Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Clayton Sanford CLAYTON@CS.COLUMBIA.EDU
Rocco A. Servedio ROCCO@CS.COLUMBIA.EDU
Emmanouil-Vasileios Vlatakis-Gkaragkounis EMVLATAKIS@CS.COLUMBIA.EDU
Columbia University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider the well-studied problem of learning intersections of halfspaces under the Gaussian distribution in the challenging *agnostic learning* model. Recent work of Diakonikolas et al. (2021b) shows that any Statistical Query (SQ) algorithm for agnostically learning the class of intersections of k halfspaces over $\mathbb{R}^n$ to constant excess error either must make queries of tolerance at most $n^{-\tilde{\Omega}(\sqrt{\log k})}$ or must make $2^{n^{\Omega(1)}}$ queries. We strengthen this result by improving the tolerance requirement to $n^{-\tilde{\Omega}(\log k)}$. This lower bound is essentially best possible since an SQ algorithm of Klivans et al. (2008) agnostically learns this class to any constant excess error using $n^{O(\log k)}$ queries of tolerance $n^{-O(\log k)}$. We prove two variants of our lower bound, each of which combines ingredients from Diakonikolas et al. (2021b) with (an extension of) a different earlier approach for agnostic SQ lower bounds for the Boolean setting due to Dachman-Soled et al. (2014). Our approach also yields lower bounds for agnostically SQ learning the class of "convex subspace juntas" (studied by Vempala, 2010a) and the class of sets with bounded Gaussian surface area; all of these lower bounds are nearly optimal since they essentially match known upper bounds from Klivans et al. (2008).

Keywords: Statistical Query learning, agnostic learning, intersections of halfspaces

1. Introduction

Linear threshold functions, or *halfspaces*, are ubiquitous in machine learning. They arise in the context of many statistical models for classification (Duda et al., 1973), and they are the focus of many well-known machine learning methods, including Perceptron (Rosenblatt, 1962), Support Vector Machines (Vapnik, 1982), and AdaBoost (Freund and Schapire, 1997). In this work, we consider the problem of agnostic learning for a natural and well-studied generalization of this function class: *intersections of halfspaces*.

Although many efficient algorithms for learning halfspaces have been developed to handle a wide variety of settings (Blumer et al., 1989; Blum et al., 1998b; Kalai et al., 2008; Awasthi et al., 2017; Diakonikolas et al., 2021a), known algorithms for intersections of halfspaces are conspicuously limited in scope and applicability. Indeed, no efficient PAC learning algorithms are known even for the case of intersections of two halfspaces. There, a learner faces a "credit assignment" problem when considering negative examples, as either of the two halfspaces may be responsible for an example being classified as negative, but the learner is not privy to this information. This prevents a straightforward formulation of the learning problem as a linear program, which had sufficed in the case of learning single halfspaces.

Because of the apparent difficulty of going beyond single halfspaces, much of the progress has come from learning under "nice" data marginal distributions, such as the uniform distribution or the

Gaussian distribution (Blum and Kannan, 1997; Vempala, 1997, 2010b; Klivans et al., 2004; Kalai et al., 2008; Klivans et al., 2008; Vempala, 2010a; Kane, 2014). The fastest algorithm to date for agnostically learning intersections of halfspaces under Gaussian marginals in $\mathbb{R}^n$ is L^1 polynomial regression (Kalai et al., 2008), which was shown by Klivans et al. (2008) to successfully learn up to any constant excess error in time $n^{O(\log k)}$. (Under the additional assumption of realizability, Vempala (2010a) showed that when $k = o(n)$, preprocessing with principal component analysis improves this running time to $\mathrm{poly}(n, k) + k^{O(\log k)}$.) Since this upper bound has resisted improvement for several years, attention has turned to trying to prove lower bounds, and such lower bounds are the subject of this paper.

The Statistical Query (SQ) model of Kearns (1998) offers an attractive setting for proving unconditional lower bounds against a broad class of learning algorithms. SQ learning algorithms can access data only through imperfect estimates of the expected values of query functions with respect to the data distribution. Nearly all known learning algorithms, including those of Kalai et al. (2008), Klivans et al. (2008) and Vempala (2010a), can be implemented within the SQ model, so lower bounds in the SQ model are evidence for the computational difficulty of a learning problem. Because these algorithmic results for agnostic learning hold only under "nice" marginal distributions, it is of interest to prove *distribution-dependent* SQ lower bounds under the same marginals.

The pioneering work of Dachman-Soled et al. (2014) provided a blueprint for proving such distribution-dependent SQ lower bounds. They proved an equivalence between the approximation resilience of functions in a concept class and the SQ agnostic learnability of that class, and used this equivalence to obtain the first super-polynomial SQ lower bounds for agnostically learning the important concept class of monotone juntas under the uniform distribution. To establish SQ lower bounds for agnostic learning under Gaussian marginals, Diakonikolas et al. (2021b) extended the approach of Dachman-Soled et al. (2014) using new duality arguments and embedding techniques. In doing so, they obtained lower bounds for agnostically learning a number of Boolean concept classes (as well as some real-valued concept classes). For intersections of k halfspaces, their agnostic SQ lower bound is $n^{\tilde{\Omega}(\sqrt{\log k})}$, which should be contrasted with the $n^{O(\log k)}$ upper bound of Klivans et al. (2008). In fact, they conjectured that it may be the upper bound that is loose.

Our results prove that the algorithmic results of Klivans et al. (2008) are indeed nearly optimal. Specifically, we show that any SQ algorithm that agnostically learns intersections of $k \leq \exp(O(n^{0.245}))$ halfspaces to any constant excess error must have complexity at least $n^{\tilde{\Omega}(\log k)}$. The notion of complexity is made more precise in the informal theorem statement below.

Theorem 1 (Informal version of Theorem 18) *Any SQ algorithm that agnostically learns intersections of k halfspaces to excess error ϵ under Gaussian marginals requires either $2^{n^{\Omega(1)}}$ queries or at least one query of tolerance $n^{-\tilde{\Omega}(\log k + 1/\epsilon^2)}$.*

This result is nearly optimal for any constant ϵ, up to a $\log \log k$ factor in the exponent, because the $n^{O(\log k)}$ time and sample complexity upper bounds from Klivans et al. (2008) can be achieved by an SQ algorithm. We note that by the AM-GM inequality the exponent $\tilde{\Omega}(\log k + 1/\epsilon^2)$ in our lower bound is always at least $\tilde{\Omega}(\sqrt{\log k}/\epsilon)$, which is the exponent from the SQ lower bound of Diakonikolas et al. (2021b), but can also be significantly stronger.

In fact, when k is small ($k = O(n^{0.49})$), we show that the hardness of learning intersections of $2k$ halfspaces is already present in the easier problem of learning a simple subset of the class: the family of k-dimensional cubes. This result, given in Theorem 7, relies on new technical facts about the L^1-error approximation degree of cube functions under Gaussian marginals. While Theorem 18

has the strength of applying to much larger choices of k, Theorem 7 provides an explicit class of rotated cubes that are difficult to learn in the agnostic SQ model, in contrast to the other existential result.

The hardness of learning rotated cubes is of further interest because there exists a $\mathrm{poly}(n)$-time algorithm for *realizably* learning k-dimensional cubes that is far more efficient than the best known $\mathrm{poly}(n,k) + nk^{O(\log(k))}$-time algorithm by Vempala (2010a) for agnostically learning general intersections of halfspaces. This algorithm uses the gradient descent approach of Nguyen and Regev (2006) and Frieze et al. (1996) to learn each orthogonal direction. Theorem 18 implies that no such algorithm for learning rotated cubes exists in the agnostic setting.

Our bounds additionally imply new hardness results on learning functions with bounded Gaussian surface area and convex subspace juntas (see Theorems 20 and 21 respectively).

1.1. Techniques

Our proofs follow the blueprints of Dachman-Soled et al. (2014) and Diakonikolas et al. (2021b) and build upon them by using weak learning lower bounds from De and Servedio (2021) and new technical innovations for proving resilience with respect to continuous measures. Put roughly, Dachman-Soled et al. (2014):

(a) introduced a notion of *approximate resilience* on the Boolean cube and established an equivalence to L^1 approximate degree using linear programming duality;
(b) used a combinatorial argument to show that if a k-dimensional function f is approximately resilient, then there exists a family of k-juntas (n-dimensional embeddings of f for $n \gg k$) that is hard to agnostically learn in the SQ model;
(c) used Boolean Fourier analysis to prove approximate resilience for the Tribes function (a monotone read-once DNF); and
(d) proved a tighter approximate resilience bound for other monotone Boolean functions by combining a hardness result on weak learning of Blum et al. (1998a) with an agnostic learning algorithm based on L^1 polynomial approximation by Kalai et al. (2008).

To transfer this methodology to the Gaussian measure on $\mathbb{R}^n$, Diakonikolas et al. (2021b):

(a′) extended the equivalence of approximate resilience and L^1 approximate degree to Gaussian marginals with a more technical argument involving an infinite linear program and the Hahn-Banach Theorem;
(b′) showed that L^1 polynomial inapproximability of a k-dimensional function implies the hardness of SQ-learning a family of n-dimensional embeddings of f applied to k-dimensional subspaces[1]; and
(c′) lower-bounded the L^1 approximate degree of an intersection of k halfspaces using a new connection with Gaussian noise sensitivity.

Our results are obtained using a hybrid of the Dachman-Soled et al. (2014) and Diakonikolas et al. (2021b) approaches. More precisely, we rely on (a′) and (b′) to establish agnostic SQ lower bounds over Gaussian marginals for approximately resilient functions, but we draw inspiration from (c) instead of (c′) to bound the approximate resilience of the Cube_k function by directly analyzing its Hermite representation. We also draw inspiration from (d) when we lower-bound the approximate

1. The underlying hard problem is distinguishing a standard (multivariate) Gaussian from a distribution that differs from the standard Gaussian only in the high-order moments of a k-dimensional projection (Diakonikolas et al., 2017).

resilience of other intersections of halfspaces by using a recent hardness result from De and Servedio (2021) for weak learning those functions.

In more detail, Theorem 7 proves the hardness of learning the restricted class of k-dimensional cubes for $k = O(n^{0.49})$ in n-dimensional space by directly bounding the approximate resilience of a single cube function, $\mathsf{Cube}_k : \mathbb{R}^k \to \mathbb{R}$. That is, we show that Cube_k is close in L^1-distance to a bounded function that is orthogonal to all polynomials of degree $d = \tilde{\Omega}(\log k)$. To construct this bounded function, we develop a new argument which is inspired by (c) but is significantly more technically involved. Due to the unboundedness and continuity of our $\mathcal{N}(0, I_n)$ setting, our argument requires a careful iterative construction, which involves defining a thresholding transform that reduces the low-degree Hermite coefficients of its input while maintaining its boundedness and taking the limit of applying the transform an infinite number of times. The key properties of Cube_k for this argument are the boundedness of its outputs and its small low-degree Hermite weight. The approximate resilience of Cube_k provides an almost-tight bound on the L^1 approximate degree of the function, and the main result follows by direct application of (a′) and (b′).

Theorem 14, which shows the hardness of learning to constant accuracy the broader classes of all intersections of k halfspaces for any $k = \exp(O(n^{0.245}))$, instead relies on the combination of recent lower bounds on the number of queries needed to weakly learn intersections of k halfspaces from De and Servedio (2021) and well-known algorithmic results of Kalai et al. (2008) for agnostically learning functions with bounded L^1 approximate degree. This approach draws inspiration from (d). We show that the L^1 approximate degree of a random intersection of k halfspaces must be at least $\tilde{\Omega}(\log k)$ with high probability, since otherwise there would be a contradiction between the aforementioned works: Kalai et al. (2008) would provide an algorithm to weakly learn intersections of halfspaces using fewer queries than the lower bound established by De and Servedio (2021). As before, these bounds on polynomial inapproximability translate to SQ learning lower bounds via the machinery of Diakonikolas et al. (2021b).

All of the above arguments are for constant excess error (constant ϵ). We introduce the dependence on $\frac{1}{\epsilon}$ in Theorem 18 by augmenting the previously-considered intersections of k halfspaces with a single halfspace (in an additional dimension) that passes through the origin. Ganzburg (2002) showed that a single halfspace has L^1 ϵ-approximate degree $\Omega(\frac{1}{\epsilon^2})$, and we use this to show that our new intersection of halfspaces has approximate degree $\tilde{\Omega}(\log k + \frac{1}{\epsilon^2})$.

1.2. Related work

Efficient algorithms are known for PAC learning intersections of halfspaces under certain marginal distributions. Baum (1990) gave an algorithm for learning two homogeneous halfspaces under origin-symmetric distributions, and the same algorithm is now known to also succeed under mean-zero log-concave distributions (Klivans et al., 2009). For PAC learning intersections (and other functions) of k general halfspaces, algorithms are known for the uniform distribution on the unit ball (Blum and Kannan, 1997), the uniform distribution on the Boolean cube (Klivans et al., 2004; Kalai et al., 2008; Kane, 2014), Gaussian distributions (Klivans et al., 2008; Vempala, 2010a), and general log-concave distributions (Vempala, 1997, 2010b). In most of these cases, the dependence on k in the running time is $n^{\Omega(k)}$ or worse (the exceptions are the algorithms for Gaussian or uniform on $\{-1, 1\}^n$ marginals). In fact, only the L^1 polynomial regression algorithm is known to succeed in the agnostic setting, and only under Gaussian or uniform on $\{-1, 1\}^n$ marginals (Klivans et al., 2008; Kane, 2014). Finally, efficient algorithms are also known for PAC learning intersections of any constant number of halfspaces under marginals satisfying a geometric

margin condition (Arriaga and Vempala, 2006; Klivans and Servedio, 2008), and also for learning intersections (and other functions) of halfspaces using membership queries (Kwek and Pitt, 1998; Gopalan et al., 2012).

