \beta)$. For any $\gamma' > 0$, any SQ algorithm for $\mathcal{B}$ requires queries of tolerance at most $\sqrt{\gamma + \gamma'}$ or makes at least $s\gamma'/(\beta - \gamma)$ queries.*

2. SQ Lower Bound for Boolean-Valued Concepts: Proof of Theorem 4

The idea of our construction is to find a function $g : \mathbb{R}^m \to [-1, 1]$ whose low-degree moments vanish and is non-trivially close to f. Our hard distribution will then embed g in a random m-dimensional subspace. Given this construction, we can apply Lemma 12 to prove Theorem 4. The following result establishes the existence of such a function g.

Proposition 13 *Let $f : \mathbb{R}^m \to \{\pm 1\}$ be such that for any polynomial $p : \mathbb{R}^m \to \mathbb{R}$ of degree at most $d{-}1$, it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[|p(\mathbf{x}) - f(\mathbf{x})|] \ge 2\epsilon$. There exists a function g: $\mathbb{R}^m \to [-1, 1]$ such that:*

1. *For any degree at most $d-1$ polynomial $P : \mathbb{R}^m \to \mathbb{R}$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0$, i.e., g has zero low-degree moments, and,*
2. *$\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[|g(\mathbf{x}) - f(\mathbf{x})|] \le 1 - 2\epsilon$, i.e., g is non-trivially close to f.*

Proof Such a function g would be a solution to the infinite linear program $(*)$ below, which we claim that is equivalent to the linear program $(**)$:

$$(*)\begin{cases} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[|g(\mathbf{x}) - f(\mathbf{x})|] \le 1 - 2\epsilon \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 \;\; \forall P \in \mathcal{P}^m_{d-1} \\ |g(\mathbf{x})| \le 1 \;\; \forall \mathbf{x} \in \mathbb{R}^m \end{cases} \qquad (**)\begin{cases} -\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})f(\mathbf{x})] + 2\epsilon \le 0 \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 \;\; \forall P \in \mathcal{P}^m_{d-1} \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})h(\mathbf{x})] - \|h\|_1 \le 0 \;\; \forall h \in L^1(\mathbb{R}^m) \end{cases}$$

We now show the equivalence between the two formulations. We claim that the third constraint of $(*)$ is equivalent with the third constraint of $(**)$. This follows by introducing the "dual variable" $h : \mathbb{R}^m \to \mathbb{R}$. The forward direction follows from Hölder's inequality and the inverse follows from the definition of dual norms as suprema. Finally, for the first constraints, note that since f is Boolean-valued and $\|g\|_\infty \le 1$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[|g(\mathbf{x}) - f(\mathbf{x})|] = 1 - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})f(\mathbf{x})]$.

At this point, we would like to use "LP duality" to argue that $(**)$ is feasible if and only if its "dual LP" is infeasible. While such a statement turns out to be true, it requires some care to prove since we are dealing with infinite LPs (both in number of variables and constraints). The proof requires a version of the geometric Hahn-Banach theorem from functional analysis.

Lemma 14 (Informal) *The LP defined by $(**)$ is feasible if only if there is no conical combination of the inequalities of $(**)$ that yields the contradictory inequality $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})\cdot 0]+1\leq 0$.*

A proof of this lemma can be found on Appendix F. Using Lemma 14, the LP defined by $(**)$ is feasible if and only if the following "dual" LP is infeasible:

$$(**') \begin{cases} \|h\|_1 - 2\lambda\,\epsilon < 0 \\ h(\mathbf{x}) + P(\mathbf{x}) - \lambda\, f(\mathbf{x}) = 0 & \qquad\qquad \forall \mathbf{x}\in\mathbb{R}^m \\ \lambda \geq 0, h\in L^1(\mathbb{R}^m), P\in\mathcal{P}^m_{d-1} \end{cases}$$

Suppose that such a solution (λ, h, P) exists. We can assume that $\lambda > 0$, since otherwise the first inequality is violated. Moreover, by scaling the solution, we can further assume $\lambda = 1$. Then, the second constraint becomes $h = f - P$ and the first becomes $\|f-P\|_1 < 2\epsilon$. However, this cannot happen by the definition of the degree d (since, by assumption, there is no polynomial of degree less than d such that $\|f-P\|_1 < 2\epsilon$). Therefore, the LP $(**)$ is feasible, which completes our proof. ■

Our construction will use rotated versions of the function g from Proposition 13 to create a family of distributions that is hard to distinguish from a fixed reference distribution. To bound the SQ dimension of this hypothesis testing problem, we will need a generalization of Lemma 16 in Diakonikolas et al. (2020b), which bounds the correlation of two rotated versions of g. To formally state our lemma, we will need one additional piece of terminology. If $g(\mathbf{x}) = \sum_{J\in\mathbb{N}^m}\hat{g}(J)H_J(\mathbf{x})$ is the Hermite expansion of g, the degree-t Hermite part of g is the sum of the terms corresponding to the Hermite polynomials of degree exactly t. (For background in multilinear algebra and Hermite analysis, see Appendices A.2 and A.3.) Our main correlation lemma is the following.

Lemma 15 (Correlation Lemma) *Let $g:\mathbb{R}^m\to\mathbb{R}$ and $\mathbf{U},\mathbf{V}\in\mathbb{R}^{m\times n}$ be linear maps such that $\mathbf{U}\mathbf{U}^\intercal = \mathbf{V}\mathbf{V}^\intercal = \mathbf{I}_m$. Then, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[g(\mathbf{U}\mathbf{x})g(\mathbf{V}\mathbf{x})] \leq \sum_{t=0}^{\infty}\|\mathbf{U}\mathbf{V}^\intercal\|_2^t\,\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[(g^{[t]}(\mathbf{x}))^2]$, where $g^{[t]}$ denotes the degree-t Hermite part of g.*

We consider high-dimensional distributions that encode a function in a subspace and are Gaussian in the orthogonal complement. Using Lemma 15, we can bound their pairwise correlations.

Corollary 16 *Let $d\geq 2$ and D be a distribution over $\mathbb{R}^m$ such that the first $(d-1)$ moments of D match the corresponding moments of $\mathcal{N}_m$. Let $G(\mathbf{x}) = D(\mathbf{x})/\phi_m(\mathbf{x})$ be the ratio of the corresponding probability density functions. For matrices $\mathbf{U},\mathbf{V}\in\mathbb{R}^{m\times n}$ such that $\mathbf{U}\mathbf{U}^\intercal = \mathbf{V}\mathbf{V}^\intercal = \mathbf{I}_m$, define $D_{\mathbf{U}}$ and $D_{\mathbf{V}}$ to have probability density functions $G(\mathbf{U}\mathbf{x})\phi_n(\mathbf{x})$ and $G(\mathbf{V}\mathbf{x})\phi_n(\mathbf{x})$, respectively. Then, we have that $|\chi_{\mathcal{N}_n}(D_{\mathbf{U}}, D_{\mathbf{V}})| \leq \|\mathbf{U}\mathbf{V}^\intercal\|_2^d\chi^2(D,\mathcal{N}_m)$.*

See Appendix B.2 for the proof. Note that $D_{\mathbf{U}}$ and $D_{\mathbf{V}}$ are copies of D in the subspaces defined by $\mathbf{U}$ and $\mathbf{V}$ respectively, and independent Gaussians in the orthogonal component. In order to create our hard family of distributions, we will need the following lemma which states that there exist exponentially many linear operators from $\mathbb{R}^n$ to $\mathbb{R}^m$ that are nearly orthogonal. Its proof is deferred to Appendix B.3.

Lemma 17 *Let $0 < a, c < 1/2$ and $m, n \in \mathbb{Z}_+$ such that $m \leq n^a$. There exists a set S of $2^{\Omega(n^c)}$ matrices in $\mathbb{R}^{m \times n}$ such that every $\mathbf{U} \in S$ satisfies $\mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_m$ and every pair $\mathbf{U}, \mathbf{V} \in S$ with $\mathbf{U} \neq \mathbf{V}$ satisfies $\|\mathbf{U}\mathbf{V}^\intercal\|_F \leq O(n^{2c-1+2a})$.*

We now formally define the family of distributions that we use to prove our hardness result.

Definition 18 *Given a function $g : \mathbb{R}^m \to [-1, 1]$, we define $\mathcal{D}_g$ to be the class of distributions over $\mathbb{R}^n \times \{\pm 1\}$ of the form $(\mathbf{x}, y)$ such that $\mathbf{x} \sim \mathcal{N}_n$ and $\mathbf{E}[y|\mathbf{x} = \mathbf{z}] = g(\mathbf{U}\mathbf{z})$, where $\mathbf{U} \in \mathbb{R}^{m \times n}$ with $\mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_m$.*

In the following, we show that if g has low-degree moments equal to zero, then distinguishing $\mathcal{D}_g$ from the distribution $(\mathbf{x}, y)$ with $\mathbf{x} \sim \mathcal{N}_n$, $y \sim \mathcal{U}(\{\pm 1\})$ is hard in the SQ model.

Proposition 19 *Let $g : \mathbb{R}^m \to [-1, 1]$ be such that $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}_m}[g(\mathbf{x})p(\mathbf{x})] = 0$, for every polynomial $p : \mathbb{R}^m \to \mathbb{R}$ of degree less than d, and $\mathcal{D}_g$ be the class of distributions from Definition 18. Then, if $m \leq n^a$, for some constant $a < 1/2$, any SQ algorithm that solves the decision problem $\mathcal{B}(\mathcal{D}_g, \mathcal{N}_n \times \mathcal{U}(\{\pm 1\}))$ must either use queries of tolerance $n^{-\Omega(d)}$ or make at least $2^{n^{\Omega(1)}}$ queries.*

Proof Consider the set of matrices S of Lemma 17, for an appropriately small value of $c > 0$. Each matrix $\mathbf{U} \in S$ is associated with a unique element of $\mathcal{D}_g$. For every pair of distinct $\mathbf{U}, \mathbf{V} \in S$, we have that $\|\mathbf{U}\mathbf{V}^\intercal\|_2 \leq \|\mathbf{U}\mathbf{V}^\intercal\|_F \leq O(n^{2c-1+2a}) \leq n^{-\Omega(1)}$, where for the last inequality we chose c to be a sufficiently small constant, e.g., $c = (1 - 2a)/4$.

Note that the distribution in $\mathcal{D}_g$ associated to a matrix $\mathbf{U}$ has probability density $(1+g(\mathbf{U}\mathbf{x}))\phi_n(\mathbf{x})$ when conditioned on $y = 1$, and density $(1 - g(\mathbf{U}\mathbf{x}))\phi_n(\mathbf{x})$ when conditioned on $y = -1$. Let $D_\mathbf{U}$ be the distribution associated to $\mathbf{U}$ and $D_\mathbf{V}$ the distribution associated to $\mathbf{V}$. Denote by $A_\mathbf{U}$ the distribution $D_\mathbf{U}$ conditioned on the event $y = 1$ and $B_\mathbf{U}$ the same distribution conditioned on $y = -1$. Similarly, let $A_\mathbf{V}$ and $B_\mathbf{V}$ denote the conditional distributions associated with $\mathbf{V}$. Using the definition of pairwise correlation and the fact that y gets each label with equal probability, it follows directly that $\chi_{\mathcal{N}_n \times \mathcal{U}(\{\pm 1\})}(D_\mathbf{U}, D_\mathbf{V}) = \frac{1}{2}\left(\chi_{\mathcal{N}_n}(A_\mathbf{U}, A_\mathbf{V}) + \chi_{\mathcal{N}_n}(B_\mathbf{U}, B_\mathbf{V})\right)$. By Corollary 16 applied to $A_\mathbf{U}, A_\mathbf{V}$ and $B_\mathbf{U}, B_\mathbf{V}$, we obtain $\chi_{\mathcal{N}_n}(A_\mathbf{U}, A_\mathbf{V}) + \chi_{\mathcal{N}_n}(B_\mathbf{U}, B_\mathbf{V}) \leq \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \left(\chi^2(A, \mathcal{N}_m) + \chi^2(B, \mathcal{N}_m)\right)$, where A is the distribution of the random variable $\mathbf{U}\mathbf{x}$ for $\mathbf{x} \sim A_\mathbf{U}$ (and similarly for B). For the χ^2-divergence terms, we have that $\chi^2(A, \mathcal{N}_m) = 1$ (see Appendix B.4). Combining the above, we get that $|\chi_{\mathcal{N}_n \times \mathcal{U}(\{\pm 1\})}(D_\mathbf{U}, D_\mathbf{V})| \leq n^{-\Omega(d)}$. This inequality implies that $\mathrm{SD}(\mathcal{B}, \gamma, \beta) = 2^{\Omega(n^c)}$, for $\gamma = n^{-\Omega(d)}$ and $\beta = O(1)$. Using Lemma 12, with $\gamma' = \gamma$, completes the proof. ■

