Proceedings of Machine Learning Research vol 125:1–26, 2020 33rd Annual Conference on Learning Theory

Algorithms and SQ Lower Bounds for PAC Learning One-Hidden-Layer ReLU Networks

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

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

Vasilis Kontonis KONTONIS@WISC.EDU
Nikos Zarifis ZARIFIS@WISC.EDU
University of Wisconsin-Madison

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We study the problem of PAC learning one-hidden-layer ReLU networks with k hidden units on $\mathbb{R}^d$ under Gaussian marginals in the presence of additive label noise. For the case of positive coefficients, we give the first polynomial-time algorithm for this learning problem for k up to $\tilde{O}(\sqrt{\log d})$. Previously, no polynomial time algorithm was known, even for $k = 3$. This answers an open question posed by Klivans (2017). Importantly, our algorithm does not require any assumptions about the rank of the weight matrix and its complexity is independent of its condition number. On the negative side, for the more general task of PAC learning one-hidden-layer ReLU networks with arbitrary real coefficients, we prove a Statistical Query lower bound of $d^{\Omega(k)}$. Thus, we provide a separation between the two classes in terms of efficient learnability. Our upper and lower bounds are general, extending to broader families of activation functions.

Keywords: PAC learning, one-hidden-layer networks, statistical query model

1. Introduction

1.1. Background and Motivation

In recent years, the impressive practical success of deep learning has motivated the development of provably efficient learning algorithms for various classes of neural networks. A large body of research (see Section 1.4 for a brief overview) has resulted in efficient learning algorithms for shallow networks with common activation functions (e.g., ReLUs or sigmoids) under various assumptions on the underlying distribution and the weight structure of the network. Despite intensive investigation, the broad question of whether deep neural networks are efficiently learnable with provable guarantees remains an outstanding theoretical challenge in machine learning. In particular, the class of networks for which efficient learners are known is relatively limited, even in the realizable case (i.e., when the data is drawn from a neural network in the class).

In this work, we continue this line of investigation by studying the learnability of a simple class of networks without imposing strong restrictions on the structure of its weights. Specifically, we focus on the problem of learning one-hidden-layer ReLU networks under the Gaussian distribution

in the presence of additive random label noise. Our goal is to understand the complexity of this problem *in the PAC learning model without assumptions on the weight matrix of the network.*

Definition 1 (One-hidden-layer ReLU networks) *Let $\mathcal{C}_k$ denote the concept class of one-hidden-layer ReLU networks on $\mathbb{R}^d$ with k hidden units. That is, $f_{\alpha,\mathbf{W}} \in \mathcal{C}_k$ if and only if there exist weight vectors $\mathbf{w}^{(i)} \in \mathbb{R}^d$ and real coefficients α_i, $i \in [k]$, such that $f_{\alpha,\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^k \alpha_i \phi(\langle \mathbf{w}^{(i)}, \mathbf{x} \rangle)$, where $\phi(t) = \max\{0, t\}$, $t \in \mathbb{R}$. We will denote by $\alpha = (\alpha_i)_{i=1}^k$ the vector of coefficients and by $\mathbf{W} = [\mathbf{w}^{(i)}]_{i=1}^k$ the weight matrix of the network. We will use $\mathcal{C}_k^+$ to denote the subclass of $\mathcal{C}_k$ where $\alpha \in \mathbb{R}_+^k$.*

The (distribution-specific) PAC learning problem for a concept class $\mathcal{C}$ of real-valued functions is the following: The input is a multiset of i.i.d. labeled examples $(\mathbf{x}, y)$, where $\mathbf{x}$ is generated from the standard Gaussian distribution on $\mathbb{R}^d$ and $y = f(\mathbf{x}) + \xi$, where $f \in \mathcal{C}$ is the unknown target concept and ξ is some type of random observation noise. The goal of the learner is to output a hypothesis $h : \mathbb{R}^d \to \mathbb{R}$ that with high probability is close to f in L_2-norm. The hypothesis h is allowed to lie in any efficiently representable hypothesis class $\mathcal{H}$. If $\mathcal{H} = \mathcal{C}$, the PAC learning algorithm is called *proper*.