Our work focuses on hardness of learning intersections of halfspaces. Besides the SQ lower bounds of Diakonikolas et al. (2021b) for (agnostic) learning under Gaussian marginals (which built on the closely related work of Dachman-Soled et al. (2014)), there is other evidence for the difficulty of this learning problem. First, distribution-free PAC learning—both proper learning and improper learning with certain hypothesis classes—is known to be NP-hard (Blum and Rivest, 1992; Megiddo, 1988), and lower bounds on the threshold degree of intersections of two halfspaces due to Sherstov (2013) rule out efficient algorithms that use polynomial threshold functions as hypotheses. Cryptographic lower bounds (Klivans and Sherstov, 2006) give further evidence that distribution-free PAC learning is hard even if the learner is permitted to output any polynomial-time computable hypothesis. (The distribution-free correlational SQ lower bounds of Gollakota et al. (2020) give similar evidence for restricted types of learners.) These lower bounds leave open the possibility that fixed-distribution PAC learning is tractable, but again there is evidence against this, at least for certain classes of learning algorithms. Klivans and Sherstov (2007) showed that there is a (non-uniform) marginal distribution on the Boolean cube under which the SQ dimension of intersections of $\sqrt{n}$ halfspaces is at least $2^{\Omega(\sqrt{n})}$; this implies lower bounds for (weak) SQ learning under that distribution. Finally, Klivans et al. (2008) gave membership query lower bounds for learning certain convex bodies under Gaussian marginals. These lower bounds are exhibited by intersections of k halfspaces for sufficiently large k, but they do not rule out $\mathrm{poly}(n)$ query algorithms unless k is at least polynomially large in n. Moreover, these lower bounds are insensitive to the error parameter ϵ sought by the learner, and in particular do not become higher for subconstant ϵ.

1.3. Organization

In Section 3 we prove Theorem 7, which gives an $n^{\tilde{\Omega}(\log k)}$ SQ lower bound for agnostically learning intersections of k halfspaces (in fact, k-dimensional cubes) to constant excess error when $k = O(n^{0.49})$. Section 4 gives a similar SQ lower bound for larger values of k (using different arguments and less structured intersections of halfspaces which are not cubes). Section 5 improves the quantitative results of both these sections by allowing for subconstant excess error, thereby establishing Theorem 1 (see Theorem 18 in Section 5 for a detailed theorem statement). Appendix A extends our results to the concept class of functions with bounded Gaussian surface area and convex subspace juntas, and gives some observations on lower bounds for L^1 polynomial approximation.

2. Preliminaries

2.1. Functions in Gaussian space

For any $k \in \mathbb{N}$, the standard Gaussian distribution on $\mathbb{R}^k$ is denoted by $\mathcal{N}(0, I_k)$. For $q \geq 1$, let $\|f\|_q = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I_k)}[|f(\mathbf{x})|^q]^{1/q}$ denote the L^q-norm of $f \in L^q(\mathcal{N}(0, I_k))$, and let $\langle f, g\rangle = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I_n)}[f(\mathbf{x})g(\mathbf{x})]$ denote the inner product between $f, g \in L^2(\mathcal{N}(0, I_k))$. For a multi-index $J \in \mathbb{N}^k$, let $\#J$ denote the number of nonzero elements of J, and let $|J| = J_1 + \cdots + J_k$. Let $\mathcal{P}_{k,d}$ denote the family of all polynomials $p : \mathbb{R}^k \to \mathbb{R}$ of degree at most d.

In Appendix B, we recall basic facts about the *Hermite polynomials* $\{H_J\}_{J\in\mathbb{N}^k}$, which form an orthogonal basis for $L^2(\mathcal{N}(0, I_k))$, as well as some tools based on Gaussian hypercontractivity.

2.2. Agnostic learning under Gaussian marginals and Statistical Query learning

We recall the framework of agnostic learning under Gaussian marginals. Given a concept class $\mathcal{C}$ of functions from $\mathbb{R}^n$ to $\{-1,1\}$, an agnostic learning algorithm is given access to i.i.d. labeled examples $(\mathbf{x},\mathbf{y})$ drawn from a distribution $\mathcal{D}$ over $\mathbb{R}^n \times \{-1,1\}$, where the marginal of $\mathcal{D}$ over the first n coordinates is $\mathcal{N}(0,I_n)$. Intuitively, a successful agnostic learning algorithm for $\mathcal{C}$ is one which can find a hypothesis that correctly predicts the label $\mathbf{y}$ almost as well as the best predictor in $\mathcal{C}$. More precisely, an *agnostic learning algorithm for $\mathcal{C}$ under Gaussian marginals with excess error* ϵ is an algorithm which, with high probability, outputs a hypothesis function $h : \mathbb{R}^n \to \{-1,1\}$ such that $\Pr_{(\mathbf{x},\mathbf{y})\in\mathcal{D}}[h(\mathbf{x}) \neq \mathbf{y}] \leq \mathrm{OPT} + \epsilon$, where $\mathrm{OPT} = \inf_{f\in\mathcal{C}} \Pr_{(\mathbf{x},\mathbf{y})\in\mathcal{D}}[f(\mathbf{x}) \neq \mathbf{y}]$. The special case of $\mathrm{OPT} = 0$ corresponds to (realizable) PAC learning under the Gaussian distribution.

The above definition is for a learning scenario in which the learner has access to individual random examples. In the well-known Statistical Query (SQ) learning model, the learning algorithm cannot access individual examples from $\mathcal{D}$ but instead has access to a "STAT oracle."

Definition 2 *A learning algorithm $\mathcal{A}$ has access to a* STAT oracle *if $\mathcal{A}$ makes queries with a function $g : \mathbb{R}^n \times \{-1,1\} \to [-1,1]$ and a tolerance parameter $\tau > 0$ and recieves an estimate of the expectation $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[g(\mathbf{x},\mathbf{y})]$ that is accurate up to additive error $\pm\tau$. An algorithm $\mathcal{A}$ with access to a STAT oracle is an* SQ agnostic learning algorithm *for concept class $\mathcal{C}$ if it returns with high probability a hypothesis $h : \mathbb{R}^n \to \{-1,1\}$ such that $\Pr_{(\mathbf{x},\mathbf{y})\in\mathcal{D}}[h(\mathbf{x}) \neq \mathbf{y}] \leq \mathrm{OPT} + \epsilon$.*

2.3. Resilience and L^1 polynomial approximation

Dachman-Soled et al. (2014) established a useful connection between lower bounds for SQ agnostic learning under the uniform distribution on $\{-1,1\}^n$ and the notion of *resilience* for bounded functions. This connection was extended to the Gaussian setting by Diakonikolas et al. (2021b), and it also plays an essential role in our results.

Intuitively, a function is "resilient" if it has zero correlation with all low-degree basis functions. More formally, we have the following:

Definition 3 *A function $g : \mathbb{R}^n \to [-1,1]$ is* d-resilient *if $\langle g,p\rangle = 0$ for every $p \in \mathcal{P}_{n,d}$ (equivalently, $\langle g, H_J\rangle = 0$ for every $|J| \leq d$). For $0 \leq \alpha < 1$, a function $f : \mathbb{R}^n \to [-1,1]$ is said to be* α-approximately d-resilient *if there exists a d-resilient* witness *$g : \mathbb{R}^n \to [-1,1]$ such that $\|f-g\|_1 \leq \alpha$.*

Next we define the notion of L^1 polynomial approximation:

Definition 4 *Given $0 \leq \epsilon < 1$ and $f : \mathbb{R}^n \to [-1,1]$, we say that the L^1 ϵ-approximate degree of f is the smallest value $d \geq 0$ such that there exists a polynomial $p \in \mathcal{P}_{n,d}$ satisfying $\|f-p\|_1 \leq \epsilon$.*

Definition 4 is of course equivalent to d being the largest value such that every polynomial p of degree at most $d-1$ has $\|f-p\|_1 > \epsilon$.

Using linear programming duality, for the setting of functions $f : \{-1,1\}^n \to [-1,1]$ and the uniform distribution over $\{-1,1\}^n$, Dachman-Soled et al. (2014) established an *equivalence* between the L^1-distance to the closest d-resilient bounded function (cf. Definition 3) and the best possible accuracy of L^1 polynomial approximation by degree-d polynomials (cf. Definition 4). They did this by showing (see their Theorem 1.2) that for $f : \{-1,1\}^n \to [-1,1]$, if the former quantity is α then the latter quantity is $1-\alpha$.

This equivalence was extended to the setting of Gaussian space (our domain of interest in the current work) by Diakonikolas et al. (2021b); a more involved argument is required for this setting, essentially because now the linear programming duality involves an infinitely large linear program, but the result still goes through. The proof of their Proposition 2.1 establishes the following:[2]

Lemma 5 (Equivalence of approximate resilience and L^1 approximate degree) *A function $f : \mathbb{R}^n \to \{-1,1\}$ is α-approximately d-resilient if and only if its L^1 $(1-\alpha)$-approximate degree is d.*

In the Boolean hypercube setting, Dachman-Soled et al. (2014) combined their L^1 polynomial approximation characterization of resilience with standard SQ lower bounds and standard results on the existence of combinatorial designs to show the following: if $f : \{-1,1\}^k \to \{-1,1\}$ is an α-approximately d-resilient function, then (roughly speaking; see their Lemma 2.1 for a precise statement) any concept class of functions from $\{-1,1\}^n$ to $\{-1,1\}$ containing all "embeddings" of f (according to the combinatorial design) admits a Statistical Query lower bound.

Diakonikolas et al. (2021b) carried out a similar-in-spirit argument in the setting of Gaussian space. We note that their result is significantly technically more challenging than the analogous argument of Dachman-Soled et al. (2014); it builds on recent SQ lower bounds for distinguishing distributions due to Diakonikolas et al. (2017), and uses embeddings of low-dimensional functions in hidden low-dimensional subspaces rather than combinatorial designs. We record their key result below, which will be crucially used for all of our agnostic SQ lower bounds:

Lemma 6 (Diakonikolas et al., 2021b, Theorem 1.4) *Let $n, m \in \mathbb{N}$ with $m \leq n^a$ for any $0 < a < 1/2$, and let $\epsilon \geq n^{-c}$ for a suitably small absolute constant $c > 0$. Given any function $f : \mathbb{R}^m \to \{-1,1\}$, let d be the L^1 (2ϵ)-approximate degree of f.*[3] *Let $\mathcal{C}$ be a class of $\{-1,1\}$-valued functions on $\mathbb{R}^n$ which includes all functions of the form $F(x) = f(Px)$ for all $P \in \mathbb{R}^{m \times n}$ such that $PP^{\mathsf{T}} = I_m$. Any SQ algorithm that agnostically learns $\mathcal{C}$ under $\mathcal{N}(0, I_n)$ to error $\mathrm{OPT} + \epsilon$ either requires queries with tolerance at most $n^{-\Omega(d)}$ or makes at least $2^{n^{\Omega(1)}}$ queries.*

3. Hardness of SQ learning to constant excess error via approximate resilience

The main result of this section is Theorem 7, which, roughly speaking, shows that any SQ algorithm that makes a sub-exponential number of statistical queries and agnostically learns the concept class of "embedded k-dimensional cubes" (for $k = O(n^{0.49})$) to any constant excess error that is bounded below $\frac{1}{2}$ must make queries of tolerance $n^{-\Omega(\log(k)/\log\log k)}$. (Note that this gives a special case of Theorem 1 in which the excess error ϵ is constant and $k = O(n^{0.49})$.) This is done by establishing that the k-dimensional cube function is approximately resilient; recall that by Definition 3, this means that it is close in L^1 distance to a bounded function that is orthogonal to all low-degree polynomials.

We define the function $\mathsf{Cube}_k : \mathbb{R}^k \to \{-1,1\}$ as $\mathsf{Cube}_k(y) := \mathrm{sign}(\theta_k - \|y\|_\infty)$. (Note that this is equivalent to $\mathsf{Cube}_k(y) = 2\prod_{i=1}^k \mathbb{1}\{|y_i| \leq \theta_k\} - 1$.) In words, $\mathsf{Cube}_k^{-1}(1)$ is the axis-aligned origin-centered solid cube with side length $2\theta_k$, where $\theta_k \geq 0$ is chosen to ensure that $\mathbb{E}_{\mathbf{y} \sim \mathcal{N}(0,I_k)}[\mathsf{Cube}_k(\mathbf{y})] = 0$. Note that Cube_k is an intersection of $2k$ halfspaces.

2. The statement of Proposition 2.1 of Diakonikolas et al. (2021b) only goes in one direction (that L^1 polynomial approximate degree implies approximate resilience), but the proof establishes both directions.
3. By Lemma 5, this condition is equivalent to f being $(1-2\epsilon)$-approximately d-resilient.

Theorem 7 *For sufficiently large n and k, with $k = O(n^{0.49})$, define the concept class $\mathcal{C} = \{x \mapsto \mathsf{Cube}_k(Px) : P \in \mathbb{R}^{k\times n}, PP^{\mathsf{T}} = I_k\}$. Any SQ algorithm that agnostically learns $\mathcal{C}$ to excess error $\frac{1}{2}\left(1 - \frac{1}{k^{0.49}}\right)$ requires $2^{n^{\Omega(1)}}$ queries or at least one query of tolerance $n^{-\Omega(\log(k)/\log\log k)}$.*

This strengthens the bounds of Diakonikolas et al. (2021b) for the regime of $k = O(n^{0.49})$ and constant excess error, improving the $n^{-\tilde{\Omega}(\sqrt{\log k})}$ tolerance requirement to $n^{-\tilde{\Omega}(\log k)}$. Theorem 7 follows directly from Lemmas 5 and 6 and the following lemma:

Lemma 8 *For sufficiently large k, the function Cube_k is α-approximately d-resilient for $\alpha = k^{-0.49}$ and $d = \Omega(\log(k)/\log\log k)$.*

The proof of Lemma 8 has two main ingredients: a bound on the Hermite weight of Cube_k that is contained in its low-degree coefficients (Lemma 9), and an approximate resilience guarantee for functions with bounded low-degree Hermite weight (Lemma 10). We prove Lemma 8 in Section 3.2 by applying those two lemmas and choosing an appropriate setting for d in terms of k.