Proof [Proof of Theorem 4] Let $\mathcal{A}$ be an agnostic learner for $\mathcal{C}$. We use $\mathcal{A}$ to solve the decision problem $\mathcal{B}(\mathcal{D}_g, \mathcal{N}_n \times \mathcal{U}(\{\pm 1\}))$, where $g : \mathbb{R}^m \to [-1, 1]$ is the function from Proposition 13 and $\mathcal{D}_g$ the family of Definition 18. Let D' be the target distribution, i.e., $D' = \mathcal{N}_n \times \mathcal{U}(\{\pm 1\})$ if the null hypothesis is true or $D' \in \mathcal{D}_g$ otherwise. We feed $\mathcal{A}$ examples drawn from D' and it outputs a hypothesis $h : \mathbb{R}^n \to \{\pm 1\}$ such that $\mathbf{Pr}_{(\mathbf{x},y) \sim D'}[h(\mathbf{x}) \neq y] \leq \mathrm{OPT} + \frac{\epsilon}{2}$. If $D' \in \mathcal{D}_g$, then for a matrix $\mathbf{U} \in \mathbb{R}^{m \times n}$ with $\mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_m$, we have that $\mathrm{OPT} \leq \mathbf{Pr}_{(\mathbf{x},y) \sim D'}[f(\mathbf{U}\mathbf{x}) \neq y] = \frac{1}{2}\|f - g\|_1 \leq \frac{1}{2}(1 - 2\epsilon)$, where in the equality we used the fact that the expectation of y conditioned on $\mathbf{x}$ is $g(\mathbf{x})$ and the last inequality is due to Proposition 13. Combining the above, we get that $\mathbf{Pr}_{(\mathbf{x},y) \sim D'}[h(\mathbf{x}) \neq y] \leq (1 - \epsilon)/2$, or equivalently that $\mathbf{E}_{(\mathbf{x},y) \sim D'}[h(\mathbf{x})y] \geq \epsilon$. On the other hand, if the labels were drawn uniformly at random, this correlation would be exactly 0. Therefore, we can distinguish between the two cases by performing a final query of tolerance $\epsilon/2$ for the correlation of h with y. ■

3. Explicit SQ Lower Bounds for Boolean Concept Classes

3.1. LTFs and Degree-k PTFs

Linear threshold functions (LTFs) are Boolean functions of the form $F(\mathbf{x}) = \mathrm{sign}(\langle \mathbf{w}, \mathbf{x}\rangle + \theta)$, where $\mathbf{w} \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$. A degree-k PTF is any Boolean function of the form $F(\mathbf{x}) = \mathrm{sign}(q(\mathbf{x}))$, where $q : \mathbb{R}^n \to \mathbb{R}$ is a real degree-k polynomial. In this section, we show:

Theorem 20 (Degree Lower Bound for PTFs) *There exists a degree-k PTF $f : \mathbb{R} \to \{\pm 1\}$ such that any degree-d polynomial $p : \mathbb{R} \to \mathbb{R}$ with $\|f - p\|_1 < \epsilon$ must have $d = \Omega(k^2/\epsilon^2)$.*

Theorems 4 and 20 imply that any SQ algorithm that agnostically learns the class of degree-k PTFs on $\mathbb{R}^n$ under the Gaussian distribution must have complexity at least $n^{\Omega(k^2/\epsilon^2)}$.

Lower Bound for LTFs The L^1-regression algorithm Kalai et al. (2008) is known to be an agnostic learner for LTFs under Gaussian marginals with complexity $n^{O(1/\epsilon^2)}$. This upper bound uses the known fact that the L^1 polynomial ϵ-approximate degree of LTFs under the Gaussian distribution is $d = O(1/\epsilon^2)$ (see, e.g., Diakonikolas et al. (2010c)). This upper bound is tight. Specifically, known results in approximation theory (see Appendix C.1) imply that, any polynomial that ϵ-approximates the function $\mathrm{sign}(t)$ in L^1-norm, under the standard Gaussian distribution, requires degree $\Omega(1/\epsilon^2)$. Given this structural result, an application of Theorem 4, for $m = 1$ and $f(t) = \mathrm{sign}(t)$ gives the tight SQ lower bound of $n^{\Omega(1/\epsilon^2)}$. This bound improves on the best previous bound of $n^{\Omega(1/\epsilon)}$ Goel et al. (2020b); Diakonikolas et al. (2020c). Importantly, our approach is much simpler and generalizes to any concept class satisfying the mild assumptions of Theorem 4.

Lower Bound for Degree-k PTFs The L^1-regression algorithm is known to be an agnostic learner for degree-k PTFs under Gaussian marginals with complexity $n^{O(k^2/\epsilon^4)}$. This bound uses the known upper bound of $O(k\sqrt{\epsilon})$ on the Gaussian noise sensitivity of degree-k PTFs Kane (2010), which implies an upper bound of $O(k^2/\epsilon^2)$ on the L^2 polynomial ϵ-approximate degree, and therefore an upper bound of $O(k^2/\epsilon^4)$ on the L^1 polynomial ϵ-approximate degree. This upper bound is not known to be optimal (it is sub-optimal for k=1) and it is a plausible conjecture that the right answer is $\Theta(k^2/\epsilon^2)$. We prove a lower bound of $\Omega(k^2/\epsilon^2)$, which applies even for the univariate case.

Proposition 21 *There exists a $(k + 1)$-piecewise-constant function $f : \mathbb{R} \to \{0, 1\}$ such that any degree-d polynomial $p : \mathbb{R} \to \mathbb{R}$ that satisfies $\|f - p\|_1 < \epsilon$ must have $d = \Omega(k^2/\epsilon^2)$.*

An application of Theorem 4, for $m = 1$ and $f(t)$ being the piecewise constant function of Proposition 21, implies an SQ lower bound of $n^{\Omega(k^2/\epsilon^2)}$. Before we provide the formal proof, we sketch the proof of Proposition 21. The hard function f consists of $k/2$ intervals with the same carefully chosen length; we split each interval in half and we let $f = 0$ in the first half, and $f = 1$ in the second half. We construct a distribution D that puts almost all of its mass in the first half of each interval, matches the first d moments with the standard Gaussian, and $D(x) \leq 2\phi(x)$ for all $\mathbf{x} \in \mathbb{R}$. Then, by construction $\mathbf{E}_{x\sim\mathcal{N}}[f(x)]$ is much larger than the same expectation under D. We show that, in fact, this difference bounds from below the error of any degree-d polynomial approximation to the function f. The main technical lemma we establish in this context is given below. The proofs of Proposition 21 and Lemma 22 can be found in Appendices C.2 and C.3.

Lemma 22 *There exists a distribution D that (i) matches its first d moments with $\mathcal{N}$, (ii) the pdf of D is at most 2 times the pdf of $\mathcal{N}$ pointwise in $\mathbb{R}$, and (iii) for some $\alpha = \Theta(1/\sqrt{d})$ it holds that* $\mathbf{Pr}[(X \mod a) \in (a/2, a)] = 2^{-\Omega(d)}$.

3.2. Intersections of Halfspaces: Degree Lower Bound via Gaussian Noise Sensitivity

An intersection of k halfspaces on $\mathbb{R}^n$ is any function $f : \mathbb{R}^n \to \{\pm 1\}$ such that there exist k LTFs $h_i : \mathbb{R}^n \to \{\pm 1\}$, $i \in [k]$, such that $f(\mathbf{x}) = 1$ if and only if $h_i(\mathbf{x}) = 1$ for all $i \in [k]$.

The L^1-regression algorithm Kalai et al. (2008) is known to be an agnostic learner for intersection of k halfspaces on $\mathbb{R}^n$ under Gaussian marginals with complexity $n^{O((\log k)/\epsilon^4)}$. This upper bound uses the known tight upper bound of $O(\sqrt{\epsilon \log k})$ on the Gaussian noise sensitivity of this concept class Klivans et al. (2008), which implies an upper bound of $O(\log k/\epsilon^4)$ on the L^1 polynomial ϵ-degree. This degree upper bound is not known to be optimal (in fact, it is provably suboptimal for $k = 1$) and it is a plausible conjecture that the right answer is $\Theta(\sqrt{\log k}/\epsilon^2)$. Here we prove a lower bound of $\tilde{\Omega}(\sqrt{\log k}/\epsilon)$, which applies even for k-dimensional functions.

Theorem 23 (Degree Lower Bound for Intersections of Halfspaces) *There exists an intersection of k halfspaces f on $\mathbb{R}^k$ such that the following holds: Any degree-d polynomial $p : \mathbb{R}^k \to \mathbb{R}$ that satisfies $\|f - p\|_1 < \epsilon$ must have $d = \tilde{\Omega}(\sqrt{\log k}/\epsilon)$.*

Theorem 23 combined with Theorem 4, applied for $m = k$ and f being the function from Theorem 23, implies that any SQ algorithm that agnostically learns intersections of k halfspaces on $\mathbb{R}^n$ under the Gaussian distribution must have complexity at least $n^{\tilde{\Omega}(\sqrt{\log k}/\epsilon)}$. To prove Theorem 23, we make essential use of our structural result, Theorem 5, combined with the following tight lower bound on the Gaussian noise sensitivity of a well-chosen family of intersection of halfspaces (see Appendix C.4 for the proof).

Lemma 24 *There exists an intersection of k halfspaces on $\mathbb{R}^k$, $f : \mathbb{R}^k \to \{\pm 1\}$, such that $\mathrm{GNS}_\epsilon(f) = \Omega(\sqrt{\epsilon \log k})$.*

3.3. Proof of Theorem 5

Proposition 25 *Let $p(\theta)$ be a degree-d polynomial on the circle, i.e., a degree at most d polynomial in $\sin\theta$ and $\cos\theta$, and let $B(\theta)$ be a Boolean-valued function that is periodic modulo 2π. Then, for t being a sufficiently small multiple of $\log d/d$, it holds $\frac{1}{2\pi}\int_0^{2\pi} |p(\theta) - B(\theta)| \mathrm{d}\theta = \tilde{\Omega}(1/\log d)\, \mathbf{Pr}_{\phi\sim\mathcal{U}([0,2\pi])}[B(\phi - t) \neq B(\phi + t)]$.*

Proof We can assume that $\frac{1}{2\pi}\int_0^{2\pi} |p(\theta)| \mathrm{d}\theta$ is at most 2, since otherwise the $\frac{1}{2\pi}\int_0^{2\pi} |p(\theta) - B(\theta)| \mathrm{d}\theta$ is at least 1. Let k be an odd integer proportional to $\log d$. We start with the following technical claim (see Appendix C.5 for the proof).

Claim 26 *For any $\theta \in [0, 2\pi]$, it holds $|p^{(k)}(\theta)| = O(d)^k$.*

We next pick t a small multiple of $\log d/d$ and $\phi \in [0, 2\pi]$. Let $z_m = t\cos(\pi m/k) + \phi$, for $m = 0, 1, \ldots, k$, and let $q(z) = \sum_{j=0}^k c_j z^j$ be the unique degree-k polynomial such that $q(z_m) = p(z_m)$, for $m = 0, 1, \ldots, k$. Observe that $q - p$ has $k + 1$ zeroes. Therefore, iterating Rolle's theorem we obtain that there is a point $\phi - t \le z \le \phi + t$ such that $p^{(k)}(z) = q^{(k)}(z)$, and thus $|q^{(k)}(z)| = O(d)^k$, or equivalently $c_k = 2^k O(d/k)^k$.

Let $R(\theta) = q(t\cos\theta + \phi)$. For some constants b_n (which depend on t and ϕ), we have that $R(\theta) = \sum_{n=-k}^k b_n e^{ni\theta}$. Since $R(\theta)$ is an even function, its Fourier coefficients are real numbers. The following claim provides an upper bound on the coefficient b_k (proof in Appendix C.5).

Claim 27 *It holds that* $|b_k| \leq 1/(4k)$.