Perhaps surprisingly, the complexity of PAC learning one-hidden-layer ReLU networks (even with positive weights) has remained open, even in the realizable setting, under Gaussian marginals, and for $k = 3$ Klivans (2017)[1]. A line of prior work Ge et al. (2018); Bakshi et al. (2019); Ge et al. (2019) had studied the task of *parameter estimation* for this concept class, i.e., the task of recovering the unknown coefficients α_i and weight vectors $\mathbf{w}^{(i)}$ of the data generating network within small accuracy. It should be noted that for parameter estimation to even be information-theoretically possible, some assumptions on the target function are necessary. The aforementioned prior works made the common assumption that the weight matrix $\mathbf{W} = [\mathbf{w}^{(i)}]_{i=1}^k$ is *full-rank*. Under this assumption, they provided efficient parameter learning algorithms with respect Gaussian marginals for the case of *positive coefficients*, i.e., for $\mathcal{C}_k^+$. Importantly, the sample and computational complexity of these algorithms scale polynomially with the condition number of $\mathbf{W}$. In contrast, no such algorithm is known for general coefficients, i.e., for $\mathcal{C}_k$, even under the aforementioned strong assumptions on the weights.

In contrast to parameter estimation, PAC learning one-hidden-layer ReLU networks does not require any assumptions on the structure of the weight matrix. The PAC learning problem for this class is information-theoretically solvable with polynomially many samples. The question is whether a computationally efficient algorithm exists. It should also be noted that proper PAC learning is not generally equivalent to parameter estimation, as it is in principle possible to have two networks that define close-by functions and whose parameters are significantly different.

1.2. Our Results

We are ready to describe the main contributions of this work. Our main positive result is the first PAC learning algorithm for $\mathcal{C}_k^+$ (one-hidden-layer Relu networks with positive coefficients) under Gaussian marginals that runs in polynomial time for any $k = \tilde{O}(\sqrt{\log d})$. On the lower bound side, we establish a Statistical Query (SQ) lower bound suggesting that no such algorithm is possible for

1. Formally speaking, the $k = 2$ case does not appear explicitly in the literature, but an efficient algorithm easily follows from prior work on parameter estimation (e.g., Ge et al. (2018)).

$\mathcal{C}_k$ (general coefficients) for any $k = \omega(1)$ (also under Gaussian marginals). Our SQ lower bound provides a separation between $\mathcal{C}_k^+$ and $\mathcal{C}_k$ in terms of efficient learnability.

Before we state our main theorems, we formally define the PAC learning problem.

Definition 2 (Distribution-Specific PAC Learning) *Let $\mathcal{F}$ be a concept class of real-valued functions over $\mathbb{R}^d$, $\mathcal{D}$ be a distribution on $\mathbb{R}^d$, $\mathcal{F} \in L_2(\mathcal{D}, \mathbb{R}^d)$, and $0 < \epsilon < 1$. Let f be an unknown target function in $\mathcal{F}$. A* noisy example oracle*, $\mathrm{EX}^{\mathrm{noise}}(f, \mathcal{F})$, works as follows: Each time $\mathrm{EX}^{\mathrm{noise}}(f, \mathcal{F})$ is invoked, it returns a labeled example $(\mathbf{x}, y)$, such that: (a) $\mathbf{x} \sim \mathcal{D}$, and (b) $y = f(\mathbf{x}) + \xi$, where ξ is a zero-mean and standard deviation σ subgaussian random variable that is independent of $\mathbf{x}$. A learning algorithm is given i.i.d. samples from the noisy oracle and its goal is to output a hypothesis h such that with high probability h is ϵ-close to f in L_2-norm, i.e., it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[(f(\mathbf{x}) - h(\mathbf{x}))^2] \leq \epsilon^2 \left(\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[f^2(\mathbf{x})] + \sigma^2\right)$.*

Our main positive result is the first computationally efficient PAC learning algorithm for $\mathcal{C}_k^+$.