Lemma 9 *For any sufficiently large k, and any $d \geq 0$,*[4]

$$\sum_{|J|\leq d} \widetilde{\mathsf{Cube}_k}(J)^2 \leq \frac{20d(3\ln k)^d}{k}.$$

We prove this lemma in Appendix C.1 by exactly computing the Hermite coefficients of one-dimensional centered interval functions (Lemma 27) and using those values to carefully bound the Cube_k Hermite coefficients. At a high level, our bounds on the low-degree Hermite coefficients of Cube_k are similar in flavor to the bounds of Mansour (1992) on the low-degree Fourier coefficients of the read-once "tribes" CNF over the Boolean hypercube.

Lemma 10 *For sufficiently large k, $d \geq 2$, and $f : \mathbb{R}^k \to \{-1, 1\}$, let $\gamma := \sum_{|J|\leq d} \widetilde{f}(J)^2$. Then f is α-approximately d-resilient for $\alpha = \gamma^{0.498}(72\ln k)^{d/2}$.*

The proof of Lemma 10 is given in Section 3.1 and is somewhat technically involved. Our argument modifies and extends a proof idea from Dachman-Soled et al. (2014), which they use to show that the function $\mathsf{Tribes} : \{-1, 1\}^k \to \{-1, 1\}$ is approximately resilient. Starting with the Tribes function, their approach is essentially to (i) discard its low-degree Fourier component; (ii) truncate the resulting function so it does not take very large values; (iii) again discard the low-degree Fourier component of (ii) (since the truncation could have reintroduced some low-degree Fourier component); and (iv) normalize the result of (iii) to give an L^∞ norm of at most 1. They show that this yields a new function that (1) has zero low-degree Fourier weight, (2) takes output values that are bounded in $[-1, 1]$, and (3) is close to the original Tribes function in L^1 distance.

In our setting we have the same high level goals of achieving (1-3), but achieving boundedness is significantly more difficult on the unbounded domain $\mathbb{R}^k$ than on the finite hypercube $\{-1, 1\}^k$. Our witness to the approximate resilience of Cube_k is not constructed in a single shot (in contrast to Dachman-Soled et al.), but rather is constructed gradually through an iterative process.

4. See Appendix B for notation for Hermite coefficients.

3.1. Approximate resilience of functions with small low-degree weight (Proof of Lemma 10)

In this section, we prove Lemma 10. Our key tool is the $\mathrm{TruncHigh}_{d,\tau}$ transformation, defined below, and a careful iterative application of $\mathrm{TruncHigh}_{d,\tau}$ to produce a witness to the approximate resilience of a given Boolean function f with small low-degree Hermite weight.

Definition 11 *For any $f \in L^2(\mathcal{N}(0,I_k))$ and $d \in \mathbb{N}$, let Low_d and High_d be $L^2(\mathcal{N}(0,I_k)) \to L^2(\mathcal{N}(0,I_k))$ transformations that reduce a function to its low-degree and high-degree Hermite components respectively, i.e.*

$$\mathrm{Low}_d[f] := \sum_{|J|\le d} \tilde{f}(J) H_J, \quad \text{and} \quad \mathrm{High}_d[f] := \sum_{|J|>d} \tilde{f}(J) H_J = f - \mathrm{Low}_d[f].$$

For any $\tau > 0$, the truncation transformation $\mathrm{TruncHigh}_{d,\tau} : L^2(\mathcal{N}(0,I_k)) \to L^2(\mathcal{N}(0,I_k))$ is

$$\begin{aligned}\mathrm{TruncHigh}_{d,\tau}[f](x) &:= \mathrm{High}_d[f](x) - \mathrm{High}_d[f](x)\mathbb{1}\{|\mathrm{Low}_d[f](x)| > \tau\} \\ &= \begin{cases} \mathrm{High}_d[f](x) & \text{if } |\mathrm{Low}_d[f](x)| \le \tau, \\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

The purpose of $\mathrm{TruncHigh}_{d,\tau}[f]$ is to shrink the low-degree weight of f while staying bounded in L^∞ and close to f in L^1. These properties are given in the following propositions.

Proposition 12 *If $\|f\|_\infty < \infty$, then $\|\mathrm{TruncHigh}_{d,\tau}[f]\|_\infty \le \|f\|_\infty + \tau$.*

Proof. $\mathrm{TruncHigh}_{d,\tau}[f](x)$ is non-zero only if $|\mathrm{Low}_d[f](x)| \le \tau$. In that case, it is clear that $\left|\mathrm{TruncHigh}_{d,\tau}[f](x)\right| = |\mathrm{High}_d[f](x)| = |f(x) - \mathrm{Low}_d[f](x)| \le |f(x)| + \tau$. ■

Proposition 13 *For any $k \ge 1$ and $d \ge 2$, fix some $a > 1$ and $\rho \ge \|\mathrm{Low}_d[f]\|_2$ and let*

$$\tau := \rho\left(4e\ln(3k) + \frac{8e}{d}\ln\left(\frac{a\,\|f\|_2}{\rho}\right)\right)^{d/2}. \tag{1}$$

Then, (i) $\|\mathrm{Low}_d[\mathrm{TruncHigh}_{d,\tau}[f]]\|_2 \le \frac{\rho}{a}$, and (ii) $\|\mathrm{TruncHigh}_{d,\tau}[f] - f\|_1 \le 2\rho$.

We prove Proposition 13 in Appendix C.2.

Note that there is a tension in the choice of the truncation parameter τ. If τ is too large, then $\mathrm{TruncHigh}_{d,\tau}[f]$ might still take large values. But if τ is too small, then the low-degree weight of $\mathrm{TruncHigh}_{d,\tau}[f]$ might not become much smaller compared to that of f. The proof of Lemma 10 works by applying $\mathrm{TruncHigh}_{d,\tau}$ iteratively with a carefully chosen decreasing schedule of τ-values. This process converges to a function that is bounded, has zero low-degree weight, and is sufficiently close to f, and this function certifies the α-approximate d-resilience of f.

Proof of Lemma 10. Since any f with $\|f\|_\infty \le 1$ is trivially 1-approximately d-resilient for all $d \ge 0$, we may assume that $\alpha < 1$. We define a sequence of functions $(f_i)_{i\in\mathbb{N}}$ by $f_0 := f$ and $f_i := \mathrm{TruncHigh}_{d,\tau_i}[f_{i-1}]$ for $i \ge 1$, where

$$\tau_i := \frac{\|\mathrm{Low}_d[f_0]\|_2}{4^{(i-1)d}}\left(4e\ln(3k) + \frac{8e}{d}\ln\left(\frac{4^{id}\,\|f_{i-1}\|_2}{\|\mathrm{Low}_d[f_0]\|_2}\right)\right)^{d/2}. \tag{2}$$

We'll show that the sequence $(f_i)_{i\in\mathbb{N}}$ has a limit in $L^2(\mathcal{N}(0,I_k))$ that yields a witness to the α-approximate d-resiliance of f. To do this, it will suffice to show the following claims for all $i \ge 1$:

Claim 1. $\tau_i \le \frac{\alpha}{3\cdot 2^i}$.

Claim 2. $\|f_i\|_\infty \le 1 + \frac{\alpha}{3}\sum_{\iota=1}^{i}\frac{1}{2^\iota} \le 1 + \frac{\alpha}{3}$.

Claim 3. $\|\mathrm{Low}_d[f_i]\|_2 \le \frac{1}{4^{id}}\|\mathrm{Low}_d[f_0]\|_2 \le \frac{\alpha}{6\cdot 4^{id}}$ and $\|f_i - f_{i-1}\|_1 \le \frac{\alpha}{3\cdot 4^{(i-1)d}}$.

We now explain why this is enough to prove the lemma. Claim 2 ensures that $\|f_i\|_\infty \le 1 + \alpha/3$, while Claim 3 (for all i) ensures that $\|f_i - f_0\|_1 \le \sum_{\iota=1}^{i}\|f_\iota - f_{\iota-1}\|_1 \le 2\alpha/3$ (by the triangle inequality), and also $\lim_{i\to\infty}\|\mathrm{Low}_d[f_i]\|_2 = 0$. By a limit argument (Proposition 31), the sequence $(f_i)_{i\in\mathbb{N}}$ converges in $L^2(\mathcal{N}(0,I_k))$ to some $f^* \in L^2(\mathcal{N}(0,I_k))$ with $\|f^*\|_\infty \le 1 + \alpha/3$, $\mathrm{Low}_d[f^*] = 0$, and $\|f^* - f\|_1 \le 2\alpha/3$. This proves the lemma because one of f^* and $f^{**} := f^*/\|f^*\|_\infty$ witnesses that f is α-approximately d-resilient. Indeed, if $\|f^*\|_\infty > 1$, then $\|f^{**}\|_\infty = 1$, $\mathrm{Low}_d[f^{**}] = 0$, and

$$\|f - f^{**}\|_1 \le \|f - f^*\|_1 + \|f^* - f^{**}\|_\infty \le \frac{2\alpha}{3} + \left(1 - \frac{1}{\|f^*\|_\infty}\right)\|f^*\|_\infty \le \alpha,$$

where the first inequality uses the triangle inequality and comparison of $\|\cdot\|_1$ and $\|\cdot\|_\infty$.

It remains to prove Claim 1, Claim 2, and Claim 3 for all $i \ge 1$ by induction on i.

For the base case $i = 1$, $\tau_1 \le \frac{\alpha}{6}$ (Claim 1) is an immediate consequence of the upper bound on τ_1 from Fact 30 in Appendix C.3, which relies on having $\|f_0\|_\infty = 1 < \frac{4}{3}$. Proposition 12 and the bound on τ_1 imply $\|f_1\|_\infty \le 1 + \frac{\alpha}{6}$ (Claim 2). By taking $a = 4^d$ and $\rho = \|\mathrm{Low}_d[f_0]\|_2$, Proposition 13 implies that $\|\mathrm{Low}_d[f_1]\|_2 \le \frac{1}{4^d}\|\mathrm{Low}_d[f_0]\|_2$ and $\|f_1 - f_0\|_1 \le 2\|\mathrm{Low}_d[f_0]\|_2$. We conclude the base case of Claim 3 by observing that $\|\mathrm{Low}_d[f_0]\|_2 \le \tau_1 \le \frac{\alpha}{6}$ by Fact 30.

We prove the inductive step by assuming that the three claims all hold for some fixed $i \ge 1$ and showing that they also hold for $i+1$. By applying Fact 30 with $\|f_i\|_\infty \le 1 + \frac{\alpha}{3} \le \frac{4}{3}$ from step i of Claim 2, we have $\tau_{i+1} \le \frac{\alpha}{3\cdot 2^{i+1}}$ (step $i+1$ of Claim 1). Step $i+1$ of Claim 2 is immediate from Proposition 12, the bound on τ_{i+1}, and a geometric sum:

$$\|f_{i+1}\|_\infty \le \|f_i\|_\infty + \tau_{i+1} \le 1 + \frac{\alpha}{3}\sum_{\iota=1}^{i+1}\frac{1}{2^\iota} \le 1 + \frac{\alpha}{3}.$$

We apply Proposition 13 with $a = 4^d$ and $\rho = \frac{1}{4^{id}}\|\mathrm{Low}_d[f_0]\|_2$[5] to obtain

$$\|\mathrm{Low}_d[f_{i+1}]\|_2 \le \frac{1}{4^{(i+1)d}}\|\mathrm{Low}_d[f_0]\|_2 \quad \text{and} \quad \|f_{i+1} - f_i\|_1 \le \frac{2}{4^{id}}\|\mathrm{Low}_d[f_0]\|_2.$$

Combining this with the bound $\|\mathrm{Low}_d[f_0]\|_2 \le \frac{\alpha}{6}$ completes step $i+1$ of Claim 3.

Hence, the three claims hold for all $i \ge 1$ by induction, which concludes the proof. ■

3.2. Approximate resilience of Cube_k (Proof of Lemma 8)

Let $d = \frac{\ln k}{125\ln\ln k}$. By Lemma 9, for sufficiently large k we have

$$\gamma := \sum_{|J|\le d}\widetilde{\mathsf{Cube}_k}(J)^2 \le \frac{20d(3\ln k)^d}{k}.$$

5. Note that $\rho \ge \|\mathrm{Low}_d[f_i]\|_2$ by step i of Claim 3, which is necessary for Proposition 13.

Lemma 10 guarantees that Cube_k is α-approximately d-resilient for

$$\alpha = \gamma^{0.498}(72 \ln k)^{d/2} \leq \exp\left(-0.498 \ln k + 0.00799 \ln k + o(\ln k)\right) \leq k^{-0.49}.$$

This completes the proof of Lemma 8. ■

4. Hardness of SQ learning to constant excess error via weak learning lower bounds

In this section we give a different proof of our main agnostic SQ hardness result for learning intersections of k halfspaces to constant excess error. While Theorem 7 established hardness for a highly structured subclass of this concept class (consisting of suitable embeddings of the Cube_k function), the current argument only applies to the broader class of all intersections of k halfspaces. However, an advantage of the current argument is that it holds for a wider range of values of k (up to $2^{O(n^{0.245})}$). In more detail, in this section we prove the following:

Theorem 14 *For sufficiently large n and any $k = 2^{O(n^{0.245})}$, any SQ algorithm that agnostically learns the class of intersections of k halfspaces over $\mathbb{R}^n$ to excess error c requires either $2^{n^{\Omega(1)}}$ queries or at least one query of tolerance $n^{-\Omega(\log(k)/\log\log k)}$. (Here $c > 0$ is an absolute constant independent of all other parameters.)*

As discussed in Section 1.1, the proof of Theorem 14 follows the high-level approach of Theorem 1.4 of Dachman-Soled et al. (2014). Rather than analyzing the Hermite spectrum of the hard-to-learn functions (as was done in Section 3), the argument combines the agnostic learning algorithm of Kalai et al. (2008) with a (slight extension of a) recently-established lower bound on the ability of membership query (MQ) algorithms to weakly learn intersections of halfspaces.