On the other hand, by doing a filtering using the $(2k)$-th roots of unity, we get that $\sum_{m=0}^{2k-1} R(2\pi m/(2k)) = 2kb_k$, and this is equivalent to $\sum_{m=-k+1}^{k} q(t\cos(\pi m/k) + \phi)(-1)^m = 2kb_k$. Therefore,

$$\begin{aligned} b_k &= \frac{1}{2k} \sum_{m=-k+1}^{k} q(t\cos(\pi m/k) + \phi)(-1)^m = \frac{1}{2k} \sum_{m=-k+1}^{k} p(z_{|m|})(-1)^m \\ &= \frac{1}{2k}\Big(\sum_{m=-k+1}^{k} (p(z_{|m|}) - B(z_{|m|}))(-1)^m + \sum_{m=-k+1}^{k} B(z_{|m|})(-1)^m + (B(\phi+t) - B(\phi - t))\Big)\,. \end{aligned}$$

Since $k-1$ is even and B is Boolean, $2\sum_{m=1}^{k-1} B(z_m)(-1)^m$ is a multiple of 4. If $B(\phi+t) \neq B(\phi-t)$, the reverse triangle inequality gives $\left|B(\phi+t) - B(\phi-t) + 2\sum_{m=1}^{k-1} B(z_m)(-1)^m\right| \geq 2$. Therefore, in this case, we have that $\frac{1}{4k} > |b_k| \geq \frac{1}{2k}\left(2 - \sum_{m=-k+1}^{k} |p(z_{|m|}) - B(z_{|m|})|\right)$, or in other words, $\sum_{m=-k+1}^{k} |p(z_{|m|}) - B(z_{|m|})| \geq \mathbb{1}\{B(\phi+t) \neq B(\phi-t)\}$. Integrating this over ϕ from 0 to 2π gives $\int_0^{2\pi} |p(\theta) - B(\theta)| d\theta \geq \frac{\pi}{k} \mathbf{Pr}_{\phi\sim\mathcal{U}([0,2\pi])}[B(\phi - t) \neq B(\phi+t)]$. The result follows from our assumption that k is proportional to $\log d$. ■

Proof [Proof of Theorem 5] The latter bound follows from the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - p(\mathbf{x})|] \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - \mathrm{sign}(p(\mathbf{x}))|/2]$. On the other hand, we can write $\mathrm{GNS}_\epsilon(f) - \mathrm{GNS}_\epsilon(\mathrm{sign}(p)) = \mathbf{Pr}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_d^\epsilon}[f(\mathbf{x}) \neq f(\mathbf{y})] - \mathbf{Pr}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_d^\epsilon}[\mathrm{sign}(p(\mathbf{x})) \neq \mathrm{sign}(p(\mathbf{y}))] \leq \mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}_n}[f(\mathbf{x}) \neq \mathrm{sign}(p(\mathbf{x}))] + \mathbf{Pr}_{\mathbf{y}\sim\mathcal{N}_n}[f(\mathbf{y}) \neq \mathrm{sign}(p(\mathbf{y}))] = 2\,\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - \mathrm{sign}(p(\mathbf{x}))|$. Combining these, we get that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - p(\mathbf{x})|] \geq (\mathrm{GNS}_\epsilon(f) - \mathrm{GNS}_\epsilon(\mathrm{sign}(p)))/4$. Note that $\mathrm{sign}(p)$ is a degree-d PTF, and therefore by Kane (2010) it holds that $\mathrm{GNS}_\epsilon(\mathrm{sign}(p)) = O(d\sqrt{\epsilon})$.

For the first bound, let $\mathbf{y}$ and $\mathbf{z}$ be independent Gaussians and let $\mathbf{x}(\phi) = \cos\phi\,\mathbf{y} + \sin\phi\,\mathbf{z}$. Let a be a sufficiently small multiple of $\log\, d/d$. For any $\phi \in [0, 2\pi]$, $\mathbf{x}(\phi - a)$ and $\mathbf{x}(\phi+a)$ are $(1-\delta)$-correlated Gaussian random variables, where $\delta = \Theta(\log d/d)^2$. We have that

$$\begin{aligned} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - p(\mathbf{x})|] &= \mathop{\mathbf{E}}_{\phi\in\mathcal{U}([0,2\pi])}\left[\mathop{\mathbf{E}}_{\mathbf{y},\mathbf{z}\sim\mathcal{N}_n}[|f(\mathbf{x}(\phi)) - p(\mathbf{x}(\phi))|]\right] = \mathop{\mathbf{E}}_{\mathbf{y},\mathbf{z}\sim\mathcal{N}_n}\left[\mathop{\mathbf{E}}_{\phi\in\mathcal{U}([0,2\pi])}[|f(\mathbf{x}(\phi)) - p(\mathbf{x}(\phi))|]\right] \\ &\geq \Omega(1/\log(d)) \mathop{\mathbf{E}}_{\mathbf{y},\mathbf{z}\sim\mathcal{N}_n}[\mathop{\mathbf{Pr}}_{\phi\in\mathcal{U}([0,2\pi])}[f(\mathbf{x}(\phi - a)) \neq f(\mathbf{x}(\phi+a))]]\,, \end{aligned}$$

where in the inequality we used Proposition 25. Moreover, using Fubini's theorem, we have

$$\begin{aligned} &\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[|f(\mathbf{x}) - p(\mathbf{x})|] \geq \Omega(1/\log(d)) \mathop{\mathbf{E}}_{\phi\in\mathcal{U}([0,2\pi])}\left[\mathop{\mathbf{Pr}}_{\mathbf{y},\mathbf{z}\sim\mathcal{N}_n}[f(\mathbf{x}(\phi - a)) \neq f(\mathbf{x}(\phi+a))]\right] \\ &= \Omega(1/\log(d)) \mathop{\mathbf{E}}_{\phi\in\mathcal{U}([0,2\pi])}[\mathrm{GNS}_\delta(f)] = \Omega(1/\log(d))\mathrm{GNS}_\delta(f) = \Omega(1/\log(d))\mathrm{GNS}_{(\log(d)/d)^2}(f)\,. \end{aligned}$$

■

Acknowledgments

Ilias Diakonikolas was supported by NSF Award CCF-1652862 (CAREER), a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant. Daniel M. Kane was supported by NSF Award CCF-1553288 (CAREER) and a Sloan Research Fellowship. Nikos Zarifis was supported in part by a DARPA Learning with Less Labels (LwLL) grant.

References

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.

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

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.

D. Dachman-Soled, V. Feldman, L.-Y. Tan, A. Wan, and K. Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In Piotr Indyk, editor, *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 498–511. SIAM, 2015.

A. Daniely. A PTAS for agnostically learning halfspaces. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, pages 484–502, 2015.

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.

I. Diakonikolas, P. Gopalan, R. Jaiswal, R. Servedio, and E. Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010a.

I. Diakonikolas, P. Harsha, A. Klivans, R. Meka, P. Raghavendra, R. A. Servedio, and L. Y. Tan. Bounding the average sensitivity and noise sensitivity of polynomial threshold functions. In *STOC*, pages 533–542, 2010b.

I. Diakonikolas, D. M. Kane, and J. Nelson. Bounded independence fools degree-2 threshold functions. In *FOCS*, pages 11–20, 2010c.

I. Diakonikolas, P. Raghavendra, R. A. Servedio, and L. Y. Tan. Average sensitivity and noise sensitivity of polynomial threshold functions. *SIAM J. Comput.*, 43(1):231–253, 2014.

I. Diakonikolas, R. Jaiswal, R. A. Servedio, L. Y. Tan, and A. Wan. Noise stable halfspaces are close to very small juntas. *Chicago Journal OF Theoretical Computer Science*, 4:1–13, 2015.

I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84, 2017. doi: 10.1109/FOCS.2017.16.

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.

I. Diakonikolas, S. Goel, S. Karmalkar, A. R. Klivans, and M. Soltanolkotabi. Approximation schemes for relu regression. In *Conference on Learning Theory, COLT 2020*, volume 125 of *Proceedings of Machine Learning Research*, pages 1452–1485. PMLR, 2020a.

I. Diakonikolas, D. M. Kane, V. Kontonis, and N. Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In *Conference on Learning Theory*, pages 1514–1539. PMLR, 2020b.

I. Diakonikolas, D. M. 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, December 6-12, 2020, virtual*, 2020c.

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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020d.

K. Fan. On infinite systems of linear inequalities. *Journal of Mathematical Analysis and Applications*, 1968. doi: https://doi.org/10.1016/0022-247X(68)90255-2.

V. Feldman. Statistical query learning. In *Encyclopedia of Algorithms*, pages 2090–2095. 2016.

V. Feldman. A general characterization of the statistical query complexity. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 785–830. PMLR, 2017.

V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami. New results for learning noisy parities and halfspaces. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 563–574. IEEE Computer Society, 2006.

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.

V. Feldman, W. Perkins, and S. Vempala. On the complexity of random satisfiability problems with planted solutions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC, 2015*, pages 77–86, 2015.

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.

M. I. Ganzburg. Limit theorems for polynomial approximation with hermite and freud weights. *Approximation Theory X: Abstract and Classical Analysis (CK Chui, et al, eds.)*, pages 211–221, 2002.

M. I. Ganzburg and J. Rognes. *Limit theorems of polynomial approximation with exponential weights*. American Mathematical Soc., 2008.

S. Goel, V. Kanade, A. R. Klivans, and J. Thaler. Reliably learning the relu in polynomial time. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 1004–1042, 2017.

S. Goel, S. Karmalkar, and A. R. Klivans. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 8582–8591, 2019.

S. Goel, A. Gollakota, Z. Jin, S. Karmalkar, and A. Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pages 3587–3596. PMLR, 2020a.

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*, 2020b.

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

P. Harsha, A. R. Klivans, and R. Meka. Bounding the sensitivity of polynomial threshold functions. *Theory of Computing*, 10:1–26, 2014.

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

A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008. Special issue for FOCS 2005.

D. M. Kane. The average sensitivity of an intersection of half spaces. In *Symposium on Theory of Computing, STOC 2014*, pages 437–440, 2014.

D.M. Kane. The Gaussian surface area and noise sensitivity of degree-d polynomial threshold functions. In *CCC*, pages 205–210, 2010.

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

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

A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via Gaussian surface area. In *Proc. 49th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 541–550, Philadelphia, Pennsylvania, 2008.

A. Klivans, P. Long, and R. Servedio. Learning Halfspaces with Malicious Noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

P. Manurangsi and D. Reichman. The computational complexity of training relu (s). *arXiv preprint arXiv:1810.04207*, 2018.

E. Nelson. The free markoff field. *Journal of Functional Analysis*, 12(2):211–227, 1973.

R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. ISBN 978-1-10-703832-5.

G. Szegö. *Orthogonal Polynomials*. Number τ. 23 in American Mathematical Society colloquium publications. American Mathematical Society, 1967. ISBN 9780821889527. URL https://books.google.com/books?id=3hcW8HBh7gsC.

B. Szörényi. Characterizing statistical query learning: Simplified notions and proofs. In *Algorithmic Learning Theory, 20th International Conference, ALT 2009*, volume 5809 of *Lecture Notes in Computer Science*, pages 186–200. Springer, 2009.

L. G. Valiant. A theory of the learnable. In *Proc. 16th Annual ACM Symposium on Theory of Computing (STOC)*, pages 436–445. ACM Press, 1984.

Appendix A. Omitted Background

A.1. Correlational Statistical Query (CSQ) Model

For some of our lower bounds in the real-valued setting, we consider *correlational* or inner product queries. The CSQ model is a restriction of the SQ model, where the algorithm is allowed to choose any bounded query function, and obtain estimates for its correlation with the labels. Specifically, for $f, h : X \to \mathbb{R}$ and a distribution D over the domain X, we denote by $\langle f, h \rangle_D$ the quantity $\mathbf{E}_{\mathbf{x}\sim D}[f(\mathbf{x})h(\mathbf{x})]$ and refer to it as the correlation of f and h under D. While it is commonly assumed that the query function h is pointwise bounded, it is in fact sufficient to assume that it has bounded L^2-norm. If $\mathcal{D}$ is the joint distribution on points and labels, a correlational query takes h and a parameter $t > 0$, and outputs a value $v \in [\mathbf{E}_{(\mathbf{x},y)\sim D}[h(\mathbf{x})y] - \tau, \mathbf{E}_{(\mathbf{x},y)\sim D}[h(\mathbf{x})y] + \tau]$. Similarly to the general SQ model, we consider the following notions of statistical dimension.

Definition 28 (Correlational Statistical Query Dimension) *For $\beta, \gamma > 0$, a probability distribution D over domain X and a family $\mathcal{C}$ of functions $f : X \to \mathbb{R}$, let s be the maximum integer for which there exists a finite set of functions $\{f_1, \dots, f_s\} \subseteq \mathcal{C}$ such that $|\mathbf{E}_{\mathbf{x}\sim D}[f_i^2(\mathbf{x})]| \le \beta$ for all $i \in [s]$ and $|\mathbf{E}_{\mathbf{x}\sim D}[f_i(\mathbf{x})f_j(\mathbf{x})]| \le \gamma$ for all $i, j \in [s]$ with $i \neq j$. We define the* Correlational Statistical Query Dimension *with pairwise correlations (γ, β) of $\mathcal{C}$ to be s and denote it by* $\mathrm{CSD}_D(\mathcal{C}, \gamma, \beta)$.

Definition 29 (Average Correlational Statistical Query Dimension) *Let $\rho > 0$, let D be a probability distribution over some domain X, and let $\mathcal{C}$ be a family of functions $f : X \to \mathbb{R}$. We define the average pairwise correlation of functions in $\mathcal{C}$ to be $\rho(\mathcal{C}) = \frac{1}{|\mathcal{C}|^2} \sum_{g,r\in\mathcal{C}} |\mathbf{E}_{\mathbf{x}\sim D}[g(\mathbf{x})r(\mathbf{x})]|$. The* Average Correlational Statistical Query Dimension *of $\mathcal{C}$ relative to D with parameter γ, denoted by $\mathrm{CSDA}_D(\mathcal{C}, \gamma)$, is defined to be the largest integer s such that every subset $\mathcal{C}' \subseteq \mathcal{C}$ of size $|\mathcal{C}'| \ge |\mathcal{C}|/s$, satisfies $\rho(\mathcal{C}') \ge \rho$.*

In most of the cases, it suffices to bound the correlational statistical query dimension, since by simple calculations this implies a bound on the average statistical query dimension.