Theorem 3 (Proper PAC Learner for $\mathcal{C}_k^+$) *There is a proper PAC learning algorithm for $\mathcal{C}_k^+$ with respect to the standard Gaussian distribution on $\mathbb{R}^d$ with the following performance guarantee: The algorithm draws $\mathrm{poly}(k/\epsilon) \cdot \tilde{O}(d)$ noisy labeled examples from an unknown target $f \in \mathcal{C}_k^+$, runs in time $\mathrm{poly}(d/\epsilon) + (k/\epsilon)^{O(k^2)}$, and outputs a hypothesis $h \in \mathcal{C}_k^+$ that with high probability is ϵ-close to f in L_2-norm.*

Theorem 3 gives the first polynomial-time PAC learning algorithm for one-hidden-layer ReLU networks under any natural distributional assumptions, answering a question posed by Klivans (2017). Our algorithm runs in polynomial time for some $k = \tilde{\Omega}(\sqrt{\log d})$. The existence of such an algorithm was previously open, even for $k = 3$.

We remark that our main algorithmic result is more general, in the sense that it immediately extends to positive coefficient one-hidden-layer networks composed of any non-negative Lipschitz activation function. See Theorem 5 for a detailed statement.

Some additional remarks are in order: As stated in Theorem 3, our learning algorithm is proper, i.e., $h \in \mathcal{C}_k^+$. An important distinguishing feature of our algorithm from prior related work is that it requires no assumptions on the weight matrix of the network, and in particular that its sample complexity is independent of its condition number. Prior work had given parameter estimation algorithms for this concept class with sample complexity (and running time) polynomial in the condition number. On the other hand, the running time of our algorithm scales with $\exp(k)$, while previous parameter estimation algorithms had $\mathrm{poly}(k)$ dependence. The existence of a $\mathrm{poly}(k)$ time PAC learning algorithm remains an outstanding open question. An additional advantage of our algorithm is that it also immediately extends to the agnostic setting and in particular is robust to a small (dimension-independent) amount of adversarial L_2-error.

The algorithm of Theorem 3 crucially uses the assumption that the coefficients of the target network are positive. A natural question is whether an algorithm with similar guarantees can be obtained for unrestricted coefficients. Perhaps surprisingly, we provide evidence that such an algorithm does not exist. Specifically, our second main result is a correlational Statistical Query (SQ) lower bound ruling out a broad family of $\mathrm{poly}(d)$-time algorithms for $\mathcal{C}_k$ for $\epsilon = \Omega(1)$, for *any* $k = \omega(1)$.

Specifically, we prove a lower bound for PAC learning $\mathcal{C}_k$ under Gaussian marginals in the correlational SQ model. A correlational SQ algorithm has query access to the target concept $f : \mathbb{R}^d \to \mathbb{R}$

via the following oracle: The oracle takes as input any bounded query function $q : \mathbb{R}^d \to [-1, 1]$ and an accuracy parameter $\tau > 0$, and outputs an estimate γ of the expectation $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x})q(\mathbf{x})]$ such that $|\gamma - \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x})q(\mathbf{x})]| \leq \tau$. We note that the correlational SQ model captures a broad family of algorithms, including first-order methods (e.g., gradient-descent), dimension-reduction, and moment-based methods. (In particular, our algorithm establishing Theorem 3 can be easily simulated in this model.) We establish the following:

Theorem 4 (Correlational SQ Lower Bound for $\mathcal{C}_k$) *Any correlational SQ learning algorithm for $\mathcal{C}_k$ under the standard Gaussian distribution on $\mathbb{R}^d$ that guarantees error $\epsilon = \Omega(1)$ requires either queries of accuracy $d^{-\Omega(k)}$ or $2^{d^{\Omega(1)}}$ many queries.*

The natural interpretation of Theorem 4 is the following: If the SQ algorithm uses statistical queries of accuracy $d^{-\Omega(k)}$, then simulating a single query with iid samples would require $d^{\Omega(k)}$ samples (hence time). Otherwise, the algorithm would require $2^{d^{\Omega(1)}}$ time (since each query requires at least one unit of time). Theorem 4, combined with our Theorem 3, provides a (super-polynomial) computational separation between the PAC learnability of $\mathcal{C}_k$ and $\mathcal{C}_k^+$ in the correlational SQ model.

We note that the statement of our general SQ lower bound (Theorem 13) is much more general than Theorem 4. Specifically, we obtain a correlational SQ lower bound for PAC learning (under Gaussian marginals) a class of functions of the form $\sigma(\sum_{i=1}^k \alpha_i \phi(\mathbf{w}^{(i)}, \mathbf{x}))$, where roughly speaking σ is any odd non-vanishing function and ϕ is not a low-degree polynomial.