We first recall the following lower bound from De and Servedio (2021):

Lemma 15 (De and Servedio, 2021, Theorem 2) *For sufficiently large m, for any $q \geq m$, there is a distribution $\mathcal{D}_{\mathrm{actual}}$ over centrally symmetric convex sets (specifically, intersections of $O(q^{100})$ halfspaces) of $\mathbb{R}^m$ with the following property: for a target convex set $\boldsymbol{f} \sim \mathcal{D}_{\mathrm{actual}}$ for any MQ algorithm A making at most q many queries to $\boldsymbol{f}$, the expected error of A (the probability over $\boldsymbol{f} \sim \mathcal{D}_{\mathrm{actual}}$, any internal randomness of A, and a Gaussian $\mathbf{x} \sim \mathcal{N}(0, I_n)$, that the output hypothesis h of A is wrong on $\mathbf{x}$) is at least $\frac{1}{2} - \frac{O(\log q)}{\sqrt{m}}$.*

We require the following corollary of Lemma 15, which we prove in Appendix D.1.

Corollary 16 *For sufficiently large n, for all $q \geq m$, there is a distribution $\mathcal{D}$ over intersections of q^{101} halfspaces such that for a target function $\boldsymbol{f} \sim \mathcal{D}$, any MQ algorithm $\mathcal{A}$ making at most q queries to $\boldsymbol{f}$ has expected error at least $\frac{1}{2} - \frac{O(\log q)}{\sqrt{m}}$ (where the expectation is over $\boldsymbol{f} \sim \mathcal{D}$ and any internal randomness of $\mathcal{A}$, and the the accuracy is with respect to $\mathcal{N}(0, I_n)$).*

Theorem 14 follows immediately from Lemma 6 and the following lemma.

Lemma 17 *For any $k = 2^{O(n^{0.245})}$, there exists an intersection of k halfspaces $f : \mathbb{R}^m \to \{-1, 1\}$ that has L^1 $\frac{1}{2}$-approximate degree $d = \Omega(\log(k)/\log\log k)$, where $m = O(n^{0.49})$.*

Proof. First we note that we may assume k is at least some sufficiently large absolute constant as specified below through the choice of q (since otherwise, because of the $\Omega(\cdot)$ in the specification of d, there is nothing to prove). Suppose that every intersection of k halfspaces f over $\mathbb{R}^m$ has L^1 $\frac{1}{2}$-approximate degree at most $d-1$; we will prove the lemma by showing that d must be $\Omega(\log(k)/\log\log k)$.

Let $q = k^{1/101}$ and let $\mathcal{S} \subseteq \mathbb{R}^n$ be the subspace of $\mathbb{R}^n$ spanned by the first $m = c_1 \ln^2 q$ coordinates, where c_1 is a sufficiently large universal constant specified below and q is chosen sufficiently large (relative to c_1) so that $q \geq m$ and m satisfies the "sufficiently large" requirement of Corollary 16. By Corollary 16, there is a distribution $\mathcal{D}$ over intersections of at most $k = q^{101}$ halfspaces over $\mathcal{S}$ such that any membership query algorithm making at most q queries to an unknown $\boldsymbol{f} \sim \mathcal{D}$ outputs a hypothesis with expected error at least $\frac{1}{2} - \frac{O(\log q)}{\sqrt{m}}$. For a sufficiently large setting of c_1, this expected error is at least $\frac{1}{2} - \frac{O(\log q)}{\sqrt{c_1 \ln q}} \geq 0.49$.

By the assumption that every intersection of k halfspaces has L^1 $\frac{1}{2}$-approximate degree at most $d-1$, if the agnostic learner of Theorem 5 of Kalai et al. (2008) is run on any intersection of k halfspaces over the first m coordinates, then it uses $s := \mathrm{poly}(m^d/\epsilon)$ labeled examples from $\mathcal{N}(0, I_n)$, runs in $\mathrm{poly}(s)$ time, and with probability at least (say) 0.9 outputs a hypothesis h with error at most

$$\epsilon + \frac{1}{2} \min_{p \in \mathcal{P}_{n,d}} \|f - p\|_1 \leq \epsilon + \frac{1}{4}$$

(see Theorem 1.3 of Dachman-Soled et al. (2014)). Taking $\epsilon = 0.15$ and observing that a labeled example from $\mathcal{N}(0, I_n)$ can be simulated using a single membership query, we see that for the concept class of intersections of k halfspaces over the first m coordinates, there is a membership query algorithm $\mathcal{A}$ that makes at most $m^{c_2 d}$ many membership queries and with probability at least 0.9 achieves error at most 0.4; hence the expected error of this MQ algorithm is at most $0.9 \cdot 0.4 + 0.1 \cdot 1 = 0.46$.

Comparing the conclusions of the previous two paragraphs, we see that $m^{c_2 d} \geq q$, and hence (recalling that $m = c_1 \ln^2 q$ and $q = k^{1/101}$), we get that

$$d \geq \frac{\ln q}{c_2 \ln m} = \Omega(\log(k)/\log\log k),$$

which proves the lemma. ∎

5. Hardness of SQ learning to arbitrary excess error

In this section, we strengthen both of the SQ lower bounds from Sections 3 and 4 by combining them with lower bounds on the L^1 ϵ-approximate degree of halfspaces due to Ganzburg (2002). By doing so, we improve the lower bounds to $n^{\Omega(\log(k)/\log\log(k) + 1/\epsilon^2)}$ for agnostically learning intersections of k halfspaces to excess error ϵ for any $\epsilon \geq n^{-c}$ (cf. Lemma 6). By the arithmetic-geometric mean inequality, this lower bound is always at least as strong as the $n^{-\tilde{\Omega}(\log^{1/2}(k)/\epsilon)}$ lower bound of Diakonikolas et al. (2021b).

Let k, m be as described in Lemma 17. We construct an intersection of $k+1$ halfspaces over $\mathbb{R}^{m+1}$ by taking the intersection of

- the k halfspaces identified in Lemma 17 over $\mathbb{R}^m$; and
- an origin-centered halfspace orthogonal to the $(m+1)$-st coordinate basis vector.

Appendix E formally bounds the L^1 approximate degree of intersections of this construction of half spaces and proves a strengthening of Theorem 14, which is our main agnostic SQ lower bound:

Theorem 18 (Formal version of Theorem 1) *For any $k = 2^{O(n^{0.245})}$ and any $\epsilon \geq n^{-c}$ for a suitably small absolute constant $c > 0$, any SQ algorithm that agnostically learns intersections of k halfspaces to excess error ϵ under Gaussian marginals requires either $2^{n^{\Omega(1)}}$ queries or at least one query of tolerance $n^{-\Omega(\log(k)/\log\log k + 1/\epsilon^2)}$.*

Acknowledgments

In this work, C. Sanford and D. Hsu are supported by NSF grants CCF-1740833 and IIS-1563785, NASA ATP grant 80NSSC18K109, a Sloan Research Fellowship, and a Google Faculty Award. C. Sanford, R.A. Servedio, and E.-V. Vlatakis-Gkaragkounis are supported by NSF grants CCF-1814873, IIS-1838154, CCF-1563155, and by the Simons Collaboration on Algorithms and Geometry. C. Sanford gratefully acknowledges support from the National Science Foundation Graduate Research Fellowship Program (NSF GRFP). E.-V. Vlatakis-Gkaragkounis is additionally supported by NSF grants NSF-CCF-1763970 and NSF-CCF-2107187 and by the Onassis Foundation under Scholarship ID: F ZN 010-1/2017-2018.

References

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions*. Dover, 1972.

Rosa I. Arriaga and Santosh S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

Pranjal Awasthi, Maria Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM*, 63(6):1–27, 2017.

Keith Ball. The Reverse Isoperimetric Problem for Gaussian Measure. *Discrete and Computational Geometry*, 10:411–420, 1993.

Eric B. Baum. A polynomial time algorithm that learns two hidden unit nets. *Neural Computation*, 2(4):510–522, 1990.

Avrim L. Blum and Ravindran Kannan. Learning an intersection of a constant number of halfspaces over a uniform distribution. *Journal of Computer and System Sciences*, 54(2):371–380, 1997.

Avrim L. Blum and Ronald L. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5(1):117–127, 1992.

Avrim L. Blum, Carl Burch, and John Langford. On learning monotone Boolean functions. In *Thirty-Ninth Annual Symposium on Foundations of Computer Science*, 1998a.

Avrim L. Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1):35–52, 1998b.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

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

Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2014.

Anindya De and Rocco A. Servedio. Weak learning convex sets under normal distributions. In *Thirty-Fourth Annual Conference on Learning Theory*, 2021.

Anindya De, Shivam Nadimpalli, and Rocco A. Servedio. Convex influences. *arXiv preprint arXiv:2109.03107*, 2021.

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. *Fifty-Eighth Annual Symposium on Foundations of Computer Science*, 2017.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Efficiently learning halfspaces with Tsybakov noise. In *Fifty-Third Annual ACM Symposium on Theory of Computing*, 2021a.

Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. The optimality of polynomial regression for agnostic learning under Gaussian marginals in the SQ model. In *Thirty-Fourth Annual Conference on Learning Theory*, 2021b.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification and Scene Analysis*. Wiley New York, 1973.

William Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, Inc., 3rd edition, 1968.

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

Alan Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In *FOCS*, 1996.

Michael I. Ganzburg. Limit theorems for polynomial approximations with Hermite and Freud weights. In *Approximation Theory X: Wavelets, Splines, and Applications*, Innovations in Applied Mathematics, pages 211–221. Vanderbilt University Press, 2002.

Aravind Gollakota, Sushrut Karmalkar, and Adam R. Klivans. The polynomial method is universal for distribution-free correlational SQ learning. *arXiv preprint arXiv:2010.11925*, 2020.

Parikshit Gopalan, Adam R. Klivans, and Raghu Meka. Learning functions of halfspaces using prefix covers. In *Twenty-Fifth Annual Conference on Learning Theory*, 2012.

Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

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

Daniel M. Kane. The average sensitivity of an intersection of half spaces. In *Forty-Sixth Annual ACM Symposium on Theory of Computing*, 2014.

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

Adam R. Klivans and Rocco A. Servedio. Learning intersections of halfspaces with a margin. *Journal of Computer and System Sciences*, 74(1):35–48, 2008.

Adam R. Klivans and Alexander A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. In *Forty-Seventh Annual IEEE Symposium on Foundations of Computer Science*, 2006.

Adam R. Klivans and Alexander A. Sherstov. Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning*, 69(2):97–114, 2007.

Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, 68(4):808–840, 2004.

Adam R. Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via Gaussian surface area. In *Forty-Ninth Annual IEEE Symposium on Foundations of Computer Science*, 2008.

Adam R. Klivans, Philip M. Long, and Alex K. Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. 2009.

Stephen Kwek and Leonard Pitt. PAC learning intersections of halfspaces with membership queries. *Algorithmica*, 22(1):53–75, 1998.

Yishay Mansour. An $O(n^{\log\log n})$ Learning Algorithm for DNF under the Uniform Distribution. In *Fifth Annual Workshop on Computational Learning Theory*, 1992.

Nimrod Megiddo. On the complexity of polyhedral separability. *Discrete & Computational Geometry*, 3(4):325–337, 1988.

Fedor Nazarov. *On the Maximal Perimeter of a Convex Set in $\mathbb{R}^n$ with Respect to a Gaussian Measure*, pages 169–187. Springer Berlin Heidelberg, 2003.

Edward Nelson. Construction of quantum fields from Markoff fields. *Journal of Functional Analysis*, 12(1):97–112, 1973.

Phong Q. Nguyen and Oded Regev. Learning a parallelepiped: Cryptanalysis of GGH and NTRU signatures. In *Eurocrypt*, 2006.

Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

Frank Rosenblatt. *Principles of Neurodynamics*. Spartan Books, 1962.

Alexander A. Sherstov. Optimal bounds for sign-representing the intersection of two halfspaces by polynomials. *Combinatorica*, 33(1):73–96, 2013.

Jeffrey D. Vaaler. Some extremal functions in Fourier analysis. *Bulletin of the American Mathematical Society*, 12(2):183–216, 1985.

Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, 1982.

Santosh S. Vempala. A random sampling based algorithm for learning the intersection of half-spaces. In *Thirty-Eighth Annual Symposium on Foundations of Computer Science*, 1997.

Santosh S. Vempala. Learning convex concepts from Gaussian distributions with PCA. In *Fifty-First Annual Symposium on Foundations of Computer Science*, 2010a.

Santosh S. Vempala. A random-sampling-based algorithm for learning intersections of halfspaces. *Journal of the ACM*, 57(6):1–14, 2010b.

Appendix A. Discussion

A.1. Agnostic SQ lower bounds for learning functions of bounded Gaussian surface area and convex m-subspace juntas

In this appendix, we note that our arguments imply agnostic SQ lower bounds for several classes of $\{-1,1\}$-valued functions over $\mathbb{R}^n$ that were studied by Vempala (2010a) and Klivans et al. (2008). Our lower bounds essentially match the upper bounds for those classes by Klivans et al. (2008).

Functions with bounded Gaussian Surface Area. Recall the definition of Gaussian Surface Area:

Definition 19 *Let $f : \mathbb{R}^n \to \{-1,1\}$ be such that $\{x \in \mathbb{R}^n : f(x) = 1\}$ is a Borel set. The* Gaussian surface area *of f is defined to be*

$$\Gamma(f) := \liminf_{\delta \to 0} \frac{\Pr_{\mathbf{x} \sim \mathcal{N}(0,I_n)}\left[f(\mathbf{x}) = -1 \text{ and } \exists y \in f^{-1}(1) \text{ s.t. } \|\mathbf{x} - y\|_2 \leq \delta\right]}{\delta}.$$

Let $\mathcal{C}_s$ denote the class of all Borel sets in $\mathbb{R}^n$ with Gaussian surface area at most s. The main result of Klivans et al. (2008, Theorem 25) is that $\mathcal{C}_s$ is agnostically learnable to accuracy $\mathrm{OPT} + \epsilon$ by an SQ algorithm that makes $n^{O(s^2/\epsilon^4)}$ queries, each of tolerance $n^{-O(s^2/\epsilon^4)}$. Their agnostic learning algorithm for intersections of k halfspaces, mentioned earlier, is obtained from this result by combining it with the fact, due to Nazarov (2003), that any intersection of k halfspaces has Gaussian surface area at most $O(\sqrt{\log k})$.