Lemma 30 *Let $\mathcal{C}$ be a class of functions and D be a distribution and suppose that $\mathrm{CSD}_D(\mathcal{C}, \gamma, \beta) = d$ for some $\gamma, \beta > 0$. Then, for all $\gamma' > 0$, $\mathrm{CSD}_D(\mathcal{C}, \gamma + \gamma') \ge d\gamma'/(\beta - \gamma)$.*

The following result Szörényi (2009); Goel et al. (2020a) relates the Average Correlational SQ dimension of a concept class with the complexity of any CSQ algorithm for the class.

Lemma 31 (Theorem B.1 from Goel et al. (2020a)) *Let D be a distribution over a domain X and let $\mathcal{C}$ be a real-valued concept class over X such that $0 \in \mathcal{C}$, and $\|f\|_2 \geq \eta$ for all $f \in \mathcal{C}, f \neq 0$. Suppose that for some $\gamma > 0$ we have $s = \mathrm{CSDA}_D(\mathcal{C}, \gamma)$. Any CSQ algorithm that outputs a hypothesis h such that $\|h - f\|_2 < \eta$ needs at least $s/2$ queries or queries of tolerance $\sqrt{\gamma}$.*

A.2. Preliminaries: Multilinear Algebra

Here we introduce some multilinear algebra notation. An order k tensor $\mathbf{A}$ is an element of the k-fold tensor product of subspaces $\mathbf{A} \in \mathcal{V}_1 \otimes \ldots \otimes \mathcal{V}_k$. We will be exclusively working with subspaces of $\mathbb{R}^d$ so a tensor A can be represented by a sequence of coordinates, that is $A_{i_1,\ldots,i_k}$. The tensor product of a order k tensor $\mathbf{A}$ and an order m tensor $\mathbf{B}$ is an order $k+m$ tensor defined as $(\mathbf{A} \otimes \mathbf{B})_{i_1,\ldots,i_k,j_1,\ldots,j_m} = \mathbf{A}_{i_1,\ldots,i_k}\mathbf{B}_{j_1,\ldots,j_m}$. We are also going to use capital letters for multi-indices, that is tuples of indices $I = (i_1, \ldots, i_k)$. We denote by E_i the multi-index that has 1 on its i-th co-ordinate and 0 elsewhere. For example the previous tensor product can be denoted as $\mathbf{A}_I\mathbf{B}_J$. To simplify notation we are also going to use Einstein's summation where we assume that we sum over repeated indices in a product of tensors. For example if $\mathbf{A} \in \mathbb{R}^d \otimes \mathbb{R}^d$, $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{u} \in \mathbb{R}^d$ we have $\sum_{i,j=1}^d \mathbf{v}_i\mathbf{u}_j\mathbf{A}_{ij} = \mathbf{v}_i\mathbf{u}_j\mathbf{A}_{ij}$. We define the dot product of two tensors (of the same order) to be $\langle \mathbf{A}, \mathbf{B} \rangle = \mathbf{A}_{i_1,\ldots,i_k}\mathbf{B}_{i_1,\ldots,i_k} = \mathbf{A}_I\mathbf{B}_I$. We also denote the ℓ_2-norm of a tensor by $\|\mathbf{A}\|_2 = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}$. We denote by $\mathbf{A}(\mathbf{X})$ a function that maps the tensor $\mathbf{X}$ to a tensor $\mathbf{A}(\mathbf{X})$. Let $\mathcal{V}$ be a vector space and let $\mathbf{A}(\mathbf{x}) : \mathbb{R}^d \to \mathcal{V}^{\otimes k}$ be a tensor valued function. We denote by $\partial_i\mathbf{A}(\mathbf{x})$ the tensor of partial derivatives of $A(\mathbf{x})$, $\partial_i\mathbf{A}(\mathbf{x}) = \partial_i\mathbf{A}_J(\mathbf{x})$ is a tensor of order $k+1$ in $\mathcal{V}^{\otimes k} \otimes \mathbb{R}^d$. We also denote this tensor $\nabla\mathbf{A}(\mathbf{x}) = \partial_i\mathbf{A}_J(\mathbf{x})$. Similarly we define higher-order derivatives, and we denote

$$\nabla^m\mathbf{A}(\mathbf{x}) = \partial_{i_1}\ldots\partial_{i_m}\mathbf{A}_J(\mathbf{x}) \in \mathcal{V}^{\otimes k} \otimes (\mathbb{R}^d)^{\otimes m} .$$

A.3. Basics of Hermite Polynomials

We are also going to use the Hermite polynomials that form an orthonormal system with respect to the Gaussian measure. While, usually one considers the probabilists's or physicists' Hermite polynomials, in this work we define the *normalized* Hermite polynomial of degree i to be $H_0(x) = 1, H_1(x) = x, H_2(x) = \frac{x^2-1}{\sqrt{2}}, \ldots, H_i(x) = \frac{He_i(x)}{\sqrt{i!}}, \ldots$ where by $He_i(x)$ we denote the probabilists' Hermite polynomial of degree i. These normalized Hermite polynomials form a complete orthonormal basis for the single dimensional version of the inner product space L^2. To get an orthonormal basis for L^2, we use a multi-index $J \in \mathbb{N}^d$ to define the d-variate normalized Hermite polynomial as $H_J(\mathbf{x}) = \prod_{i=1}^d H_{v_i}(\mathbf{x}_i)$. The total degree of H_J is $|J| = \sum_{v_i \in J} v_i$. Given a function $f \in L^2(\mathbb{R})$ we compute its Hermite coefficients as $\hat{f}(J) = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[f(\mathbf{x})H_J(\mathbf{x})]$ and express it uniquely as $\sum_{J\in\mathbb{N}^n} \hat{f}(J)H_J(\mathbf{x})$. For more details on the Gaussian space and Hermite Analysis (especially from the theoretical computer science perspective), we refer the reader to O'Donnell (2014). Most of the facts about Hermite polynomials that we use in this work are well known properties and can be found, for example, in Szegö (1967).

We denote by $f^{[k]}(x)$ the degree k part of the Hermite expansion of f, $f^{[k]}(\mathbf{x}) = \sum_{|J|=k} \hat{f}(J) \cdot H_J(\mathbf{x})$. We say that a polynomial q is harmonic of degree k if it is a linear combination of degree k

Hermite polynomials, that is q can be written as

$$q(\mathbf{x}) = q^{[k]}(\mathbf{x}) = \sum_{J:|J|=k} c_J H_J(\mathbf{x})$$

For a single dimensional Hermite polynomial it holds $H_m'(x) = \sqrt{m} H_{m-1}'(x)$. Using this we obtain that for a multivariate Hermite polynomial $H_M(\mathbf{x})$, where $M = (m_1, \ldots, m_n)$ it holds

$$\nabla H_M(\mathbf{x}) = \sqrt{m_i} H_{M-E_i}(\mathbf{x}) \in \mathbb{R}^n, \tag{1}$$

where $E_i = \mathbf{e}_i$ is the multi-index that has 1 position i and 0 elsewhere. From this fact and the orthogonality of Hermite polynomials we obtain

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[\langle \nabla H_M(\mathbf{x}), \nabla H_L(\mathbf{x})\rangle] = |M| \delta_{M,L}. \tag{2}$$

The following fact gives us a formula for the inner product of

Fact 32 *Let $p, q : \mathbb{R}^n \to \mathbb{R}$ be harmonic polynomials of degree k. Then*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}\left[\langle \nabla^\ell p(\mathbf{x}), \nabla^\ell q(\mathbf{x})\rangle\right] = k(k-1)\ldots(k-\ell+1) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[p(\mathbf{x})q(\mathbf{x})].$$

In particular,

$$\langle \nabla^k p(\mathbf{x}), \nabla^k q(\mathbf{x})\rangle = k! \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[p(\mathbf{x})q(\mathbf{x})].$$

Proof Write $p(\mathbf{x}) = \sum_{M:|M|=k} b_M H_M(\mathbf{x})$ and $q(\mathbf{x}) = \sum_{M:|M|=k} c_M H_M(\mathbf{x})$. Since the Hermite polynomials are orthonormal we obtain $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[p(\mathbf{x})q(\mathbf{x})] = \sum_{M:|M|=k} c_M b_M$. Now, using Equation (1) iteratively we obtain

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\langle \nabla^\ell H_M(\mathbf{x}), \nabla^\ell H_L(\mathbf{x})\rangle\right] = k(k-1)\ldots(k-\ell+1)\delta_{M,L}.$$

Using this equality we obtain

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\langle \nabla^\ell p(\mathbf{x}), \nabla^\ell q(\mathbf{x})\rangle\right] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\langle \sum_M b_M \nabla^\ell H_M(\mathbf{x}), \sum_L c_L \nabla^\ell H_L(\mathbf{x})\rangle\right] \\
&= \sum_{M,L} b_M c_L \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}\left[\langle \nabla^\ell H_M(\mathbf{x}), \nabla^\ell H_L(\mathbf{x})\rangle\right] \\
&= \sum_{M,L} b_M c_L k(k-1)\ldots(k-\ell+1)\delta_{M,L}. \\
&= k(k-1)\ldots(k-\ell+1) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}}[p(\mathbf{x})q(\mathbf{x})].
\end{aligned}$$

■

Observe that for every harmonic polynomial $p(\mathbf{x})$ of degree k we have that $\nabla^k p(\mathbf{x})$ is a symmetric tensor of order k. Since the degree of the polynomial is k and we differentiate k times this tensor no longer depends on $\mathbf{x}$. Using Fact 32, we observe that this operation (modulo a division by $\sqrt{k!}$) preserves the L^2-norm of the harmonic polynomial p, that is $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_n}[p^2(\mathbf{x})] = \left\|\nabla^k p(\mathbf{x})\right\|_2^2 / k!$.

Appendix B. Omitted Proofs from Section 2

B.1. Proof of Lemma 15

We restate the lemma here for convenience.

Lemma 33 *Let $f : \mathbb{R}^m \to \mathbb{R}$ and $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m\times n}$ be linear maps such that $\mathbf{U}\mathbf{U}^\intercal = \mathbf{V}\mathbf{V}^\intercal = \mathbf{I}_m$. Then,*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[f(\mathbf{U}\mathbf{x})f(\mathbf{V}\mathbf{x})] \le \sum_{t=0}^{\infty} \|\mathbf{U}\mathbf{V}^\intercal\|_2^t \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[(f^{[t]}(\mathbf{x}))^2]\,,$$

where $f^{[t]}$ denotes the degree-t Hermite part of f.

Proof To simplify notation, write $g_1(\mathbf{x}) = g(\mathbf{U}\mathbf{x})$ and $g_2(\mathbf{x}) = g(\mathbf{V}\mathbf{x})$. Moreover, we will write $g_1(\mathbf{x}) \sim \sum_{k=0}^\infty g_1^{[k]}(\mathbf{x})$ and $g_2(\mathbf{x}) \sim \sum_{k=0}^\infty g_2^{[k]}(\mathbf{x})$. Using Fact 32, we obtain

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[g_1(\mathbf{x})g_2(\mathbf{x})] &= \sum_{k=0}^{\infty} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[g_1^{[k]}(\mathbf{x})g_2^{[k]}(\mathbf{x})] = \sum_{k=0}^{\infty}\frac{1}{k!}\langle \nabla^k g_1^{[k]}(\mathbf{x}), \nabla^k g_2^{[k]}(\mathbf{x})\rangle \\ &= \sum_{k=0}^{\infty}\frac{1}{k!}\langle \nabla^k g^{[k]}(\mathbf{U}\mathbf{x}), \nabla^k g^{[k]}(\mathbf{V}\mathbf{x})\rangle\,. \end{aligned} \tag{3}$$

Denote by $\mathcal{U} \subseteq \mathbb{R}^n$ the image of the linear map $\mathbf{U}^\intercal$. Now observe that, using the chain rule, for any function $h(\mathbf{U}\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ it holds $\nabla h(\mathbf{U}\mathbf{x}) = \partial_i h(\mathbf{U}\mathbf{x})\mathbf{U}_{ij} \in \mathcal{U}$, where we used Einstein's summation notation for repeated indices. Applying the above rule k times, we have that

$$\nabla^k h(\mathbf{U}\mathbf{x}) = \partial_{i_k} \ldots \partial_{i_1} h(\mathbf{U}\mathbf{x})\mathbf{U}_{i_1 j_1}\ldots\mathbf{U}_{i_k j_k} \ \in \mathcal{U}^{\otimes k}\,.$$