1.3. Our Techniques

Here we provide an overview of our techniques in tandem with a comparison to prior work. We start with our algorithm establishing Theorem 3. Our learning algorithm for $\mathcal{C}_k^+$ employs a data-dependent dimension reduction procedure. Specifically, we give an efficient method to reduce our d-dimensional learning problem down to a k-dimensional problem, that can in turn be efficiently solved by a simple covering method.

Let $f(\mathbf{x}) = \sum_{i=1}^k \alpha_i \phi(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle)$ be the target function and observe that f depends only on the k unknown linear forms $\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle$, $i \in [k]$. If we could identify the subspace V spanned by the $\mathbf{w}^{(i)}$'s exactly, then we could also identify f by brute-force on V, noting that we only need to search a k^2-dimensional space of functions and that for any $\mathbf{x} \in \mathbb{R}^d$ it holds $f(\mathbf{x}) = f(\mathrm{proj}_V(\mathbf{x}))$. Our algorithm is based on a robust version of this idea. In particular, if we can find a subspace V' that closely approximates V, then it suffices to solve for f on V' and use this projection to obtain an approximation to f.

To find a subspace V' approximating V, we consider the matrix of degree-2 Chow parameters (second moments) of f, i.e., $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}(0,I)}[f(\mathbf{x})(\mathbf{x}\mathbf{x}^T - \mathbf{I})]$. It is not hard to see that the (normalized) second moments of f are positive in the directions along V and 0 in orthogonal directions. Thus, if we could compute the second moments exactly, we could solve for V as the span of the second moment matrix. Unfortunately, we can only approximate the true second moment matrix via samples. To deal with this approximation, we note that the true second moments will be large in the direction of $\mathbf{w}^{(i)}$ for components with large coefficients α_i and 0 in directions orthogonal to V. Using this fact, we show that if V' is the span of the k largest eigenvalues of an approximate second moment matrix (obtained via sampling), the weight vectors $\mathbf{w}^{(i)}$ corresponding to the important components of f will still be close to V'. From this point, can use a net-based argument to find a hypothesis $h \in \mathcal{C}_k^+$ with weight vectors on V' so that $f(\mathbf{x})$ is close to $h(\mathrm{proj}_{V'}(\mathbf{x}))$ in L_2-norm.

We note that the idea of using dimension-reduction to find a low-dimensional invariant subspace has been previously used in the context of PAC learning intersections of LTFs Vempala (2010); Diakonikolas et al. (2018). Our algorithm and its analysis of correctness are quite different from these prior works. We also note that Ge et al. (2018) also used information based on low-degree moments for their parameter estimation algorithm, but in a qualitatively different way. In particular, Ge et al. (2018) used tensor-decomposition techniques (based on moments of degree up to four) to uniquely identify the weight vectors, under structural assumptions on the weight matrix (full-rank and bounded condition number).

We now proceed to explain our SQ lower bound construction. As is well-known, there is a general methodology to establish such lower bounds, via an appropriate notion of SQ dimension Blum et al. (1994); Feldman et al. (2017). In our setting, to prove an SQ lower bound, it suffices to find a large collection of functions $f_1, \ldots, f_m \in \mathcal{C}_k$ with the following properties: (1) The f_i's are pairwise far away from each other, and (2) The f_i's have small pairwise correlations. The difficulty is, of course, to construct such a family. We describe our construction in the following paragraph.

First, it is not hard to see that (1) and (2) can only be simultaneously satisfied if almost all of the f_i's have nearly-matching low-degree moments. In fact, we provide a construction in which all the low-degree moments of all of the f_i's vanish. To achieve this, we build on an idea introduced in Diakonikolas et al. (2017). Roughly speaking, the idea is to define a family of functions whose interesting information is hidden in a random low-dimensional subspace, so that learning an unknown function in the family amounts to finding the hidden subspace. In more detail, we will define a function in two dimensions which has the correct moments, and then embed it in a randomly chosen subspace.