Let $s = O(n^{0.1225})$, let $m = n^{0.49}$ and let $k = 2^{s^2} = 2^{O(\sqrt{m})}$. By Lemma 17 there is an intersection of k halfspaces over $\mathbb{R}^m$ that has L^1 $\frac{1}{2}$-approximate degree $\Omega(s^2/\log s)$, and by Nazarov's upper bound on Gaussian surface area, this function has Gaussian surface area at most $O(s)$. Combining this with Lemma 6, we immediately obtain that for any $s \leq O(n^{0.1225})$, any SQ agnostic learning algorithm that achieves constant excess error under Gaussian marginals for the class $\mathcal{C}_s$ either requires queries with tolerance at most $n^{-\Omega(s^2/\log(s))}$ or makes at least $2^{n^{\Omega(1)}}$ queries. Combining this with the arguments of Section 5, we get the following result for $\mathcal{C}_s$:

Theorem 20 *For sufficiently large n, any $s = O(n^{0.1225})$, and any $\epsilon \geq n^{-c}$ for a suitably small absolute constant $c > 0$, any SQ algorithm that agnostically learns the class $\mathcal{C}_s$ to excess error ϵ requires either $2^{n^{\Omega(1)}}$ queries or at least one query of tolerance $n^{-\Omega(s^2/\log(s)+1/\epsilon^2)}$.*

Convex subspace juntas. Vempala (2010a) gave a learning algorithm (in the realizable, i.e., non-agnostic, setting) for a class of functions that we refer to as *convex m-subspace juntas*. A function $f : \mathbb{R}^n \to \{-1,1\}$ is a convex m-subspace junta if f is the indicator function of a convex set K with a normal subspace of dimension m; equivalently, f is an intersection of halfspaces all of whose normal vectors lie in some subspace of $\mathbb{R}^n$ of dimension at most m (note that the number of halfspaces in such an intersection may be arbitrarily large or even infinite).

Vempala's algorithm learns to accuracy ϵ and runs in time $\mathrm{poly}(n, 2^m/\epsilon, m^{\tilde{O}(\sqrt{m}/\epsilon^4)})$ in the realizable ($\mathrm{OPT} = 0$) setting of learning under Gaussian marginals. As alluded to in Section 1, this algorithm uses principal component analysis to do a preprocessing step and then runs the algorithm of Klivans et al. (2008). The analysis crucially relies on a Brascamp-Lieb type inequality (Lemma 4.7 of Vempala, 2010a) which, roughly speaking, makes it possible to identify the "relevant directions"); however, this breaks down in the non-realizable (agnostic) setting. The best known agnostic learning result for the class of convex m-subspace juntas under Gaussian marginals is

the SQ algorithm of Klivans et al. (2008), which makes $n^{O(\sqrt{m}/\epsilon^4)}$ statistical queries, each of tolerance at least $n^{-O(\sqrt{m}/\epsilon^4)}$. This performance bound for the algorithm follows immediately from Theorem 25 of Klivans et al. (2008) and the upper bound, due to Ball (1993), that any convex set in $\mathbb{R}^m$ has Gaussian surface area at most $O(m^{1/4})$.

Let $m \leq n^{0.49}$. By Lemma 17 there is a convex m-subspace junta (an intersection of $2^{O(\sqrt{m})}$ many halfspaces, all of whose normal vectors lie in an m-dimensional subspace of $\mathbb{R}^n$) that has L^1 $\frac{1}{2}$-approximate degree $\Omega(\sqrt{m}/\log m)$. Combining this with Lemma 6 and the arguments of Section 5, we obtain the following lower bound:

Theorem 21 *For sufficiently large n, any $m \leq n^{0.49}$, and $\epsilon \geq n^{-c}$ for a suitably small absolute constant $c > 0$, any SQ algorithm that agnostically learns the class of convex m-subspace juntas to excess error ϵ requires either $2^{n^{\Omega(1)}}$ queries or at least one query of tolerance $n^{-\Omega(\sqrt{m}/\log m + 1/\epsilon^2)}$.*

A.2. On lower bounds for L^1 polynomial approximation

One of the contributions of Diakonikolas et al. (2021b) is that it introduced new analytic techniques for obtaining lower bounds on the L^1 approximate degree of functions $f : \mathbb{R}^n \to \{-1, 1\}$. In particular, Diakonikolas et al. (2021b) established a new structural result that translates a lower bound on the Gaussian Noise Sensitivity of any function $f : \mathbb{R}^n \to \{-1, 1\}$ to a lower bound on the L^1 approximate degree of f.

Definition 22 (O'Donnell, 2014, Definition 11.9) *Given $0 \leq \rho \leq 1$ and $f : \mathbb{R}^n \to \{-1, 1\}$, the* Gaussian Noise Sensitivity of f at correlation $1 - \rho$, *written* $\mathrm{GNS}_\rho(f)$, *is*

$$\mathrm{GNS}_\rho(f) := \Pr_{(\mathbf{x},\mathbf{g}) \sim \mathcal{N}(0,I_n)^{\otimes 2}} \Big[f(\mathbf{x}) \neq f((1-\rho)\mathbf{x} + \sqrt{2\rho - \rho^2}\mathbf{g}) \Big].$$

Equivalently, $\mathrm{GNS}_\rho(f)$ is the probability that $f(\mathbf{x}) \neq f(\mathbf{y})$ where $\mathbf{x}, \mathbf{y}$ are standard n-dimensional Gaussians with correlation $1 - \rho$.

Theorem 23 (Diakonikolas et al., 2021b, Theorem 1.5) *Let $f : \mathbb{R}^n \to \{-1, 1\}$ and let $p : \mathbb{R}^n \to \mathbb{R}$ be any polynomial of degree at most d. Then*

1. $\|f - p\|_1 \geq \Omega(1/\log d) \cdot \mathrm{GNS}_{(\ln(d)/d)^2}(f)$.
2. *For any $\epsilon > 0$, we have* $\|f - p\|_1 \geq \mathrm{GNS}_\epsilon(f)/4 - O(d\sqrt{\epsilon})$.

In contrast with L^2 polynomial approximation (for which the degree required for ϵ-approximation can be "read off" of the Hermite expansion), polynomial approximation in L^1 is much less well understood. Thus it is interesting and useful to have general tools for L^1 approximate degree bounds such as Theorem 23, and conversely, it is of interest to understand the limitations of such tools.

Diakonikolas et al. (2021b) use Theorem 23 to prove an L^1 approximate degree lower bound for intersections of k halfspaces. They first show that for a particular[6] intersection of k halfspaces f' over $\mathbb{R}^k$, for each $\tau < \Theta(1/\log k)$ it holds that $\mathrm{GNS}_\tau(f') = \Theta(\sqrt{\tau \log k})$. Combining this with item (1) of Theorem 23 gives that any polynomial p for which $\|f - p\|_1 \leq \epsilon$ must have

6. This function f' is very similar to the Cube_k function; instead of upper and lower bounding each of the k coordinates $x_1, \ldots, x_k$, it only upper bounds each coordinate.

$d \geq \Omega(\frac{\log^{1/2} k}{\epsilon})$. Our resilience results for the Cube_k function give a stronger L^1 approximate degree lower bound, and combining this with the GNS bound from Diakonikolas et al. (2021b) gives an example of a function for which the bound of part (1) of Theorem 23 is not tight.

In more detail, recall that our Lemma 8 states that the Cube_k function is $k^{-0.49}$-approximately $\Theta(\log(k)/\log\log k)$-resilient. An entirely similar analysis to the proof of Lemma 8 shows that the function f' of Diakonikolas et al. (2021b) is also $k^{-0.49}$-approximately $d := \Theta(\log(k)/\log\log k)$-resilient, i.e., there is a function $g : \mathbb{R}^k \to [-1, 1]$ which has zero correlation with every polynomial of degree at most $d-1$ and which has $\|f' - g\|_1 \leq k^{-0.49}$. By Lemma 5, the existence of this resilient g implies that every polynomial p of degree at most $d-1$ must have

$$\|f' - p\|_1 \geq 1 - \frac{2}{k^{0.49}},$$

which is close to one for large k.

Now consider what can be obtained from the GNS bound of Diakonikolas et al. (2021b). Since

$$\mathrm{GNS}_{((\ln(d-1))/(d-1))^2}(f') = \sqrt{\frac{\Theta((\log\log k)^4)}{(\ln k)^2} \cdot \ln k} = \frac{\Theta((\log\log k)^2)}{\sqrt{\ln k}},$$

part (1) of Theorem 23 only gives that every polynomial p of degree at most $d-1$ has

$$\|f' - p\|_1 \geq \Omega\left(\frac{\log\log(k)}{\sqrt{\log k}}\right).$$

This bound is close to zero for large k.

Appendix B. Hermite polynomials and Gaussian hypercontractivity

Let $\{h_j\}_{j=0}^{\infty}$ be the (unnormalized) probabilists' Hermite polynomials

$$h_j(x) := (-1)^j e^{x^2/2} \frac{\mathrm{d}^j}{\mathrm{d}x^j} e^{-x^2/2}, \quad j = 0, 1, 2, \ldots . \tag{3}$$

These polynomials form an orthogonal basis for the Hilbert space $L^2(\mathcal{N}(0,1))$; more precisely, we have $\langle h_j, h_{j'} \rangle = j! \cdot \delta_{j,j'}$. For any $f \in L^2(\mathcal{N}(0,1))$, the Hermite coefficients $\widetilde{f}(j)$ of f are given by

$$\widetilde{f}(j) := \frac{1}{\sqrt{j!}} \langle f, h_j \rangle .$$

Let $\{H_J\}_{J \in \mathbb{N}^k}$ be the multivariate Hermite polynomials, which correspond to a tensor product of the univariate Hermite polynomials above. That is,

$$H_J(x) := \prod_{i=1}^{k} h_{J_i}(x_i).$$

These polynomials form an orthogonal basis for $L^2(\mathcal{N}(0, I_k))$, and we have that $\langle H_J, H_{J'} \rangle = J! \delta_{J,J'}$, where $J! = J_1! \cdots J_k!$. For any $F \in L^2(\mathcal{N}(0, I_k))$, the Hermite coefficients $\widetilde{F}(J)$ of F are given by

$$\widetilde{F}(J) := \frac{1}{\sqrt{J!}} \langle F, H_J \rangle .$$

Additional properties of the Hermite polynomials can be found in Chapter 22 of Abramowitz and Stegun (1972) and Section 11.2 of O'Donnell (2014).

Our results—particularly those in Section 3.1—rely on bounds powered by Gaussian hypercontractivity. We recall the basic Gaussian hypercontractive inequality for low-degree polynomials (Bonami, 1970; Nelson, 1973; Gross, 1975):

Fact 24 *For a polynomial $p \in \mathcal{P}_d$ and any $q \geq 2$, $\|p\|_q \leq (q-1)^{d/2} \|p\|_2$.*

In particular, we will use the following bound on the fourth moment of Hermite polynomials, which follows immediately from Fact 24 and standard bounds on the norm of Hermite polynomials:

Fact 25 $\|H_J\|_4 \leq 3^{d/2} \|H_J\|_2 \leq 3^{d/2}\sqrt{J!}$.

We will also use the following concentration bound, which follows from Gaussian hypercontractivity using Markov's inequality:

Fact 26 (O'Donnell, 2014, Theorem 9.23) *For any polynomial $p : \mathbb{R}^k \to \mathbb{R}$ of degree d and any $t \geq e^d$,*

$$\Pr_{\mathbf{x} \sim \mathcal{N}(0, I_n)} \left[|p(\mathbf{x})| \geq t \|p\|_2\right] \leq \exp\left(-\frac{d}{2e} t^{2/d}\right).$$

Proof. Consider any $q \geq 2$.

$$\Pr\left[|p(\mathbf{x})| \geq t \|p\|_2\right] = \Pr\left[|p(\mathbf{x})|^q \geq t^q \|p\|_2^q\right] \leq \frac{\|p\|_q^q}{t^q \|p\|_2^q} \leq \left(\frac{(q-1)^{d/2}}{t}\right)^q \leq \left(\frac{q^{d/2}}{t}\right)^q.$$

Let $q = \frac{t^{2/d}}{e}$, which has $q \geq 2$ because $t \geq e^d$. Then,

$$\Pr\left[|p(\mathbf{x})| \geq t \|p\|_2\right] \leq \left(\frac{1}{e^{d/2}}\right)^{t^{2/d}/e} = \exp\left(-\frac{d}{2e} t^{2/d}\right). \qquad \blacksquare$$

Appendix C. Supporting lemmas and proofs for Section 3

C.1. Small low-degree Hermite weight of Cube_k (Proof of Lemma 9)

We recall Lemma 9.

Lemma 9 *For any sufficiently large k, and any $d \geq 0$,*[7]

$$\sum_{|J| \leq d} \widetilde{\mathsf{Cube}_k}(J)^2 \leq \frac{20d(3 \ln k)^d}{k}.$$

We note that by the analysis of Cube_k by De et al. (2021, Example 14), the upper bound of Lemma 9 in the case $d = 2$ is tight up to constant factors.

Our proof of Lemma 9 uses the product structure of $\mathcal{N}(0, I_k)$ and the fact that Cube_k is essentially a product of univariate interval functions over disjoint variables. Thanks to these properties, it suffices to analyze the Hermite coefficients of interval functions of the right width.

7. See Appendix B for notation for Hermite coefficients.

For any $\theta \geq 0$, let $f_\theta : \mathbb{R} \to \{0,1\}$ be the indicator function for the interval $[-\theta, \theta]$, i.e.,

$$f_\theta(x) := \mathbb{1}\{|x| \leq \theta\}.$$

Then, Cube_k can be written as

$$\mathsf{Cube}_k(x) = 2\prod_{i=1}^{k} f_{\theta_k}(x_i) - 1.$$

Since θ_k is chosen to ensure that $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I_k)}[\mathsf{Cube}_k(\mathbf{x})] = 0$, the Hermite coefficients of Cube_k are given by

$$\widetilde{\mathsf{Cube}_k}(J) = \begin{cases} 0 & \text{if } J = 0, \\ 2\prod_{i=1}^{k} \widetilde{f_{\theta_k}}(J_i) & \text{otherwise.} \end{cases}$$

Proof of Lemma 9. We may assume that $d \leq k/(2e^2 \ln k)$, since otherwise the claimed bound on $\sum_{|J|\leq d} \widetilde{\mathsf{Cube}_k(J)}^2$ is more than one.