We denote $\mathbf{R} = \nabla^k g^{[k]}(\mathbf{x})$ and observe that this tensor does not depend on $\mathbf{x}$. Moreover, denote $\mathbf{M} = \mathbf{U}\mathbf{V}^\intercal$, $\mathbf{S} = \nabla^k g^{[k]}(\mathbf{U}\mathbf{x}) = (\mathbf{U}^\intercal)^{\otimes k}\mathbf{R} \in \mathcal{U}^{\otimes k}$, and $\mathbf{T} = \nabla^k g^{[k]}(\mathbf{V}\mathbf{x}) = (\mathbf{V}^\intercal)^{\otimes k}\mathbf{R} \in \mathcal{V}^{\otimes k}$. We have that

$$\langle \mathbf{S}, \mathbf{T}\rangle = \langle (\mathbf{U}^\intercal)^{\otimes k}\mathbf{R}, (\mathbf{V}^\intercal)^{\otimes k}\mathbf{R}\rangle = \langle \mathbf{R}, \mathbf{M}^{\otimes k}\mathbf{R}\rangle \le \left\|\mathbf{M}^{\otimes k}\right\|_2 \|\mathbf{R}\|_2^2 = k!\,\|\mathbf{M}\|_2^k \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[(g^{[k]}(\mathbf{x}))^2]\,,$$

where to get the last equality we used again Fact 32. To finish the proof, we combine this inequality with Equation (3). ■

B.2. Proof of Corollary 16

Corollary 34 *Let $d \ge 2$ and D be a distribution over $\mathbb{R}^m$ such that the first $(d-1)$ moments of D match the corresponding moments of $\mathcal{N}_m$. Let $G(\mathbf{x}) = D(\mathbf{x})/\phi_m(\mathbf{x})$ be the ratio of the corresponding probability density functions. For matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{m\times n}$ such that $\mathbf{U}\mathbf{U}^\intercal = \mathbf{V}\mathbf{V}^\intercal = \mathbf{I}_m$, define $D_{\mathbf{U}}$ and $D_{\mathbf{V}}$ to have probability density functions $G(\mathbf{U}\mathbf{x})\phi_n(\mathbf{x})$ and $G(\mathbf{V}\mathbf{x})\phi_n(\mathbf{x})$, respectively. Then, we have that $|\chi_{\mathcal{N}_n}(D_{\mathbf{U}}, D_{\mathbf{V}})| \le \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \chi^2(D, \mathcal{N}_m)$.*

Proof We compute

$$\chi_{\mathcal{N}_n}(D_{\mathbf{U}}, D_{\mathbf{V}}) = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}\left[\frac{(D_{\mathbf{U}}(\mathbf{x}) - \phi_n(\mathbf{x}))(D_{\mathbf{V}}(\mathbf{x}) - \phi_n(\mathbf{x}))}{\phi_n^2(\mathbf{x})}\right] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[(G(\mathbf{U}\mathbf{x}) - 1)(G(\mathbf{V}\mathbf{x}) - 1)] .$$

We then apply Lemma 15 to the function $g(\mathbf{x}) = G(\mathbf{x}) - 1$. Note that the assumption that D matches the first $d-1$ moments with $\mathcal{N}_m$ is equivalent to saying that $g^{[t]} = 0$ for $t < d$. Thus, Lemma 15 implies that

$$\begin{aligned} |\chi_{\mathcal{N}_n}(D_{\mathbf{U}}, D_{\mathbf{V}})| &\le \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \sum_{t=0}^{\infty} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[(g^{[t]}(\mathbf{x}))^2] = \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g^2(\mathbf{x})] \\ &\le \|\mathbf{U}\mathbf{V}^\intercal\|_2^d \chi^2(D, \mathcal{N}_m) , \end{aligned}$$

where the equality is Parseval's identity and in the last inequality we used the definition of G. ■

B.3. Proof of Lemma 17

We restate the lemma below.

Lemma 35 *Let $0 < a, c < 1/2$ and $m, n \in \mathbb{Z}_+$ such that $m \le n^a$. There exists a set S of $2^{\Omega(n^c)}$ matrices in $\mathbb{R}^{m\times n}$ such that every $\mathbf{U} \in S$ satisfies $\mathbf{U}\mathbf{U}^\intercal = \mathbf{I}_m$ and every pair $\mathbf{U}, \mathbf{V} \in S$ with $\mathbf{U} \neq \mathbf{V}$ satisfies $\|\mathbf{U}\mathbf{V}^\intercal\|_F \le O(n^{2c-1+2a})$.*

Proof Our proof relies on the following fact that there exist exponentially many nearly orthogonal unit vectors.

Fact 36 (see, e.g., Lemma 3.7 of Diakonikolas et al. (2017)) *For any $0 < c < 1/2$ there exists a set S' of $2^{\Omega(n^c)}$ unit vectors in $\mathbb{R}^n$ such that any pair $\mathbf{u}, \mathbf{v} \in S'$, with $\mathbf{u} \neq \mathbf{v}$, satisfies $|\langle \mathbf{u}, \mathbf{v}\rangle| < O(n^{c-1/2})$.*

Let S' be the set of unit vectors that Fact 36 constructs. We group them into sets of size m and use the vectors of each group as rows for each matrix that we make. Thus, we create at least $|S'|/n^a = 2^{\Omega(n^c)}$ many matrices. Next, we ortho-normalize each matrix $\mathbf{V} \in S'$ using the Gram-Schmidt process, in order to get $\mathbf{V}\mathbf{V}^\intercal = \mathbf{I}_m$. In every row of $\mathbf{V}$, the Gram-Schmidt algorithm adds at most m orthogonal vectors, each having norm $O(n^{c-1/2})$. Thus, the total correction term for each row has norm at most $\sqrt{m}O(n^{c-1/2})$. Putting everything together, we have that for all $\mathbf{U}, \mathbf{V}$ obtained that way,

$$\|\mathbf{U}\mathbf{V}^\intercal\|_F \le \left(m^2 m^2 O(n^{4(c-1/2)})\right)^{1/2} = O\left(n^{2c-1+2a}\right) .$$

■

B.4. Omitted Proof from Proposition 19

Claim 37 *It holds that $\chi^2(A, \mathcal{N}_m) = 1$.*

Proof

$$\chi^2(A, \mathcal{N}_m) = \int_{\mathbb{R}^m} \frac{A^2(\mathbf{z})}{\phi_m(\mathbf{z})} \mathrm{d}\mathbf{z} - 1 = \int_{\mathbb{R}^m} \frac{\phi_m^2(\mathbf{z}) \, \mathbf{Pr}^2[y = 1 | \mathbf{x} = \mathbf{z}]}{\phi_m(\mathbf{z}) \, \mathbf{Pr}^2[y = 1]} \mathrm{d}\mathbf{z} - 1$$
$$\leq 4 \int_{\mathbb{R}^m} \phi_m(\mathbf{z}) \, \mathbf{Pr}[y = 1 | \mathbf{x} = \mathbf{z}] \mathrm{d}\mathbf{z} - 1 = 4 \, \mathbf{Pr}[y = 1] - 1 = 1 \; ,$$

where we used the definition of A, Bayes' rule and the fact that $\mathbf{Pr}[y = 1] = 1/2$. ■

Appendix C. Omitted Proofs from Section 3

C.1. Low-Degree Polynomial Approximation to the Sign Function

In this subsection, we record a known powerful theorem from the approximation theory literature by Ganburg Ganzburg (2002); Ganzburg and Rognes (2008). This result can be used to derive tight polynomial degree lower bounds for the sign function and the ReLU function, used in a subsequent section. Let $A_\sigma(f)_p = \inf_{g \in B_\sigma} \|f - g\|_p$, where B_σ, $\sigma > 0$ is the class of all entire functions of exponential type σ, i.e., the class consisting of every entire function g such that for every $\epsilon > 0$ there exists a C for which $|g(z)| \leq C e^{\sigma(1+\epsilon)|z|}$.

Fact 38 *For any function $f : \mathbb{R} \to \mathbb{R}$ of polynomial growth*

$$\lim_{n \to \infty} \left(\frac{b_n}{\sigma} \right)^{1/p} \inf_{p \in \mathcal{P}_n} \left\| f\left(\frac{b_n}{\sigma} x \right) - p(x) \right\|_p = A_\sigma(f)_p \; ,$$

where $b_n = 2\sqrt{n}$, $p \in [1, 2]$ and $A_\sigma(f)_p$ is the error of best approximation of f by entire functions of exponential type σ in $L^p(\mathbb{R})$.

By selecting $f(t) = \mathrm{sign}(t)$ and $p = 1$, we get that any polynomial that achieves error at most ϵ with respect to the L^1-norm must have degree at least $\Omega(1/\epsilon^2)$.

Corollary 39 *Let $f : \mathbb{R} \to \{\pm 1\}$ with $f(t) = \mathrm{sign}(t)$. Any polynomial $p : \mathbb{R} \to \mathbb{R}$ satisfying $\|f - p\|_1 \leq \epsilon$ must have degree $d = \Omega(1/\epsilon^2)$.*

C.2. Proof of Lemma 22

We restate the lemma below.

Lemma 40 *There exists a distribution D that (i) matches d moments with $\mathcal{N}$, (ii) the pdf of D is at most 2 times the pdf of $\mathcal{N}$ on every point in $\mathbb{R}$ and (iii) for some $\alpha = \Theta(1/\sqrt{d})$*

$$\mathbf{Pr}[(X \mod a) \in (a/2, a)] = 2^{-\Omega(d)} \; .$$

First, we need the following lemma.

Lemma 41 *There is a d-wise independent family of $t = O(d)$ standard Gaussians $X_1, X_2, \ldots, X_t$ such that $\left(\sum_{i=1}^t X_i\right) \mod 1 \in [0, 1/2]$ with probability $1 - 2^{-\Omega(d)}$. Furthermore, such a distribution can be obtained by rejection sampling a set of independent standard Gaussians, where a sample is rejected with probability $1/2$.*

Proof The standard Gaussian distribution can be decomposed into a uniform component and a remaining term. That is, $\mathcal{N} = c\mathcal{U}([0, 1]) + (1 - c)E$, where $\mathcal{U}([0, 1])$ is the uniform distribution in $[0, 1]$, E is another distribution, and $c > 0$ is a constant. Let $t \in \mathbb{N}$ such that $t > d/c$. We generate this d-wise independent family $X_1, \ldots, X_t$ as follows.

First, we sample $Y_1, \ldots, Y_t$ independent standard Gaussians, writing each Y_i either as a sample from $\mathcal{U}([0, 1])$ or a sample from E. Then, two complementary cases are considered.

> **Case 1**. The number of Y_i's that came from $\mathcal{U}([0, 1])$ is at most d. In this case, the sample is rejected with probability $1/2$.
>
> **Case 2**. Otherwise, the sample is rejected if and only if $(\sum_{i=1}^t Y_i) \mod 1 \in (1/2, 1]$.

Let $X_1, \ldots, X_t$ be the output of this rejection sampling procedure. The probability that the sample is generated by the first case of the algorithm is exponentially small. To see this, define $Z_i \in \{0, 1\}$ to be one if and only Y_i is drawn from $\mathcal{U}([0, 1])$. If C_1 denotes the event of being in Case 1, then by standard Chernoff bounds we have that

$$\begin{aligned}\mathbf{Pr}[C_1] = \mathbf{Pr}\left[\sum_{i=1}^t Z_i \le d\right] &= \mathbf{Pr}\left[\sum_{i=1}^t Z_i \le \mathbf{E}\left[\sum_{i=1}^t Z_i\right]\left(1 - \left(1 - \frac{d}{tc}\right)\right)\right] \\ &\le \exp\left(-\frac{(1 - d/(tc))^2 tc}{2}\right) = 2^{-\Omega(d)} ,\end{aligned}$$

where we used that $t > d/c$. Therefore, the probability that $(\sum_{i=1}^t X_i) \mod 1 \in [0, 1/2]$ is $1 - 2^{-\Omega(d)}$.

Moreover, the probability of accepting the sample is exactly $1/2$ independently of the Y_i's. To see this, let $C_1, C_2 = \overline{C_1}$ be the events of Case 1 and Case 2 being true respectively, and A be the event of accepting the sample. For Case 1, we have $\mathbf{Pr}[A|C_1] = 1/2$. In Case 2, we know that at least one element is drawn from $\mathcal{U}([0, 1])$, which means that the $(\sum_{i=1}^t X_i) \mod 1$ is going to be uniform in $[0, 1]$. Thus, $\mathbf{Pr}[A|C_2] = 1/2$. Therefore, $\mathbf{Pr}[C_1|A] = \mathbf{Pr}[A|C_1]\mathbf{Pr}[C_1]/\mathbf{Pr}[A] = \mathbf{Pr}[C_1]$ and $\mathbf{Pr}[C_2|A] = \mathbf{Pr}[A|C_2]\mathbf{Pr}[C_2]/\mathbf{Pr}[A] = \mathbf{Pr}[C_2]$, i.e., accepting is independent of C_1, C_2, and thus independent of the sample itself. This means that the output $X_1, \ldots, X_t$ remains Gaussian.