For simplicity, we explain our 2-dimensional construction for ReLU activations, even though our SQ lower bound is more general. We provide an explicit 2-dimensional construction of a mixture F of $2k$ ReLUs whose first $k-1$ moments vanish exactly. For any 2-dimensional subspace V, we can define $F_V(\mathbf{x}) = F(\mathrm{proj}_V(\mathbf{x}))$. From there, we can show that if U and V are two subspaces that are far apart — in the sense that no unit vector in U has large projection in V — then F_U and F_V will have small correlation — on the order of the k-th power of the closeness parameter between the defining subspaces. Moreover, it is not hard to show that two randomly chosen U and V are far from each other with high probability. This allows us to find an exponentially large family of F_V's that have pairwise exponentially small correlation.

1.4. Related Work

In recent years, there has been an explosion of research on provable algorithms for learning neural networks in various settings, see, e.g., Janzamin et al. (2015); Sedghi et al. (2016); Daniely et al. (2016); Zhang et al. (2016); Zhong et al. (2017); Ge et al. (2018, 2019); Bakshi et al. (2019); Goel et al. (2017); Manurangsi and Reichman (2018); Goel and Klivans (2019); Vempala and Wilmes (2019) for some works on the topic. The majority of these works focused on parameter learning, i.e., the problem of recovering the weight matrix of the data generating neural network. In contrast, the focus of this paper is on PAC learning. We also note that PAC learning of simple classes of neural networks has been studied in a number of recent works Goel et al. (2017); Manurangsi and Reichman (2018); Goel and Klivans (2019); Vempala and Wilmes (2019). However, the problem of PAC learning linear combinations of (even) 3 ReLUs under any natural distributional assumptions (and in particular under the Gaussian distribution) has remained open. At a high-level, prior works

either rely on tensor decompositions Sedghi et al. (2016); Zhong et al. (2017); Ge et al. (2018, 2019); Bakshi et al. (2019) or on kernel methods Zhang et al. (2016); Daniely et al. (2016); Goel et al. (2017); Goel and Klivans (2019). In the following paragraphs, we describe in detail the prior works more closely related to the results of this paper.

The work of Ge et al. (2018) studies the parameter learning of positive linear combinations of ReLUs under the Gaussian distribution in the presence of additive (mean zero sub-gaussian) noise. That is, they consider the same concept class and noise model as we do, but study parameter learning as opposed to PAC learning. Ge et al. (2018) show that the parameters can be approximately recovered efficiently, under the assumption that the weight matrix is full-rank with bounded condition number. The sample complexity and running time of their algorithm scales polynomially with the condition number. More recently, Bakshi et al. (2019); Ge et al. (2019) obtained efficient parameter learning algorithms for vector-valued depth-2 ReLU networks under the Gaussian distribution. Similarly, the algorithms in these works have sample complexity and running time scaling polynomially with the condition number. We note that the algorithmic results in the aforementioned works do not apply to $\mathcal{C}_k$, i.e., the class of arbitrary linear combinations of ReLUs.

Vempala and Wilmes (2019) show that gradient descent agnostically PAC learns low-degree polynomials using neural networks as the hypothesis class. Their approach has implications for (realizable) PAC learning of certain neural networks under the uniform distribution on the sphere. We note that their method implies an algorithm with sample complexity and running time exponential in $1/\epsilon$, even for a single ReLU. Goel and Klivans (2019) give an efficient PAC learning algorithm for certain 2-hidden-layer neural networks under arbitrary distributions on the unit ball. We emphasize that their algorithm does not apply for (positive) linear combinations of ReLUs. In fact, recent work has shown that the problem we solve in this paper is NP-hard under arbitrary distributions, even for $k = 2$ Goel et al. (2020b).

The SQ model was introduced by Kearns (1998) in the context of learning Boolean-valued functions as a natural restriction of the PAC model Valiant (1984). A recent line of work Feldman et al. (2013, 2015b,a); Feldman (2016) extended this framework to general search problems over distributions. One can prove unconditional lower bounds on the computational complexity of SQ algorithms via an appropriate notion of *Statistical Query dimension*. A lower bound on the SQ dimension of a learning problem provides an unconditional lower bound on the computational complexity of any SQ algorithm for the problem.