By Lemma 27 (stated and proved below), $\widetilde{f_{\theta_k}}(J_i) = 0$ for any odd J_i. Hence, the only Hermite coefficients that may be non-zero are those corresponding to multi-indices $J \in \mathbb{N}^k$ with (i) only even components, and (ii) $1 \leq |J| \leq d$. Let $\mathcal{J}$ denote this set of multi-indices. For any such $J \in \mathcal{J}$,

$$\begin{aligned}
\widetilde{\mathsf{Cube}_k}(J)^2 = 4\prod_{i=1}^{k} \widetilde{f_{\theta_k}}(J_i)^2 &= 4\prod_{i:J_i=0} \widetilde{f_{\theta_k}}(J_i)^2 \prod_{i:J_i\geq 2} \widetilde{f_{\theta_k}}(J_i)^2 \\
&\leq 4\prod_{i:J_i\geq 2} \left[\left(1+\sqrt{\frac{e}{J_i}}\theta_k\right)^{2(J_i-1)} e^{-\theta_k^2}\right] \\
&\leq 4\left(1+\sqrt{\frac{e}{2}}\theta_k\right)^{2(|J|-\#J)} e^{-\theta_k^2 \#J},
\end{aligned}$$

where the first inequality uses the fact that $|\widetilde{f_{\theta_k}}(0)| \leq 1$ (Lemma 27) and the bound from Lemma 28 (stated and proved below). To bound the sum $\sum_{|J|\leq d} \widetilde{\mathsf{Cube}_k}(J)^2 = \sum_{J\in\mathcal{J}} \widetilde{\mathsf{Cube}_k}(J)^2$, we partition the terms by the value of $\#J$. Note that $\#J$ must satisfy $1 \leq \#J \leq \lfloor d/2 \rfloor$, since J is not all zeros, and every non-zero component of J is at least two. Therefore,

$$\begin{aligned}
\sum_{|J|\leq d} \widetilde{\mathsf{Cube}_k}(J)^2 &= \sum_{t=1}^{\lfloor d/2\rfloor} \sum_{J\in\mathcal{J}:\#J=t} \widetilde{\mathsf{Cube}_k}(J)^2 \\
&\leq 4\left(1+\sqrt{\frac{e}{2}}\theta_k\right)^{2(d-1)} \sum_{t=1}^{\lfloor d/2\rfloor} |\{J\in\mathcal{J} : \#J = t\}| \cdot e^{-\theta_k^2 t}. \qquad (4)
\end{aligned}$$

The definition of $\mathcal{J}$ and standard binomial coefficient inequalities provide a bound on the number of $J \in \mathcal{J}$ with $\#J = t$ for $t \geq 1$:

$$\begin{aligned}
|\{J \in \mathcal{J} : \#J = t\}| &= \binom{k}{t} \left|\{S \in \mathbb{N}^t : |S| \leq \lfloor d/2 \rfloor, S_i > 0 \text{ for all } i \in [t]\}\right| \\
&= \binom{k}{t} \left|\{S \in \mathbb{N}^t : |S| \leq \lfloor d/2 \rfloor - t\}\right| \\
&= \binom{k}{t}\binom{\lfloor d/2 \rfloor}{t} \leq \left(\frac{e^2 k d}{2t^2}\right)^t.
\end{aligned}$$

Therefore, we can bound the final expression from (4) by

$$\begin{aligned}
4\left(1 + \sqrt{\frac{e}{2}}\theta_k\right)^{2(d-1)} \sum_{t=1}^{\lfloor d/2 \rfloor} \left(\frac{e^2 k d}{2t^2} e^{-\theta_k^2}\right)^t &\leq 4\left(1 + \sqrt{e \ln k}\right)^{2(d-1)} \sum_{t=1}^{\lfloor d/2 \rfloor} \left(\frac{e^2 d \ln k}{t^2 k}\right)^t \\
&\leq 8\left(1 + \sqrt{e \ln k}\right)^{2(d-1)} \cdot \frac{e^2 d \ln k}{k} \leq \frac{20 d (3 \ln k)^d}{k},
\end{aligned}$$

where the first inequality uses the bounds on θ_k from Lemma 29, and the second inequality uses the assumption $d \leq k/(2e^2 \ln k)$. ■

The preceding proof relies on three supporting lemmas: Lemma 27 and Lemma 28 compute and bound the Hermite coefficients of f_θ; Lemma 29 gives upper- and lower-bounds on θ_k.

Let $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ denote the probability density function of the one-dimensional Gaussian distribution $\mathcal{N}(0, 1)$.

Lemma 27 *For all $j \geq 0$, the Hermite coefficients of f_θ are as follows:*

$$\widetilde{f_\theta}(j) = \begin{cases} \int_{-\theta}^{\theta} \phi(x)\,\mathrm{d}x & \text{if } j = 0, \\ 0 & \text{if } j \text{ is odd}, \\ -\frac{2}{\sqrt{j!}} h_{j-1}(\theta)\phi(\theta) & \text{if } j \geq 2 \text{ is even}. \end{cases}$$

Proof. Recalling the definition of univariate Hermite polynomials from Appendix B, the degree-0 coefficient $\widetilde{f_\theta}(0)$ is

$$\widetilde{f_\theta}(0) = \int_{-\infty}^{\infty} f_\theta(x)\phi(x)\,\mathrm{d}x = \int_{-\theta}^{\theta} \phi(x)\,\mathrm{d}x.$$

The degree-j coefficient, for $j \geq 1$, is

$$\begin{aligned}
\widetilde{f_\theta}(j) &= \frac{1}{\sqrt{j!}} \int_{-\infty}^{\infty} f_\theta(x) h_j(x)\phi(x)\,\mathrm{d}x = \frac{1}{\sqrt{j!}} \int_{-\theta}^{\theta} h_j(x)\phi(x)\,\mathrm{d}x \\
&= \frac{1}{\sqrt{j!}} \int_{-\theta}^{\theta} -\frac{\mathrm{d}}{\mathrm{d}x}\left[h_{j-1}\phi(x)\right]\phi(x)\,\mathrm{d}x = \frac{1}{\sqrt{j!}}\Big\{-h_{j-1}(x)\phi(x)\Big\}\Big|_{-\theta}^{\theta} \\
&= \frac{1}{\sqrt{j!}}\left(h_{j-1}(-\theta) - h_{j-1}(\theta)\right)\phi(\theta).
\end{aligned}$$

The third equality follows from the identity $h_j(x)\phi(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\left[h_{j-1}\phi(x)\right]$ for $j \geq 1$, which follows from the definition in (3). The last equality uses that $\phi(x)$ is an even function.

Furthermore, if j is odd, then $h_{j-1}(-\theta) = h_{j-1}(\theta)$, and hence $\widetilde{f_\theta}(j) = 0$. If j is even and $j \geq 2$, then $h_{j-1}(-\theta) = -h_{j-1}(\theta)$, and hence $\widetilde{f_\theta}(j) = -\frac{2}{\sqrt{j!}} h_{j-1}(\theta)\phi(\theta)$. ■

We bound the even-degree Hermite coefficients of the interval function by bounding each univariate Hermite polynomial, which provides the following coefficient bound.

Lemma 28 *For any even $j \geq 2$ and any $\theta \geq 0$,*

$$\widetilde{f_\theta}(j)^2 = \frac{4}{j!} h_{j-1}(\theta)^2 \phi(\theta)^2 \leq \left(1 + \theta\sqrt{\frac{e}{j}}\right)^{2(j-1)} e^{-\theta^2}.$$

Proof. The equality is by Lemma 27. For the inequality, we define the following values:

$$A_{j,\theta} := \frac{1}{\sqrt{j!}} \left|h_{j-1}(\theta)\right|, \qquad B_{j,\theta} := \sqrt[4]{\frac{2e^2}{\pi j^3}} \left(1 + \theta\sqrt{\frac{e}{j}}\right)^{j-1}.$$

We show that $A_{j,\theta} \leq B_{j,\theta}$. Since $\widetilde{f_\theta}(j)^2 = (2/\pi) A_{j,\theta}^2 \cdot e^{-\theta^2}$, this inequality implies that $\widetilde{f_\theta}(j)^2$ is at most $(2/\pi) B_{j,\theta}^2 \cdot e^{-\theta^2}$, which is easily verified to be at most the claimed upper bound in the statement of Lemma 28.

We expand $A_{j,\theta}$ using an explicit formula for the Hermite polynomial (Abramowitz and Stegun, 1972, Equation 22.3.11), followed by a change of variable:

$$\begin{aligned}
A_{j,\theta} &= \frac{1}{\sqrt{j!}} \left|h_{j-1}(\theta)\right| \\
&= \frac{(j-1)!}{\sqrt{j!}} \left| \sum_{m=0}^{j/2-1} \frac{(-1)^m \theta^{j-1-2m}}{2^m m!(j-1-2m)!} \right| && \text{(explicit formula for } h_{j-1}(\theta)) \\
&= \frac{(j-1)!}{\sqrt{j!}} \left| \sum_{\text{odd } \ell=1}^{j-1} \frac{(-1)^{\frac{j-1-\ell}{2}} \theta^\ell}{2^{\frac{j-1-\ell}{2}} \left(\frac{j-1-\ell}{2}\right)! \ell!} \right|. && \text{(change of variable)}
\end{aligned}$$

Thus, by the triangle inequality,

$$A_{j,\theta} \leq \sum_{\text{odd } \ell=1}^{j-1} \frac{(j-1)!}{\sqrt{j!}} \cdot \frac{\theta^\ell}{2^{\frac{j-1-\ell}{2}} \left(\frac{j-1-\ell}{2}\right)! \ell!} = \sum_{\text{odd } \ell=1}^{j-1} \frac{\sqrt{2}}{2^{j/2}} \binom{j-1}{\ell} \frac{(j-1-\ell)!}{\sqrt{j!}\left(\frac{j-1-\ell}{2}\right)!} (\sqrt{2}\theta)^\ell. \quad (5)$$

We employ Stirling's approximation $\sqrt{2\pi n}(n/e)^n e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n}(n/e)^n e^{1/(12n)}$ to bound each term in the sum from (5). For any odd $\ell \in [1, j-3]$:

$$\begin{aligned}
\frac{\sqrt{2}}{2^{j/2}} \binom{j-1}{\ell} \frac{(j-1-\ell)!}{\sqrt{j!}\left(\frac{j-1-\ell}{2}\right)!} (\sqrt{2}\theta)^\ell &\leq \frac{\sqrt{2}}{2^{j/2}} \binom{j-1}{\ell} \sqrt[4]{\frac{2}{\pi j}} \left(\sqrt{\frac{e}{j}}\right)^j \left(\frac{2(j-1-\ell)}{e}\right)^{\frac{j-1-\ell}{2}} (\sqrt{2}\theta)^\ell \\
&= \binom{j-1}{\ell} \sqrt[4]{\frac{2e^2}{\pi j^3}} \left(\theta\sqrt{\frac{e}{j}}\right)^\ell \left(1 - \frac{1+\ell}{j}\right)^{\frac{j-1-\ell}{2}} \\
&\leq \sqrt[4]{\frac{2e^2}{\pi j^3}} \binom{j-1}{\ell} \left(\theta\sqrt{\frac{e}{j}}\right)^\ell.
\end{aligned}$$

We handle the final term, $\ell = j-1$, separately:

$$\frac{\sqrt{2}}{2^{j/2}}\binom{j-1}{\ell}\frac{(j-1-\ell)!}{\sqrt{j!}\left(\frac{j-1-\ell}{2}\right)!}(\sqrt{2}\theta)^{\ell} = \frac{\sqrt{2}}{2^{j/2}}\frac{1}{\sqrt{j!}}(\sqrt{2}\theta)^{j-1} \le \frac{\sqrt{2}}{2^{j/2}}\frac{1}{(2\pi j)^{1/4}}\left(\sqrt{\frac{e}{j}}\right)^{j}(\sqrt{2}\theta)^{j-1}$$
$$= \frac{1}{(2\pi j)^{1/4}}\sqrt{\frac{e}{j}}\left(\theta\sqrt{\frac{e}{j}}\right)^{j-1} \le \sqrt[4]{\frac{2e^2}{\pi j^3}}\binom{j-1}{\ell}\left(\theta\sqrt{\frac{e}{j}}\right)^{\ell}.$$

Therefore, we upper-bound the summation from (5) term-by-term, and then further simplify bound by including additional non-negative terms in the summation:

$$A_{j,\theta} \le \sqrt[4]{\frac{2e^2}{\pi j^3}}\sum_{\text{odd } \ell=1}^{j-1}\binom{j-1}{\ell}\left(\theta\sqrt{\frac{e}{j}}\right)^{\ell}$$
$$\le \sqrt[4]{\frac{2e^2}{\pi j^3}}\sum_{\ell=0}^{j-1}\binom{j-1}{\ell}\left(\theta\sqrt{\frac{e}{j}}\right)^{\ell}$$
$$= \sqrt[4]{\frac{2e^2}{\pi j^3}}\left(1+\theta\sqrt{\frac{e}{j}}\right)^{j-1} = B_{j,\theta}. \qquad \blacksquare$$

Lemma 29 *For sufficiently large k, $\sqrt{2\ln k - \ln(2\ln k)} \le \theta_k \le \sqrt{2\ln k}$.*

Proof. Recall that θ_k is defined so that $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(0,I_k)}[\mathsf{Cube}_k(\mathbf{x})] = 0$. In other words, it is the median value of $\mathbf{y} := \max_{i\in[k]}|\mathbf{x}_i|$, where $(\mathbf{x}_1,\dots,\mathbf{x}_k) \sim \mathcal{N}(0,1)^{\otimes k}$. Therefore, it suffices to show that for $l_k := \sqrt{2\ln k - \ln(2\ln k)}$ and $u_k := \sqrt{2\ln k}$, we have $\Pr[\mathbf{y} < l_k] \le 1/2 \le \Pr[\mathbf{y} < u_k]$. Note that for any $t \ge 0$, $\Pr[\mathbf{y} < t] = (1 - \Pr_{\mathbf{x}_1\sim\mathcal{N}(0,1)}[|\mathbf{x}_1| \ge t])^k$. Using the Mills ratio bound (see, e.g., Feller, 1968, Lemma 2 on page 175) and $1 - x \le e^{-x}$ for all $x \in \mathbb{R}$,

$$\Pr[\mathbf{y} < l_k] \le \left(1 - \left(\frac{1}{l_k} - \frac{1}{l_k^3}\right)\sqrt{\frac{2}{\pi}}e^{-l_k^2/2}\right)^k$$
$$= \left(1 - \frac{1}{\sqrt{2\ln k - \ln(2\ln k)}}(1-o(1))\sqrt{\frac{2}{\pi}}\cdot\frac{\sqrt{2\ln k}}{k}\right)^k$$
$$\le \exp\left(-(1-o(1))\sqrt{\frac{2}{\pi}}\right) \le \frac{1}{2}$$

by the choice of l_k and assumption that k is sufficiently large. Similarly (but now using $1 - x \ge e^{-x/(1-x)}$ for $x < 1$),

$$\Pr[\mathbf{y} < u_k] \ge \left(1 - \frac{1}{u_k}\sqrt{\frac{2}{\pi}}e^{-u_k^2/2}\right)^k$$
$$= \left(1 - \frac{1}{\sqrt{\pi\ln k}}\cdot\frac{1}{k}\right)^k$$
$$\ge \exp\left(-(1+o(1))\frac{1}{\sqrt{\pi\ln k}}\right) \ge \frac{1}{2}. \qquad \blacksquare$$

C.2. Properties of $\mathrm{TruncHigh}_{d,\tau}$ for sufficiently large τ (Proof of Proposition 13)

We recall Proposition 13.