For the d-wise independence of the variables $X_1, \ldots, X_t$, let $\mathcal{I}$ be an arbitrary set of at most d indices from $\{1, \ldots, t\}$. We claim that $\{X_i\}_{i \in \mathcal{I}}$ are independent. Case 1 is trivial, since we accept independently of the values of the Y_i's. For Case 2, note that in that case there are more than d Y_i's drawn from $\mathcal{U}([0, 1])$. This means that there exists one $j \notin \mathcal{I}$ such that Y_j is uniform and forces the $(\sum_{i=1}^t X_i)$ to be uniform in $[0, 1]$. Thus, the event $(\sum_{i=1}^t X_i) \in [0, 1/2]$ is independent of $\{Y_i\}_{i \in \mathcal{I}}$, and therefore $\{X_i\}_{i \in \mathcal{I}}$ is a set of independent random variables. ■

Proof [Proof of Lemma 22] Consider the random variable $X = \sum_{i=1}^t X_i/\sqrt{t}$ for the X_i's of Lemma 41. For (i), note that the d-th moment involves the expectation of at most d of the X_i's, which are independent. Note that (ii) holds because the distribution of X puts almost all of its mass on half of the real line, and (iii) follows from our scaling of $1/\sqrt{t}$. ■

C.3. Proof of Proposition 21

Proof We can assume that k is even. Let f be 1 on the $k/2$ intervals $(ia + a/2, (i+1)a)$, for $i = 0, \ldots, k/2 - 1$, and zero elsewhere. Denote by D the distribution of Lemma 22. From property (iii), we have that $\mathbf{E}_{x\sim D}[f(x)] = 2^{-\Omega(d)}k$. On the other hand, assuming that $k = O(\sqrt{d})$, we have that $\mathbf{E}_{x\sim\mathcal{N}}[f(x)] = \Omega(k/\sqrt{d})$. This is because the regions where f is 1 are contained in the interval $[0, \Theta(k/\sqrt{d})] \subseteq [0, O(1)]$, where the pdf of the standard Gaussian is bounded below by some constant.

Let $D(x)$ and $\phi(x)$ denote the density on point x of the distribution D and $\mathcal{N}$ respectively. For every polynomial $p : \mathbb{R} \to \mathbb{R}$ of degree at most d, it holds

$$\mathop{\mathbf{E}}_{x\sim\mathcal{N}}[f(x)] - \mathop{\mathbf{E}}_{x\sim D}[f(x)] = \mathop{\mathbf{E}}_{x\sim\mathcal{N}}\left[f(x)\left(1 - \frac{D(x)}{\phi(x)}\right)\right] = \mathop{\mathbf{E}}_{x\sim\mathcal{N}}\left[(f(x) - p(x))\left(1 - \frac{D(x)}{\phi(x)}\right)\right]$$
$$\leq \mathop{\mathbf{E}}_{x\sim\mathcal{N}}[|f(x) - p(x)|] ,$$

where the second equality follows from the fact that D matches its first d moments with $\mathcal{N}$, and in the last inequality we used that $0 \leq D(x) \leq 2\phi(x)$ for all $x \in \mathbb{R}$. Thus, if f could be L^1-approximated to error ϵ by a degree-d polynomial, then $\mathbf{E}_{x\sim\mathcal{N}}[f(x)] - \mathbf{E}_{x\sim D}[f(x)]$ would be at most ϵ. But we already showed that this is $\Omega(k/\sqrt{d})$, which implies that $d = \Omega(k^2/\epsilon^2)$. ■

C.4. Proof of Lemma 24

We restate the lemma below.

Lemma 42 *There exists an intersection of k halfspaces on* $\mathbb{R}^k$, $f : \mathbb{R}^k \to \{\pm 1\}$ *such that* $\mathrm{GNS}_\epsilon(f) = \Omega(\sqrt{\epsilon \log k})$.

Proof We will exhibit a family of k halfspaces whose intersection has the claimed Gaussian noise sensitivity. In particular, these halfspaces will be orthogonal. For $i \in [k]$, let $f_i : \mathbb{R}^n \to \{\pm 1\}$ with $f_i(\mathbf{x}) = \mathrm{sign}(-\langle \mathbf{e}_i, \mathbf{x}\rangle + \theta)$, where $\mathbf{e}_i$ is the vector having 1 in the i-th coordinate and 0 elsewhere, and $\theta > 0$ is the bias. That is, f_i is 1 if and only if the i-th coordinate is less than θ.

Fix an index $i \in [k]$. The Gaussian noise sensitivity of a single halfspace is $\mathrm{GNS}_\epsilon(f_i) = \Omega(e^{-\frac{\theta^2}{2(1-\epsilon/2)}}\sqrt{\epsilon})$ (see, e.g., (Diakonikolas et al., 2015, Lemma 3.4) for a proof). Let $\mathbf{x}, \mathbf{y}$ be two $(1-\epsilon)$-correlated n-dimensional standard Gaussian random variables. Then, the inner products $\langle \mathbf{e}_i, \mathbf{x}\rangle$ and $\langle \mathbf{e}_i, \mathbf{y}\rangle$ are $(1-\epsilon)$-correlated univariate Gaussians. Since the Gaussian noise sensitivity of f_i is proportional to the probability that $\langle \mathbf{e}_i, \mathbf{x}\rangle < \theta < \langle \mathbf{e}_i, \mathbf{y}\rangle$, we have that

$$\mathop{\mathbf{Pr}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^{1-\epsilon}}[\langle \mathbf{e}_i, \mathbf{x}\rangle < \theta < \langle \mathbf{e}_i, \mathbf{y}\rangle] = \Omega(e^{-\frac{\theta^2}{2(1-\epsilon/2)}}\sqrt{\epsilon}) .$$

Let θ be the threshold for which $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}_n}[\langle \mathbf{e}_i, \mathbf{x}\rangle > \theta] = 1/k$. The standard bound for the Gaussian tail is $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{N}_n}[\langle \mathbf{e}_i, \mathbf{x}\rangle > \theta] = \Theta(e^{-\theta^2/2}/\theta)$. Therefore, for the θ that we selected it holds $\mathbf{Pr}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^{1-\epsilon}}[\langle \mathbf{e}_i, \mathbf{x}\rangle < \theta < \langle \mathbf{e}_i, \mathbf{y}\rangle] = \Omega(\theta\sqrt{\epsilon}/k) = \Omega(\sqrt{\epsilon \log k}/k)$.

Let $f : \mathbb{R}^n \to \{\pm 1\}$ be 1 if and only if f_i is 1 for all $i \in [k]$. Then, we have that

$$\mathrm{GNS}_\epsilon(f) = 2 \mathop{\mathbf{Pr}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^{1-\epsilon}}[f(\mathbf{x}) = 1, f(\mathbf{y}) = -1] = \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}_n}[f(\mathbf{x}) = 1] - \mathop{\mathbf{Pr}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{N}_n^{1-\epsilon}}[f(\mathbf{x}) = f(\mathbf{y}) = 1]$$
$$= \left(1 - \frac{1}{k}\right)^k - \left(1 - \frac{1}{k} - \Omega\left(\frac{\sqrt{\epsilon \log k}}{k}\right)\right)^k ,$$

where the k-th powers are due to the fact that $\langle \mathbf{e}_i, \mathbf{x}\rangle$ and $\langle \mathbf{e}_j, \mathbf{x}\rangle$ are independent for $i \neq j$. We can use the Taylor expansion to show that the above difference is $\Omega(\sqrt{\epsilon \log k})$. Let the function $h(t) = (1-1/k+t)^k$. By Taylor's theorem, $h(0)-h(t) = -h'(0)t - h''(\xi)t^2/2$, for some ξ between t and 0. By calculating the derivatives, setting $t = -\Omega(\sqrt{\epsilon \log k}/k)$ and noting that the second term of the approximation is less than the first one, we get that $h(0) - h(t) = \Omega\left(\frac{\sqrt{\epsilon \log k}}{k}\right) k \left(1 - \frac{1}{k}\right)^{k-1}$.
■

C.5. Proof of Claims 26 and 27

Claim 43 *For any $\theta \in [0, 2\pi]$, it holds $|p^{(k)}(\theta)| = O(d)^k$.*

Proof Using $\cos\theta = \left(e^{i\theta} + e^{-i\theta}\right)/2$ and $\sin\theta = \left(e^{i\theta} - e^{-i\theta}\right)/2$, we write $p(\theta) = \sum_{n=-\infty}^{\infty} a_n e^{ni\theta}$, for some coefficients a_n, where $a_n = \frac{1}{2\pi}\int_0^{2\pi} p(\phi)e^{-ni\theta}\mathrm{d}\phi$. Since p has degree at most d, it holds that $a_n = 0$, for all $n > d$ and $n < -d$. Therefore, we have that $p(\theta) = \sum_{n=-d}^{d} \frac{1}{2\pi}\int_0^{2\pi} p(\phi)e^{ni(\theta-\phi)}\mathrm{d}\phi$. Taking the k-th derivative (using Leibniz's rule) gives

$$p^{(k)}(\theta) = \sum_{n=-d}^{d} \frac{1}{2\pi}\int_0^{2\pi} p(\phi)(ni)^k e^{ni(\theta-\phi)}\mathrm{d}\phi \,.$$

This implies that

$$|p^{(k)}(\theta)| \leq \sum_{n=-d}^{d} \frac{1}{2\pi}\int_0^{2\pi} |p(\phi)| n^k \mathrm{d}\phi \leq 2\sum_{n=-d}^{d} n^k = O(d^{k+1})\,.$$

Moreover, k is proportional to $\log d$, thus $|p^{(k)}(\theta)| = O(d)^k$, for all $\theta \in [0, 2\pi]$. ■

Claim 44 *It holds that $|b_k| \leq 1/(4k)$.*

Proof Note that $b_k = (1/2\pi)\int_0^{2\pi} R(\theta)e^{-ki\theta}\mathrm{d}\theta$. Using the orthogonality of the trigonometric polynomials, only terms containing $\cos(k\theta)$ are non-zero. Moreover, $\cos^k\theta = \sum_{j=0}^{k} u_j \cos(j\theta)$ with $u_k = 2^{-k+1}$, which can be verified using the identity $\cos\theta = (e^{i\theta} + e^{-i\theta})/2$. Therefore, we have that

$$b_k = \frac{1}{2\pi}\int_0^{2\pi} R(\theta)e^{-ki\theta}\mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi} c_k u_k t^k \cos(k\theta)e^{-ki\theta}\mathrm{d}\theta = c_k u_k \frac{t^k}{2\pi}\pi = \left(\frac{t}{2}\right)^k c_k \,,$$

where we used that $R(\theta) = \sum_{j=0}^{k} c_j(t\cos\theta + \phi)^j$. Since $c_k = 2^k O(d/k)^k$, we have that $b_k = O(td/k)^k$; this is at most $1/(4k)$, if t is a small enough multiple of $\log\, d/d$. ■

Appendix D. Lower Bound for Real-Valued Functions

In this section, we extend our lower bounds to the case of real-valued functions. We first show that a lower bound on the degree of any polynomial that approximates the functions of the class up to L^2-error ϵ translates to a lower bound on the complexity of any CSQ learner. While the L^2-norm is the most natural norm to use for approximating real-valued functions, we show that using the L^1-norm instead yields lower bounds in the general SQ model.

D.1. CSQ Lower Bound

We start with the result involving the L^2-norm. We restate Theorem 6 below.

Theorem 45 *Let $n, m \in \mathbb{Z}_+$ with $m \leq n^a$ for any constant $0 < a < 1/2$ and $\epsilon \geq n^{-c}$ for some sufficiently small constant $c > 0$. Let $f : \mathbb{R}^m \to \mathbb{R}$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f^2(\mathbf{x})] = 1$ and d be the smallest integer such that there exists a degree at most d polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\|f - p\|_2 < \epsilon$. Let $\mathcal{C}$ be a class of real-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any matrix $\mathbf{P} \in \mathbb{R}^{m\times n}$ satisfying $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Then, any CSQ algorithm that agnostically learns $\mathcal{C}$ over $\mathcal{N}_n$ to L^2-error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

To prove our CSQ lower bound, we need to find a hard function $g : \mathbb{R}^m \to \mathbb{R}$ that is uncorrelated with low-degree polynomials and, at the same time, is close to f in the L^2-sense. Instead of using duality to establish the existence of such a function g, we let g be the orthogonal component of the truncated Hermite expansion of f.