The work of Vempala and Wilmes (2019) establishes correlational SQ lower bounds for learning a class of degree-k polynomials in d variables. Shamir (2018) shows that gradient-based algorithms (a special case of correlational SQ algorithms) cannot efficient learn certain families of neural networks under well-behaved distributions (including the Gaussian distribution). We note that the lower bound constructions in these works do not imply corresponding lower bounds for one-hidden-layer ReLU networks.

Concurrent and Independent Work. Contemporaneous work Goel et al. (2020a), using a different construction, obtained super-polynomial SQ lower bounds for learning one-hidden-layer neural networks (with ReLU and other activations) under the Gaussian distribution.

2. Preliminaries

Notation. For $n \in \mathbb{Z}_+$, we denote $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$. We will use small boldface characters for vectors. For $\mathbf{x} \in \mathbb{R}^d$, and $i \in [d]$, $\mathbf{x}_i$ denotes the i-th coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 \stackrel{\text{def}}{=} (\sum_{i=1}^d \mathbf{x}_i^2)^{1/2}$ denotes the ℓ_2-norm of $\mathbf{x}$. We denote by $\|\mathbf{A}\|_2$ the spectral norm of matrix $\mathbf{A}$. We will use $\langle \mathbf{x}, \mathbf{y} \rangle$ for the inner product between $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. We will use $\mathbf{E}[X]$ for the expectation of random variable X and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. We denote by $\mathbf{Var}[X]$ its variance.

For $d \in \mathbb{N}$, we denote $\mathbb{S}^{d-1}$ the d-dimensional sphere. Denote by $\theta(\mathbf{u}, \mathbf{v})$ the angle between the vectors $\mathbf{u}, \mathbf{v}$. For a vector of weights $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^{2k}$, and matrix $\mathbf{W} \in \mathbb{R}^{k \times d}$ we denote $f_{\alpha, \mathbf{W}}(\mathbf{x}) = \alpha^T \phi(\mathbf{W}x) = \sum_{i=1}^k \alpha_i \, \phi(\langle \mathbf{w}^{(i)}, \mathbf{x} \rangle)$. Let $\mathcal{N}$ denote the standard univariate Gaussian distribution, we also denote $\mathcal{N}^2$ the two dimensional Gaussian distribution and $\mathcal{N}^d$ the d-dimensional one.

3. Efficient Learning Algorithm

In this section, we give our upper bound for the problem of learning positive linear combinations of Lipschitz activations, thereby establishing Theorem 3. We prove the following more general statement:

Theorem 5 (Learning Sums of Lipschitz Activations) *Let $f(\mathbf{x}) = \sum_{i=1}^k \alpha_i \phi\big(\langle \mathbf{w}^{(i)}, \mathbf{x} \rangle \big)$ with $\alpha_i > 0$ for all $i \in [k]$, where $\phi(t)$ is an L-Lipschitz, non-negative activation function such that $\mathbf{E}_{t \sim \mathcal{N}}[\phi(t)] \geq C$, $\mathbf{E}_{t \sim \mathcal{N}}[\phi(t)(t^2 - 1)] \geq C$, where $C > 0$ and $\mathbf{E}_{t \sim \mathcal{N}}[\phi^2(t)]$ is finite. There exists an algorithm that given $k \in \mathbb{N}, \epsilon > 0$, and sample access to a noisy set of samples from $f : \mathbb{R}^d \to \mathbb{R}_+$, draws $m = d \cdot \mathrm{poly}(k, 1/\epsilon) \cdot \mathrm{poly}(L/C)$ samples, runs in time $\mathrm{poly}(m) + \widetilde{O}((1/\epsilon)^{k^2})$, and outputs a proper hypothesis h that, with probability at least $9/10$, satisfies*

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}^d}[(f(\mathbf{x}) - h(\mathbf{x}))^2] \leq \epsilon^2 \mathrm{poly}(L/C) \left(\sigma^2 + \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}^d}[f(\mathbf{x})^2] \right) .$$

Remark 6 *Theorem 3 follows as a corollary of the above, by noting that the ReLU satisfies $L = 1$ and $C = \frac{1}{\sqrt{2\pi}}$.*

The following fact gives formulas for the low-degree Chow parameters of a one-layer network (see Appendix A).