Proposition 13 *For any $k \geq 1$ and $d \geq 2$, fix some $a > 1$ and $\rho \geq \|\mathrm{Low}_d[f]\|_2$ and let*

$$\tau := \rho \left(4e \ln(3k) + \frac{8e}{d} \ln\left(\frac{a \|f\|_2}{\rho}\right)\right)^{d/2}. \tag{1}$$

Then, (i) $\|\mathrm{Low}_d[\mathrm{TruncHigh}_{d,\tau}[f]]\|_2 \leq \frac{\rho}{a}$, and (ii) $\|\mathrm{TruncHigh}_{d,\tau}[f] - f\|_1 \leq 2\rho$.

Proof of Proposition 13, part (i). We first bound the low-degree Hermite coefficients of $\mathrm{TruncHigh}_{d,\tau}(f)$. Fix some J with $|J| \leq d$. Then

$$\begin{aligned}
\left|\widetilde{\mathrm{TruncHigh}_{d,\tau}}[f](J)\right| &\leq \left|\widetilde{\mathrm{High}_d[f]}(J)\right| + \frac{1}{\sqrt{J!}} \left|\mathbb{E}\left[\mathrm{High}_d[f](\mathbf{x}) \mathbb{1}\left\{|\mathrm{Low}_d[f](\mathbf{x})| > \tau\right\} H_J(\mathbf{x})\right]\right| \\
&\leq \frac{1}{\sqrt{J!}} \|\mathrm{High}_d[f]\|_2 \sqrt{\mathbb{E}\left[\mathbb{1}\left\{|\mathrm{Low}_d[f](\mathbf{x}) > \tau|\right\} H_J(\mathbf{x})^2\right]} \\
&\leq \frac{1}{\sqrt{J!}} \|f\|_2 \Pr\left[\mathrm{Low}_d[f](\mathbf{x}) > \tau\right]^{1/4} \|H_J\|_4 \\
&\leq \frac{1}{\sqrt{J!}} \|f\|_2 \exp\left(-\frac{d}{8e}\left(\frac{\tau}{\|\mathrm{Low}_d[f]\|_2}\right)^{2/d}\right) 3^{d/2}\sqrt{J!} \\
&\leq \|f\|_2 \exp\left(-\frac{d}{8e}\left(\frac{\tau}{\rho}\right)^{2/d}\right) 3^{d/2}.
\end{aligned}$$

The first inequality follows from the linearity of the Hermite expansion and a triangle inequality. The second follows by Cauchy-Schwarz and the definition of $\mathrm{High}_d(f)$. The third follows from $\|\mathrm{High}_d[f]\|_2 \leq \|f\|_2$ and another application of Cauchy-Schwarz. The fourth uses Fact 25 and Fact 26 (note that (1) gives $\tau/\|\mathrm{Low}_d[f]\|_2 \geq e^d$, so Fact 26 can indeed be applied).

Now we consider the full Hermite expansion of $\mathrm{Low}_d(\mathrm{TruncHigh}_{d,\tau}(f))$ and plug in τ to retrieve the claim:

$$\begin{aligned}
\left\|\mathrm{Low}_d[\mathrm{TruncHigh}_{d,\tau}[f]]\right\|_2^2 &= \sum_{|J|\leq d} \widetilde{\mathrm{TruncHigh}_{d,\tau}}[f](J)^2 \leq k^d \|f\|_2^2 \exp\left(-\frac{d}{4e}\left(\frac{\tau}{\rho}\right)^{2/d}\right) 3^d \\
&\leq (3k)^d \|f\|_2^2 \exp\left(-d\ln(3k) - \ln\left(\frac{a^2\|f\|_2^2}{\rho^2}\right)\right) = \frac{\rho^2}{a^2}.
\end{aligned}$$

In the first inequality, we used the fact that the number of k-dimensional multi-indices J with $|J| \leq d$ is at most k^d for $d \geq 2$. ■

Proof of Proposition 13, part (ii). We have

$$\begin{aligned}
\left\|\mathrm{TruncHigh}_{d,\tau}[f] - f\right\|_1 &\leq \|f - \mathrm{High}_d[f]\|_1 + \|\mathrm{High}_d[f]\mathbb{1}\left\{|\mathrm{Low}_d[f]| > \tau\right\}\|_1 \\
&\leq \|\mathrm{Low}_d[f]\|_1 + \|\mathrm{High}_d[f]\|_2 \sqrt{\Pr\left[|\mathrm{Low}_d[f]| > \tau\right]} \\
&\leq \|\mathrm{Low}_d[f]\|_2 + \|f\|_2 \sqrt{\Pr\left[|\mathrm{Low}_d[f]| > \tau\right]} \\
&\leq \|\mathrm{Low}_d[f]\|_2 + \|f\|_2 \Pr\left[|\mathrm{Low}_d[f]| > \tau\right]^{1/4},
\end{aligned}$$

where the first inequality is by the triangle inequality and the definition of $\mathrm{TruncHigh}_{d,\tau}$, the second is Cauchy-Schwarz, and the third is monotonicity of norms and $\|\mathrm{High}_d[f]\|_2 \le \|f\|_2$. We once again use Fact 26 and τ to obtain

$$\begin{aligned}\left\|\mathrm{TruncHigh}_{d,\tau}[f] - f\right\|_1 &\le \rho + \|f\|_2 \exp\left(-\frac{d}{8e}\left(4e\log(3k) + \frac{8e}{d}\log\left(\frac{a\|f\|_2}{\rho}\right)\right)\right)\\ &\le \rho + \frac{\rho}{a} \le 2\rho.\end{aligned}$$

■

C.3. Proof of exponential decay of τ_i for Lemma 10

Fact 30 *For any fixed $i \ge 1$ and*

$$\tau_i := \frac{\|\mathrm{Low}_d[f_0]\|_2}{4^{(i-1)d}}\left(4e\ln(3k) + \frac{8e}{d}\ln\left(\frac{4^{id}\|f_{i-1}\|_2}{\|\mathrm{Low}_d[f_0]\|_2}\right)\right)^{d/2}$$

from (2), if $\|f_{i-1}\|_\infty \le \frac{4}{3}$, then $\tau_i \le \frac{\alpha}{3\cdot 2^i}$ for $\alpha = \|\mathrm{Low}_d[f_0]\|_2^{0.996}(72\ln k)^{d/2}$. In addition, $\tau_1 \ge \|\mathrm{Low}_d[f_0]\|_2$.

Proof. We first consider the case where $i = 1$. A sufficiently large choice of k yields the following:

$$\begin{aligned}\tau_1 &= \|\mathrm{Low}_d[f_0]\|_2\left(4e\ln 3k + 8e\ln 4 + \frac{8e}{d}\ln\left(\frac{1}{\|\mathrm{Low}_d[f_0]\|_2}\right)\right)^{d/2} && [\|f_0\|_2 = 1]\\ &\le \|\mathrm{Low}_d[f_0]\|_2\left(11\ln k + \frac{1000e}{\|\mathrm{Low}_d[f_0]\|_2^{1/125d}}\right)^{d/2} && [4e\ln 3k + 8e\ln 4 \le 11\ln k;\ \ln x \le x]\\ &\le \|\mathrm{Low}_d[f_0]\|_2\left(\frac{12}{\|\mathrm{Low}_d[f_0]\|_2^{1/125d}}\ln k\right)^{d/2} && [\forall x \ge 1,\ 11\ln k + 1000ex \le 12x\ln k]\\ &= \|\mathrm{Low}_d[f_0]\|_2^{0.996}(12\ln k)^{d/2} \le \frac{\alpha}{6}.\end{aligned}$$

Observe that $\tau_1 \ge \|\mathrm{Low}_d[f_0]\|_2$ for sufficiently large k, because the base of the exponent will always be at least 1.

For fixed $i \ge 2$, we prove $\tau_i \le \frac{\alpha}{e\cdot 2^i}$ by bounding $\frac{\tau_i}{\tau_1}$. Using the assumption that $\|f_{i-1}\|_2 \le \|f_{i-1}\|_\infty \le \frac{4}{3}$,

$$\begin{aligned}\frac{\tau_i}{\tau_1} &= \frac{1}{4^{(i-1)d}}\left(\frac{4e\ln(3k) + \frac{8e}{d}\ln(4^{id}\|f_{i-1}\|_2) - \ln\|\mathrm{Low}_d[f_0]\|_2}{4e\ln(3k) + \frac{8e}{d}\ln(4^d\|f_0\|_2) - \ln\|\mathrm{Low}_d[f_0]\|_2}\right)^{d/2}\\ &\le \frac{1}{4^{(i-1)d}}\left(\frac{\ln 4^{(i+1)d}}{\ln 4^d}\right)^{d/2} \le \left(\frac{i+1}{16^{i-1}}\right)^{d/2} \le \frac{1}{4^{id/2}} \le \frac{1}{2^{i-1}}.\end{aligned}$$

■

C.4. Proof of convergence of f_i's in Lemma 10

The proof of Lemma 10 constructs a sequence of functions $f_0, f_1, \cdots \in L^2(\mathcal{N}(0, I_k))$ with the following properties for any a, b, and c having $b \ge 4$ and $c \le 2$:

1. For all i, $\|f_{i+1} - f_i\|_1 \le \frac{a}{b^i}$.

2. For all i, $\|f_i\|_\infty \le c$.
3. $\lim_{i\to\infty} \|\mathrm{Low}_d(f_i)\|_2 = 0$.

We now prove that such a sequence has a limit in $L^2(\mathcal{N}(0, I_k))$ with the desired properties, as given in the following proposition.

Proposition 31 *For the sequence described above, there exists some $f^* \in L^2(\mathcal{N}(0, I_k))$ such that $\mathrm{Low}_d(f^*) = 0$, $\|f^*\|_\infty \le c$, and $\|f^* - f_i\|_1 \le \frac{2a}{b^i}$ for all i.*

Towards the proof of Proposition 31, we first show that properties (1) and (2) imply an additional property about L^2 distances between iterates.

Lemma 32 *For all i, $\|f_{i+1} - f_i\|_2 \le \sqrt{\frac{2ac}{b^i}}$.*

Proof. By the triangle inequality we have $\|f_{i+1} - f_i\|_\infty \le 2c$, and from this the bound is immediate from Holder's inequality:

$$\|f_{i+1} - f_i\|_2 \le \sqrt{\|f_{i+1} - f_i\|_\infty \|f_{i+1} - f_i\|_1} \le \sqrt{2c \cdot \frac{a}{b^i}}. \qquad \blacksquare$$

The following is immediate from Lemma 32 and the fact that $L^2(\mathcal{N}(0, I_k))$ is complete (because it is a Hilbert space).

Corollary 33 *The sequence $f_0, f_1, \ldots$ is a Cauchy sequence in $L^2(\mathcal{N}(0, I_k))$ and converges to some $f^* \in L^2(\mathcal{N}(0, I_k))$.*

Before completing the proof of Proposition 31, we recall the following topological fact concerning functional spaces L^2 and L^∞.

Lemma 34 *For any probability measure μ on $\mathbb{R}^k$ and any $\alpha > 0$,*

$$I_\alpha := \left\{ f \in L^2(\mu) : \|f\|_\infty \le \alpha \right\} \text{ is a closed set in } L^2(\mu).$$

Proof. Consider any functional sequence $(f_n)_{n\in\mathbb{N}}$ in I_α such that $f_n \to f$ in $L^2(\mu)$ as $n \to \infty$. It is clear that the limit f belongs to $L^2(\mu)$ since $L^2(\mu)$ is, by itself, closed. Thus, it suffices to prove that $\Pr_{\mathbf{x}\sim\mu}[|f(x)| \le \alpha] = 1$. Fix any $\varepsilon > 0$ and $n \in \mathbb{N}$.

$$\begin{aligned}
\Pr_{\mathbf{x}\sim\mu}[|f(\mathbf{x})| > \alpha + \varepsilon] &= \Pr[|f(\mathbf{x})| > \alpha + \varepsilon \ \wedge \ |f_n(\mathbf{x})| \le \alpha] \\
&\le \Pr[|f(\mathbf{x}) - f_n(\mathbf{x})| > \varepsilon] \\
&\le \frac{1}{\varepsilon^2} \int_{\mathbb{R}^k} |f(x) - f_n(x)|^2 \mathrm{d}\mu(x) = \frac{\|f - f_n\|_2^2}{\varepsilon^2}.
\end{aligned}$$

The final step follows from Chebyshev's inequality. By assumption, we have $\|f - f_n\|_2 \to 0$ as $n \to \infty$. For every $\varepsilon > 0$, we have $\Pr[|f(\mathbf{x})| > \alpha + \varepsilon] = 0$. Hence, $\|f\|_\infty \le \alpha$ and $f \in I_\alpha$. $\blacksquare$

Lemma 35 *f^* satisfies the properties given in Proposition 31.*

Proof. Let $B_i = \{g \in L^2(\mathcal{N}(0, I_k)) : \|g - f_i\|_2 \leq 2\sqrt{\frac{2ac}{b^i}}\}$ be the closed set containing all functions in a small L^2-ball around the i-th iterate. Note that $B_{i+1} \subset B_i$ for all $i \geq 0$ and that $\bigcap_{i \geq 0} B_i = \{f^*\}$. We prove each property of Proposition 31.