Proof [Proof of Theorem 6] Let an algorithm $\mathcal{A}$ that agnostically learns $\mathcal{C}$ up to L^2-error ϵ. Let $g(\mathbf{x}) = f(\mathbf{x}) - \sum_{i=0}^{d-1} f^{[i]}(\mathbf{x})$, i.e., g is the same as the function f without the low-degree moments up to $d-1$. Note that $\|g\|_2 \geq \epsilon$. Let $C = 2/(\epsilon\,\|g\|_2)$ and let S be the set of nearly orthogonal matrices of Lemma 17. Consider the class $\mathcal{C}_g$ that consists of all functions from $\mathbb{R}^n$ to $\mathbb{R}$ of the form $G_{\mathbf{V}}(\mathbf{x}) = Cg(\mathbf{V}\mathbf{x})$, for any matrix $\mathbf{V} \in S$. Every $G_{\mathbf{V}} \in \mathcal{C}_g$ is orthogonal to all polynomials of degree less than d, and also $\|G_{\mathbf{V}}\|_2 = 2/\epsilon$. We feed $\mathcal{A}$ with samples $(\mathbf{x}, G_{\mathbf{V}}(\mathbf{x}))$, where $\mathbf{x} \sim \mathcal{N}_n$, $\mathbf{V} \in S$. Let $\epsilon' > 0$ be the accuracy parameter used with $\mathcal{A}$. Then, $\mathcal{A}$ returns a hypothesis h satisfying

$$\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[(h(\mathbf{x}) - G_{\mathbf{V}}(\mathbf{x}))^2]} \leq \mathrm{OPT} + \epsilon' \,. \tag{4}$$

For our choice of C, the optimal error becomes

$$\begin{aligned}\mathrm{OPT} &\leq \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[(f(\mathbf{V}\mathbf{x}) - G_{\mathbf{V}}(\mathbf{x}))^2]} = \sqrt{1 + C^2\|g\|_2^2 - 2C\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})]} \\ &\leq \sqrt{1 + \frac{4}{\epsilon^2} - \frac{2\,\|g\|_2}{\epsilon}} \leq \sqrt{\frac{4}{\epsilon^2} - 1} \leq \frac{2}{\epsilon}\sqrt{1 - \frac{\epsilon^2}{4}} \leq \frac{2}{\epsilon} - \frac{\epsilon}{4}\,,\end{aligned}$$

where in the second inequality we used that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})] = \|g\|_2^2$. By choosing $\epsilon' = \epsilon/8$, Equation (4) becomes $\|h - G_{\mathbf{V}}\|_2 \leq 2/\epsilon - \epsilon/8$.

It remains to bound from above the pairwise correlation of the class $\mathcal{C}_g$. For any two different $\mathbf{U}, \mathbf{V} \in S$, we have that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_n}[G_{\mathbf{U}}(\mathbf{x})G_{\mathbf{V}}(\mathbf{x})] &\leq C^2\sum_{t=0}^{\infty}\|\mathbf{U}\mathbf{V}^\intercal\|_2^t \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[(g^{[t]}(\mathbf{x}))^2] \leq C^2\,\|\mathbf{U}\mathbf{V}^\intercal\|_2^d\sum_{t=d}^{\infty}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[(g^{[t]}(\mathbf{x}))^2] \\ &\leq 4\epsilon^{-2}\,\|\mathbf{U}\mathbf{V}^\intercal\|_F^d \leq \epsilon^{-2}n^{-\Omega(d)} \leq n^{-\Omega(d)}\,,\end{aligned}$$

where in the first inequality we used Lemma 15, in the second inequality we used the fact that g is uncorrelated with all polynomials of degree less than d, the third inequality follows from Parseval's identity and the fact that $\|g\|_2 C = 2/\epsilon$, the next one follows from Lemma 17, and the last one from our assumption $\epsilon > n^{-c}$ for an appropriate constant c. As a note, we extend our class $\mathcal{C}_g$ to include the identically zero function, which does not increase the pairwise correlations. Using Lemma 30 with $\gamma' = \gamma$, we have that $\mathrm{CSDA}_{\mathcal{N}_n}(\mathcal{C}_g, 2\gamma) = 2^{n^{\Omega(1)}}$ for $\gamma = n^{-\Omega(d)}$. An application of Lemma 31 with $\eta = 2/\epsilon$ concludes the proof. ■

D.2. SQ Lower Bound

In this section we prove lower bounds for the general SQ model. On this end, we require our hard function g to be pointwise bounded. This allows us to define a learning problem with Boolean labels, for which we have SQ lower bounds ready to be used. Because of our L^∞ constraint on g, the resulting lower bound is expressed in terms of the degrees of polynomials that approximate f in L^1 rather than L^2 sense. We restate Theorem 7 below.

Theorem 46 *Let $n, m \in \mathbb{Z}_+$ with $m \le n^a$ for any constant $0 < a < 1/2$ and $\epsilon \ge n^{-c}$ for some sufficiently small constant $c > 0$. Let $f : \mathbb{R}^m \to \mathbb{R}$ with $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[f^2(\mathbf{x})] = 1$ and d be the smallest integer such that there exists a degree at most d polynomial $p : \mathbb{R}^m \to \mathbb{R}$ satisfying $\|f - p\|_1 < \epsilon$. Let $\mathcal{C}$ be a class of real-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(\mathbf{x}) = f(\mathbf{P}\mathbf{x})$, for any matrix $\mathbf{P} \in \mathbb{R}^{m\times n}$ satisfying $\mathbf{P}\mathbf{P}^\intercal = \mathbf{I}_m$. Then, any SQ algorithm that agnostically learns $\mathcal{C}$ over $\mathcal{N}_n$ to L^2-error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

Our duality argument will now use the pair of dual norms L^1, L^∞.

Proposition 47 *Let $f \in L^2(\mathbb{R}^m)$ be such that for any degree at most $d-1$ polynomial $p : \mathbb{R}^m \to \mathbb{R}$, it holds $\|f - p\|_1 \ge \epsilon$. Then, there exists a function $g : \mathbb{R}^m \to [-1, 1]$ such that:*

1. $\mathbf{E}_{x\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})] \ge \epsilon$*, and,*
2. $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0$*, for any polynomial $P : \mathbb{R}^m \to \mathbb{R}$ with degree less than d.*

Proof The function g is a solution to the infinite system:

$$(*)\begin{cases} \mathop{\mathbf{E}}_{x\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})] \ge \epsilon & \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 & \forall P \in \mathcal{P}^m_{d-1} \\ \|g\|_\infty \le 1 & \end{cases}$$

This is equivalent to the following LP:

$$(**)\begin{cases} -\mathop{\mathbf{E}}_{x\sim\mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})] + \epsilon \le 0 & \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 & \forall P \in \mathcal{P}^m_{d-1} \\ \mathop{\mathbf{E}}_{x\sim\mathcal{N}_m}[g(\mathbf{x})h(\mathbf{x})] \le \|h\|_1 & \forall h \in L^1(\mathbb{R}^m) \end{cases}$$

From Corollary 55, the above LP is feasible unless the following is infeasible:

$$(**') \begin{cases} \|h\|_1 - \lambda\epsilon < 0 \\ h(\mathbf{x}) + P(\mathbf{x}) - \lambda f(\mathbf{x}) = 0, & \forall \mathbf{x} \in \mathbb{R}^m \\ \lambda \geq 0, h \in L^1(\mathbb{R}^m), P \in \mathcal{P}_{d-1}^m \end{cases}$$

Let (h, P, λ) be a solution to $(**')$. Note that we can assume that $\lambda = 1$ since all constraints are homogeneous. Then, the constraints become $h = f - P$ and

$$\|f - P\|_1 < \epsilon \; ,$$

which is a contradiction. Therefore, the original system $(*)$ is feasible. ■

We conclude with the proof of the main theorem for this section.

Proof [Proof of Theorem 7] Suppose that we have such an agnostic learner $\mathcal{A}$. Let $g : \mathbb{R}^m \to [-1, 1]$ be the function of Proposition 47, for a parameter $\epsilon' > 0$ to be specified. Let $\mathcal{D}_g$ be the family of distributions over $\mathbb{R}^n \times \{\pm 1\}$ from Definition 18. We use $\mathcal{A}$ to solve the problem of distinguishing between a distribution from $\mathcal{D}_g$ and the distribution where the labels are drawn uniformly at random. That is, we convert $\mathcal{A}$ into an algorithm for $\mathcal{B}(\mathcal{D}_g, \mathcal{N}_n \times \mathcal{U}(\{\pm 1\}))$, and the hardness result will follow from the hardness of that decision problem, as established by Proposition 19.

Let D' be a distribution that is either $D' = \mathcal{N}_n \times \mathcal{U}(\{\pm 1\}))$ or $D' \in \mathcal{D}_g$. We feed $\mathcal{A}$ a set of i.i.d. samples of the form $(\mathbf{x}, Cy)$, where $(\mathbf{x}, y) \sim D'$ and $C = 1/\mathbf{E}_{\mathbf{x} \sim \mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})]$. Let $\epsilon' > 0$ be the accuracy parameter used when running $\mathcal{A}$ and h be the returned hypothesis. We have that

$$\sqrt{\mathop{\mathbf{E}}_{(\mathbf{x},y) \sim D'}[(h(\mathbf{x}) - Cy)^2]} \leq \mathrm{OPT} + \epsilon' \; . \tag{5}$$

If $D' \in \mathcal{D}_g$, for the optimal error we have that

$$\mathrm{OPT} \leq \sqrt{1 + C^2 - 2C \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})]} \leq \sqrt{C^2 - 1} = C\sqrt{1 - 1/C^2} \leq C - 1/(2C) \; .$$

If we choose $\epsilon' = 1/(4C)$, Equation (5) becomes $\sqrt{\mathbf{E}_{(\mathbf{x},y) \sim D'}[(h(\mathbf{x}) - Cy)^2]} \leq C - 1/(4C)$. On the other hand, we can write $\sqrt{\mathbf{E}_{(\mathbf{x},y) \sim D'}[(h(\mathbf{x}) - Cy)^2]} \geq \sqrt{C^2 - 2C\,\mathbf{E}_{\mathbf{x} \sim \mathcal{N}_m}[h(\mathbf{x})y]}$. Combining these two, we obtain

$$2C \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}_m}[h(\mathbf{x})y] \geq C^2 - (C - 1/(4C))^2 \geq 1/3 \; ,$$

which gives that $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}_m}[h(\mathbf{x})y] \geq 1/(6C) = \mathbf{E}_{\mathbf{x} \sim \mathcal{N}_m}[f(\mathbf{x})g(\mathbf{x})]/6 \geq \epsilon/6$ from Proposition 47.

Note that if $D' = \mathcal{N}_n \times \mathcal{U}(\{\pm 1\})$, then $\mathbf{E}_{(\mathbf{x},y) \sim D'}[h(\mathbf{x})y] = 0$. Therefore, by performing a query of tolerance $\Omega(\epsilon)$ for the correlation of h with the labels, we can distinguish between the two cases of our hypothesis testing problem. By Proposition 19, this requires either $2^{n^{\Omega(1)}}$ queries or queries of tolerance $n^{-\Omega(d)}$. ■

Appendix E. Applications for Classes of Real-Valued Functions

Our applications for real-valued classes include ReLUs and sigmoids. The lower bounds established in this section are based on the degrees of the L^1 and L^2-approximating polynomials, as summarized in Table 2.

E.1. ReLU Activation

The class of Rectified Linear Unit (ReLU) functions consists of all functions of the form $\mathrm{ReLU}(\langle \mathbf{w}, \mathbf{x} \rangle)$, where $\mathbf{w} \in \mathbb{R}^n$ is any vector with $\|\mathbf{w}\|_2 = 1$ and $\mathrm{ReLU} : \mathbb{R} \to \mathbb{R}$ is defined as $\mathrm{ReLU}(t) = \max\{0, t\}$.

Upper and lower bounds for agnostically learning ReLUs were given in Goel et al. (2020b); Diakonikolas et al. (2020c). Diakonikolas et al. (2020c) established an SQ lower bound of $n^{\Omega(1/\epsilon^c)}$, for some constant $c > 0$. This constant c was not explicitly calculated in Diakonikolas et al. (2020c), but can be shown to be approximately $1/40$. Goel et al. (2020b) gave an SQ lower bound of $n^{\Omega(1/\epsilon^{1/36})}$ for this problem. We note that Goel et al. (2020b) considered a correlational type of guarantee, i.e., finding a hypothesis whose correlation with the labels is within ϵ of the optimal, as opposed to L^2-error. For this correlational guarantee, the upper bound of Goel et al. (2020b) is an L^2-regression algorithm with complexity $n^{O(\epsilon^{-4/3})}$, and the lower bound states that any SQ algorithm needs to perform queries with tolerance $\tau < n^{-\Omega(\epsilon^{-1/12})}$ or at least $2^{n^{\Omega(1)}}\epsilon$ queries. Furthermore, Goel et al. (2020b) showed that any agnostic learner with the square loss guarantee can be run with increased accuracy to satisfy the correlational guarantee. This reduction costs a "third root" in the exponent, yielding an $n^{\Omega(\epsilon^{-1/36})}$ SQ lower bound for the square loss guarantee. As a note, Goel et al. (2020b) assumes bounded labels. In this setting, agnostically learning within L^2-error $\mathrm{OPT}+\epsilon$ is equivalent to agnostically learning in squared L^2-error $\mathrm{OPT}+\epsilon'$, for $\epsilon' = \Theta(\epsilon)$.

To apply our theorems, we bound from below the degree of any polynomial that ϵ-approximates the univariate ReLU function. This can be done directly by appealing to Fact 38.

Corollary 48 *Let $f : \mathbb{R} \to \mathbb{R}$ be the ReLU function $\mathrm{ReLU}(t) = \max\{0, t\}$ and $p \in [1, 2]$. The minimum integer d for which there exists a degree-d polynomial $P : \mathbb{R} \to \mathbb{R}$ such that $\|\mathrm{ReLU} - P\|_p \le \epsilon$ is $d = \Theta\left(\epsilon^{-\frac{2}{1+1/p}}\right)$.*

Therefore, Theorems 6 and 7 imply a complexity of at least $n^{\Omega(\epsilon^{-4/3})}$ for any agnostic CSQ learner; and $n^{\Omega(\epsilon^{-1})}$ for any agnostic SQ learner respectively.

E.2. Sigmoid Activation

E.2.1. CSQ Lower Bound

We now let f be the standard sigmoid function, defined as $f(t) = 1/(1 + e^{-t})$, $t \in \mathbb{R}$. We first focus on bounding the degree of polynomials that approximate f in L^2-norm. This can be done via Hermite analysis, in particular, based on the fact that the polynomial of degree d being closest to f in L^2-norm is the truncated Hermite expansion $p_d(t) = \sum_{i=0}^{d} \hat{f}(i) H_i(t)$. The error of this approximation is $\|p_d - f\|_2^2 = \sum_{i=d+1}^{\infty} \hat{f}^2(i)$. For the asymptotic behavior of the Hermite coefficients, we use the following fact (see Goel et al. (2020a) and the references therein).

Fact 49 (Lemma A.9 from Goel et al. (2020a)) *Let $f : \mathbb{R} \to \mathbb{R}$ be the standard sigmoid function $f(t) = 1/(1 + e^{-t})$ and $\hat{f}(i)$ be its Hermite coefficients for $i \in \mathbb{Z}_+$. Then, $\hat{f}(0) = 0.5$, $\hat{f}(2i) = 0$ and $\hat{f}(2i - 1) = e^{-\Theta(\sqrt{i})}$, for $i \geq 1$.*

From this fact, we get the bound on the L^2-error of the best polynomials of degree d.

Corollary 50 (L^2-Degree Lower Bound for Sigmoid) *Let $f : \mathbb{R} \to \mathbb{R}$ be the standard sigmoid function $f(t) = 1/(1 + e^{-t})$ and d be the smallest integer for which there exists a degree-d polynomial $p : \mathbb{R} \to \mathbb{R}$ such that $\|f - p\|_2 < \epsilon$. Then $d = \Theta(\log^2(1/\epsilon))$.*

Proof Fix a degree k. From Fact 49, the best k-degree polynomial p_k achieves error

$$\|f - p_k\|_2^2 = \sum_{i=k+1}^{\infty} \hat{f}^2(i) = \sum_{i>k, i \text{ odd}} e^{-\Theta(\sqrt{i})} = \sqrt{k} e^{-\Theta(\sqrt{k})} .$$

This becomes ϵ^2 when k becomes $\Theta(\log^2(1/\epsilon))$. ■

By Theorem 6, we get that any CSQ agnostic learner for sigmoids has complexity $n^{\Omega(\log^2(1/\epsilon))}$.

E.2.2. SQ LOWER BOUND

The approach to derive lower bounds for the degrees of L^1-approximating polynomials will be to relate the L^1-norm to the L^2-norm and use the lower bounds for the latter. In particular, we will use the following fact about polynomials under the Gaussian measure.

Theorem 51 (Hypercontractivity Bogachev (1998); Nelson (1973)) *If p is a d-degree polynomial and $t > 2$, then*

$$\|p\|_t \leq (t - 1)^{d/2} \|p\|_2 .$$

Claim 52 *Let $r \in L^4(\mathbb{R})$. Then, $\|r\|_2 \leq \|r\|_1^{1/3} \|r\|_4^{2/3}$.*

Proof The proof follows from two applications of the Cauchy-Schwartz inequality.

$$\mathop{\mathbf{E}}_{t\sim\mathcal{N}}[r^2(t)] \leq \mathop{\mathbf{E}}_{t\sim\mathcal{N}}[|r(t)|]^{1/2} \mathop{\mathbf{E}}_{t\sim\mathcal{N}}\left[|r(t)|^3\right]^{1/2} \leq \mathop{\mathbf{E}}_{t\sim\mathcal{N}}[|r(t)|]^{1/2} \mathop{\mathbf{E}}_{t\sim\mathcal{N}}\left[|r(t)|^2\right]^{1/4} \mathop{\mathbf{E}}_{t\sim\mathcal{N}}\left[|r(t)|^4\right]^{1/4} .$$

Rearranging the above, yields the claimed inequality. ■

We can now show our L^1 polynomial degree lower bound.

Theorem 53 (L^1-Degree Lower Bound for Sigmoid) *Let $f : \mathbb{R} \to \mathbb{R}$ be the standard sigmoid function $f(t) = 1/(1 + e^{-t})$ and $0 < \epsilon < 1$. Any degree-d polynomial $p : \mathbb{R} \to \mathbb{R}$ that satisfies $\|f - p\|_1 < \epsilon$ must have $d = \Omega(\log(1/\epsilon))$.*

Proof Let $p : \mathbb{R} \to \mathbb{R}$ be a degree-d polynomial such that $\|f - p\|_1 < \epsilon$. Using Theorem 51 with $t = 4$ and then Claim 52 with $r(t) = p(t)$, we get that

$$\|p\|_4 \leq 3^{d/2} \|p\|_2 \leq 3^{d/2} \|p\|_1^{1/3} \|p\|_4^{2/3} .$$

After dividing both sides by $\|p\|_4^{2/3}$, we have that $\|p\|_4 \leq 3^{3d/2}\|p\|_1$. Furthermore, using the triangle inequality, $\|p\|_1 \leq \epsilon + \|f\|_1 = O(1)$. Therefore, $\|p\|_4 \leq 2^{O(d)}$. Furthermore, Claim 52 for $r(t) = f(t) - p(t)$ gives

$$\|f - p\|_2 \leq \|f - p\|_1^{1/3}\|f - p\|_4^{2/3} \leq \epsilon^{1/3} 2^{O(d)} .$$

On the other hand, for the L^2-error we have that $\|f - p\|_2 \geq \sqrt{d}e^{-\Theta(\sqrt{d})}$ (Corollary 50). Combining the two bounds, it follows that $d = \Omega(\log(1/\epsilon))$. ■

We note that Goel et al. (2020b) showed an $n^{\Omega(\log^2(1/\epsilon))}$ SQ lower bound for the correlational guarantee.

Appendix F. Duality in Infinite-Dimensional LP

We start with some basic definitions.

L^p space Let $(X, \mathcal{A}, \mu)$ be a measure space and $1 \leq p < \infty$. We will typically take $X = \mathbb{R}^n$, $n \in \mathbb{Z}_+$, and μ be the Gaussian measure, unless otherwise specified. For a function $f : X \to \mathbb{R}$, the L^p-norm of f under $\mathcal{N}_n$ is defined as $\|f\|_p := \left(\int_X |f|^p \mathrm{d}\,\mu\right)^{1/p}$. For the special case where $p = \infty$, the L^∞-norm of f is defined as the essential supremum of f on X, i.e., $\|f\|_\infty := \inf\{a \in \mathbb{R} : \mu\{\mathbf{x} \in X : f(\mathbf{x}) > a\} = 0\}$. The vector space $L^p(X, \mu)$ consists of all functions $f : X \to \mathbb{R}$ with $\|f\|_p < \infty$. We will typically use the shortened notation $L^p(\mathbb{R}^n)$ for $L^p(\mathbb{R}^n, \mathcal{N}_n)$.

Dual Norms Consider a vector space V with inner product $\langle\cdot,\cdot\rangle$ and a norm $\|\cdot\|$ on V. The *dual norm* $\|f\|_*$, $f \in V$, is defined as $\|f\|_* = \sup\{\langle f, h\rangle : \|h\| \leq 1\}$. Hölder's inequality states that for any $f, h \in V$ it holds $\langle f, h\rangle \leq \|f\|\,\|h\|_*$.

ments, we need to prove that there exists a function $g : \mathbb{R}^m \to \mathbb{R}$, such that for any function $h \in L^p(\mathbb{R}^m)$ and at most $(d-1)$-degree polynomial $P : \mathbb{R}^m \to \mathbb{R}$, it holds

$$(*)\begin{cases} -\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})f(\mathbf{x})] + c \leq 0 & 0 < c < \|f\| \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[P(\mathbf{x})g(\mathbf{x})] = 0 & \forall P \in \mathcal{P}_{d-1}^m \\ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})h(\mathbf{x})] - \|h\|_p \leq 0 & \forall h \in L^p(\mathbb{R}^m) \end{cases}$$

This is in fact an infinite dimensional linear system with respect to the unknown function $g \in (L^1(\mathbb{R}^m))^* = L^\infty(\mathbb{R}^m)$, for $p = 1$ and $g \in L^{p/(p-1)}(\mathbb{R}^m)$ for $1 \leq p < \infty$. We are going to denote $\mathcal{X}$ the metric space $L^p(\mathbb{R}^m)$.

Basics on Duality of Infinite-Dimensional LPs In our arguFor succinctness, we will use the following notation. We use $(\tilde{h}, t)$ for the inequality $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}_m}[g(\mathbf{x})\tilde{h}(\mathbf{x})] + t \leq 0$, where $\tilde{h} \in \mathcal{X}$ and $t \in \mathbb{R}$. Moreover, let $\mathcal{S}$ be the set that contains all such tuples that describe the target system. For the set $\mathcal{S}$, the closed convex cone over $\mathcal{X} \times R$ is the smallest closed set $\mathcal{S}_+$ satisfying, if $A \in \mathcal{S}_+$ and $B \in \mathcal{S}_+$ then $A + B \in \mathcal{S}_+$ and, if $A \in \mathcal{S}_+$ then $\lambda A \in \mathcal{S}_+$ for all $\lambda \geq 0$. Note that the $\mathcal{S}_+$ contains the same feasible solutions as $\mathcal{S}$. The set $\mathcal{S} = \{(h, -\|h\|_p) : h \in L^p\} \cup \{(P, 0) : P \in \mathcal{P}_{d-1}^m\} \cup \{(-f, c)\}$.

In the finite-dimensional case, we can always prove the feasibility of an LP by applying the standard Farkas' lemma (aka theorem of the alternative). However, when the system is infinite-dimensional, Farkas' lemma does not hold in general. We are going to use the following result from Fan (1968).

Lemma 54 (Theorem 1 of Fan (1968)) *If $\mathcal{X}$ is a locally convex, real separated vector space then, a linear system described by $\mathcal{S}$ for which $\mathcal{S}_+$ is feasible (i.e., there exists a $g \in \mathcal{X}^*$) if and only if* $(0, 1) \notin \mathcal{S}_+$.

One direction is trivial, but the other one needs an application of Hahn-Banach theorem which is where the assumption on $\mathcal{X}$ to be a separated space is used.

Corollary 55 *If $\mathcal{X} = L^p$ for $1 \leq p < \infty$ then, the LP described by $\mathcal{S}$ is feasible if only if* $(0, 1) \notin \mathcal{S}_+$.

Proof It is not hard to see that the positive cone is defined by

$$\mathcal{S}_+ = \{(P + h - yf, -\|h\|_p + yc - t) : P \in \mathcal{P}^m_{d-1}, h \in L^p(\mathbb{R}^m), y, t \in \mathbb{R}, y, t \geq 0\}.$$

Now if $\mathcal{S}_+$ were closed, we could simply apply Lemma 54. However, it is not immediately clear if this is the case. Instead, we note that Lemma 54 can be applied to the closure of $\mathcal{S}_+$. In particular, this means that the LP is solvable unless for any $\epsilon > 0$ we have P, h, y and t so that $\|P + h - yf\|_p, | - \|h\|_p + yc - t - 1| < \epsilon$. Letting $h' = yf - P = h - (P + h - yf)$, we find that $\mathcal{S}_+$ contains

$$(P + h' - yf, -\|h'\|_p + yc - t) = (0, -\|h'\|_p + yc - t).$$

We note that $\|h'\|_p \leq \|h\|_p + \|P + h - yf\|_p = \|h\|_p + \epsilon$. This means that $-\|h'\|_p + yc - t \geq -\|h\|_m + yc - y - \epsilon \geq 1 - 2\epsilon \geq 1/2$ if $\epsilon < 1/4$. Noting that $\mathcal{S}_+$ is scale invariant, this implies that $(0, 1) \in \mathcal{S}_+$.

Thus, the LP is solvable unless $(0, 1) \in \mathcal{S}_+$.

■