1. Suppose that $\|\mathrm{Low}_d(f^*)\|_2 \geq \epsilon$ for any fixed $\epsilon > 0$. For any sufficiently large i,

$$\|f^* - f_i\|_2 \geq \|\mathrm{Low}_d(f^*) - \mathrm{Low}_d(f_i)\|_2 \geq \|\mathrm{Low}_d(f^*)\|_2 - \|\mathrm{Low}_d(f_i)\|_2 \geq \epsilon - \frac{\epsilon}{2} = \frac{\epsilon}{2}.$$

   This would mean that exists some i' such that $\|f^* - f_{i'}\|_2 \geq 2\sqrt{\frac{2ac}{b^{i'}}}$, but then f^* would lie outside $B_{i'}$, which is a contradiction.

2. Let $I = \{g \in L^2(\mathcal{N}(0, I_k)) : \|g\|_\infty \leq c\}$. By Lemma 34 (with $\mu = \mathcal{N}(0, I_k)$), I is closed in $L^2(\mathcal{N}(0, I_k))$ and $f_0, f_1, \ldots$ is a sequence in I with limit $f^* \in L^2(\mathcal{N}(0, I_k))$, we must have that $f^* \in I$ as well. Thus, $\|f^*\|_\infty \leq c$.

3. Fix any $i \geq 0$. Choose some $i' > i$ such that $b^{i'} \geq \frac{18b^{2i}c}{a}$. Because $f^* \in B_{i'}$, it follows that $\|f_{i'} - f^*\|_1 \leq \|f_{i'} - f^*\|_2 \leq 2\sqrt{\frac{2ac}{b^{i'}}} \leq \frac{2a}{3b^i}$. Thus,

$$\|f^* - f_i\|_1 \leq \|f_{i'} - f_i\|_1 + \|f^* - f_{i'}\|_1 \leq \sum_{\iota=i}^{i'-1} \frac{a}{b^\iota} + \frac{2a}{3b^i} \leq \frac{a}{b^i} \sum_{\iota=0}^{\infty} \frac{1}{4^\iota} + \frac{2a}{3b^i} = \frac{2a}{b^i}. \qquad \blacksquare$$

Appendix D. Supporting proof for Section 4

D.1. Existence of a hard-to-weak-learn intersection of halfspaces (Proof of Corollary 16)

Corollary 16 *For sufficiently large n, for all $q \geq m$, there is a distribution $\mathcal{D}$ over intersections of q^{101} halfspaces such that for a target function $\boldsymbol{f} \sim \mathcal{D}$, any MQ algorithm $\mathcal{A}$ making at most q queries to $\boldsymbol{f}$ has expected error at least $\frac{1}{2} - \frac{O(\log q)}{\sqrt{m}}$ (where the expectation is over $\boldsymbol{f} \sim \mathcal{D}$ and any internal randomness of $\mathcal{A}$, and the the accuracy is with respect to $\mathcal{N}(0, I_n)$).*

Proof. In the proof of Lemma 15, $\mathcal{D}_{\text{actual}}$ is a distribution which is supported on intersections of finitely many halfspaces. In more detail, for $\Lambda = q^{100} \ln 2$ and some $M \gg \Lambda$ (the exact value is not important for our purposes), a draw of $\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}$ is defined in the proof of Theorem 2 of De and Servedio (2021) to be an intersection of $H_{\boldsymbol{f}} \leq M$ halfspaces from a fixed collection $\{h_1, \ldots, h_M\}$, where each halfspace h_i is independently included in the intersection with probability $\frac{\Lambda}{M}$. Note that the expected number of halfspaces included in $\boldsymbol{f}$ is $\mathbb{E}[H_{\boldsymbol{f}}] = \Lambda$.

We define $\mathcal{D}$ to be the conditional distribution of $\mathcal{D}_{\text{actual}}$ conditioned on $\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}$ being an intersection of at most q^{101} halfspaces. By Markov's inequality, we have that $H_{\boldsymbol{f}} \leq q^{101} \ln 2 \leq q^{101}$ with probability at least $1 - \frac{1}{q}$. We bound the expected accuracy of the classifer h returned by $\mathcal{A}$ for random $\boldsymbol{f} \sim \mathcal{D}$ by comparing it to the expected error of a random $\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}$:

$$\begin{aligned}
\Pr_{\boldsymbol{f} \sim \mathcal{D}, \mathcal{A}, \mathbf{x}}[h(\mathbf{x}) \neq \boldsymbol{f}(\mathbf{x})] &= \Pr_{\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}, \mathcal{A}, \mathbf{x}}\left[h(\mathbf{x}) \neq \boldsymbol{f}(\mathbf{x}) \mid H_{\boldsymbol{f}} \leq q^{101}\right] \\
&\geq \Pr_{\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}, \mathcal{A}, \mathbf{x}}\left[h(\mathbf{x}) \neq \boldsymbol{f}(\mathbf{x}), H_{\boldsymbol{f}} \leq q^{101}\right] \\
&\geq \Pr_{\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}, \mathcal{A}, \mathbf{x}}[h(\mathbf{x}) \neq \boldsymbol{f}(\mathbf{x})] - \Pr_{\boldsymbol{f} \sim \mathcal{D}_{\text{actual}}, \mathcal{A}, \mathbf{x}}\left[H_{\boldsymbol{f}} > q^{101}\right] \\
&\geq \frac{1}{2} - \frac{O(\log q)}{\sqrt{m}} - \frac{1}{q} \geq \frac{1}{2} - \frac{O(\log q)}{\sqrt{m}},
\end{aligned}$$

where in the last line we used that $\frac{1}{q} = \frac{O(\log q)}{\sqrt{m}}$ (with room to spare) since $q \geq m$. ■

Appendix E. Supporting lemmas and proofs for Section 5

In what follows we describe our approach to strengthen the SQ lower bounds from Section 4; the lower bounds from Section 3 can be similarly strengthened in an entirely analogous fashion. Recall the intersection $k+1$ halfspaces over $\mathbb{R}^{m+1}$ obtained by taking the intersection of

- the k halfspaces identified in Lemma 17 over $\mathbb{R}^m$; and
- an origin-centered halfspace orthogonal to the $(m+1)$-st coordinate basis vector.

This intersection of $k+1$ halfspaces $f : \mathbb{R}^{m+1} \to \{-1,1\}$ can be written as $f(x_1,\ldots,x_{m+1}) = f_1(x_1,\ldots,x_m) \wedge f_2(x_{m+1})$, where $f_1 : \mathbb{R}^m \to \{-1,1\}$ is the intersection of k halfspaces given in Lemma 17, $f_2 : \mathbb{R} \to \{-1,1\}$ is the sign$(\cdot)$ function which outputs 1 on an input z iff $z > 0$, and the "$\wedge$" of two values from $\{-1,1\}$ is 1 iff both of them are 1.

The following lemma gives a lower bound on the approximate degree of f in terms of the approximate degrees of f_1 and f_2.

Lemma 36 *Let $\mathbf{z}_1$ and $\mathbf{z}_2$ be independent random variables in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively. Fix any $\epsilon > 0$ and $g_i \colon \mathbb{R}^{n_i} \to \{-1,1\}$ for $i \in \{1,2\}$ with $c := \min\{\Pr_{\mathbf{z}_1}[g_1(\mathbf{z}_1) = 1], \Pr_{\mathbf{z}_2}[g_2(\mathbf{z}_2) = 1]\} > 0$. Then the function $g \colon \mathbb{R}^{n_1+n_2} \to \{-1,1\}$ defined by $g(z_1,z_2) := g_1(z_1) \wedge g_2(z_2)$ has L^1 $(c\epsilon)$-approximate degree at least $\max\{d_1,d_2\}$ (with respect to the joint distribution of $(\mathbf{z}_1,\mathbf{z}_2)$), where d_i is the L^1 ϵ-approximate degree of g_i (with respect to the marginal distribution of $\mathbf{z}_i$).*

Proof. Assume without loss of generality that $d_1 \geq d_2$. For a $\{-1,1\}$-valued function h, let $h' = \frac{h+1}{2}$ (so h' is the $\{0,1\}$-valued version of h). We observe that $c = \min\{\mathbb{E}_{\mathbf{z}_1}[g_1'(\mathbf{z}_1)], \mathbb{E}_{\mathbf{z}_2}[g_2'(\mathbf{z}_2)]\}$, and that d_i is the L^1 $(\epsilon/2)$-approximate degree of g_i'.

Let d be the L^1 $(c\epsilon/2)$-approximate degree of g', and let p' be a degree-d polynomial over $\mathbb{R}^{n_1+n_2}$ satisfying $\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[|g'(\mathbf{z}_1,\mathbf{z}_2) - p'(\mathbf{z}_1,\mathbf{z}_2)|] \leq c\epsilon/2$. For this polynomial p,

$$\min_{z_2 \in \mathbb{R}^{n_2} : g_2'(z_2)=1} \mathbb{E}_{\mathbf{z}_1}[|g'(\mathbf{z}_1,z_2) - p'(\mathbf{z}_1,z_2)|] \leq \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}\left[|g'(\mathbf{z}_1,\mathbf{z}_2) - p'(\mathbf{z}_1,\mathbf{z}_2)| \cdot \frac{g_2'(\mathbf{z}_2)}{\mathbb{E}_{\mathbf{z}_2}[g_2'(\mathbf{z}_2)]}\right]$$
$$\leq \frac{\mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}[|g'(\mathbf{z}_1,\mathbf{z}_2) - p'(\mathbf{z}_1,\mathbf{z}_2)|]}{c} \leq \frac{\epsilon}{2}.$$

So there exists $z_2 \in \mathbb{R}^{n_2}$ such that

$$\mathbb{E}_{\mathbf{z}_1}[|g'(\mathbf{z}_1,z_2) - p'(\mathbf{z}_1,z_2)|] = \mathbb{E}_{\mathbf{z}_1}[|g_1'(\mathbf{z}_1) - p'(\mathbf{z}_1,z_2)|] \leq \frac{\epsilon}{2}.$$

Letting $p = 2p' - 1$, since $g = 2g' - 1$, there exists $z_2 \in \mathbb{R}^{n_2}$ such that

$$\mathbb{E}_{\mathbf{z}_1}[|g(\mathbf{z}_1,z_2) - p(\mathbf{z}_1,z_2)|] = 2\,\mathbb{E}_{\mathbf{z}_1}[|g_1'(\mathbf{z}_1) - p'(\mathbf{z}_1,z_2)|] \leq \epsilon.$$

Since $p(\cdot,z_2)$ is a polynomial over $\mathbb{R}^{n_1}$ of degree at most d, it follows that the L^1 $(\epsilon/2)$-approximate degree of g_1' is at most d. Hence $d \geq d_1 = \max\{d_1,d_2\}$. Since the L^1 $(c\epsilon/2)$-approximate degree of g' (which is d) is the same as the L^1 $(c\epsilon)$-approximate degree of g, the lemma is proved. ■

By Lemma 17, the L^1 $\frac{1}{2}$-approximate degree of f_1 is at least $\Omega(\log(k)/\log\log k)$, and hence so is its L^1 (4ϵ)-approximate degree (for $\epsilon \leq 1/8$). A lower bound on the L^1 (4ϵ)-approximate degree of f_2 is given by the following result of Ganzburg (2002).

Lemma 37 *For any $\epsilon > 0$, the L^1 ϵ-approximate degree of the sign$(\cdot)$ function is $\Omega(1/\epsilon^2)$.*

Lemma 37 (presented as Corollary B.1 of Diakonikolas et al. (2021b)) is a direct consequence of Theorem 1 of Ganzburg (2002) and Theorem 4 of Vaaler (1985).

We are now almost ready to apply Lemma 36 to our intersection of $k+1$ halfspaces obtained via f_1 (from Lemma 17) and f_2 (the sign function). We just need to ensure that each of f_1 and f_2 takes value $+1$ with sufficient probability. First, observe that f_1 satisfies $\Pr_{\mathbf{x}\sim\mathcal{N}(0,I_m)}[f_1(\mathbf{x}) = 1] \geq 1/4$, since otherwise the $1/2$-approximate degree of f would be zero, as witnessed by the constant -1 function. Moreover, $\Pr_{\mathbf{x}\sim\mathcal{N}(0,1)}[f_2(\mathbf{x}) = 1] = 1/2$ by symmetry of $\mathcal{N}(0,1)$. So, we have established that both f_1 and f_2 take value $+1$ with probability at least $1/4$, and that also they have (4ϵ)-approximate degrees $\Omega(\log(k)/\log\log k)$ and $\Omega(1/\epsilon^2)$, respectively. Therefore, Lemma 36 implies a lower bound on the L^1 ϵ-approximate degree of f, as stated in the following lemma.

Lemma 38 *For any $k = 2^{O(n^{0.245})}$ and any $\epsilon > 0$, there is an intersection of $k+1$ halfspaces $f\colon \mathbb{R}^{m+1} \to \{-1,1\}$ with L^1 ϵ-approximate degree $\Omega(\frac{\log k}{\log\log k} + \frac{1}{\epsilon^2})$, where $m = O(n^{0.49})$.*

Lemma 38 and Lemma 6 together imply Theorem 18.

Theorem 18 (Formal version of Theorem 1) *For any $k = 2^{O(n^{0.245})}$ and any $\epsilon \geq n^{-c}$ for a suitably small absolute constant $c > 0$, any SQ algorithm that agnostically learns intersections of k halfspaces to excess error ϵ under Gaussian marginals requires either $2^{n^{\Omega(1)}}$ queries or at least one query of tolerance $n^{-\Omega(\log(k)/\log\log k + 1/\epsilon^2)}$.*