Fact 7 (Low-degree Chow Parameters) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be of the form $f(\mathbf{x}) = \sum_{i=1}^k \alpha_i \cdot \phi\left(\langle \mathbf{w}^{(i)}, \mathbf{x} \rangle\right)$. Then $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}^d}[f(\mathbf{x})] = \mathbf{E}_{t \sim \mathcal{N}}[\phi(t)] \sum_{i=1}^k \alpha_i$, $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}^d}[f(\mathbf{x})\mathbf{x}] = \mathbf{E}_{t \sim \mathcal{N}}[\phi(t)t] \cdot \sum_{i=1}^k \alpha_i \mathbf{w}^{(i)}$, and*

$$\mathbf{A} = \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}^d}\left[f(\mathbf{x})(\mathbf{x}\mathbf{x}^T - \mathbf{I})\right] = \mathop{\mathbf{E}}_{t \sim \mathcal{N}}[\phi(t)(t^2 - 1)] \sum_{i=1}^k \alpha_i \mathbf{w}^{(i)} \mathbf{w}^{(i)^T} . \tag{1}$$

The crucial formula is the one of the degree-2 Chow parameters, Equation (1). In fact, we can already describe the main idea of our upper bound. Let us assume that we have the degree-2 Chow parameters matrix $\mathbf{A}$ *exactly*. Then, by using singular value decomposition, we would obtain a basis of the vector space spanned by the parameters $\mathbf{w}^{(i)}$. The dimension of this space is at most k and

therefore in that way we essentially reduce the dimension of the problem from d down to k. To find parameters $\hat{\alpha}_i, \hat{\mathbf{w}}^{(i)}$ that give small mean squared error, we can now make a grid $\mathcal{G}$ and pick the ones that minimize the empirical mean squared error with the samples, that is

$$\min_{\beta, \mathbf{U} \in \mathcal{G}} \sum_{i=1}^{m} (f_{\beta, \mathbf{U}}(\mathbf{x}^{(i)}) - y^{(i)})^2 \ .$$

Even though we do not have access to the matrix $\mathbf{A}$ exactly, we can estimate it empirically. Since the activation function $\phi(\cdot)$ is well-behaved and the distribution of the examples is Gaussian, we can get a very accurate estimate of $\mathbf{A}$ with roughly $\widetilde{O}(dk/\epsilon^2)$ samples. We give the following lemma whose proof relies on matrix concentration and concentration of polynomials of Gaussian random variables (see Appendix B).

Lemma 8 (Estimation of degree-2 Chow parameters) *Let $f_{\alpha, \mathbf{W}}(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i \phi(\langle \mathbf{w}^{(i)}, \mathbf{x} \rangle)$, where $\phi(t)$ is an L-Lipschitz, non-negative activation function such that $\mathbf{E}_{t \sim \mathcal{N}}[\phi(t)] \geq C$. Let $\mathbf{\Sigma} = \mathbf{E}_{\mathbf{x} \sim \mathcal{N}^d}[f_{\alpha, \mathbf{W}}(\mathbf{x}) \mathbf{x} \otimes \mathbf{x}]$ be the degree-2 Chow parameters of $f_{\alpha, \mathbf{W}}$. Then, for some $N = \widetilde{O}(dk/\epsilon^2)$ samples $(\mathbf{x}^{(i)}, y^{(i)})$, where $y^{(i)} = f_{\alpha, \mathbf{W}}(\mathbf{x}^{(i)}) + \xi_i$ and ξ_i is a zero-mean, subgaussian noise with variance σ^2, it holds with probability at least 99% that*

$$\left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)} y^{(i)} - \mathbf{\Sigma} \right\|_2 \leq \epsilon \left(\sigma + \frac{L}{C} \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}^d} [f_{\alpha, \mathbf{W}}(\mathbf{x})] \right) \ .$$

The next step is to quantify how accurately we need to estimate the degree-2 Chow parameters, so that doing SVD on the empirical matrix gives us a good approximation of the subspace spanned by the true parameters $\mathbf{w}^{(i)}$. We show that that estimating the degree-2 Chow parameter matrix within spectral norm roughly ϵ/k suffices. In particular, we show that the top-k eigenvectors of our empirical estimate span approximately the subspace where the true parameters $\mathbf{w}^{(i)}$ lie. For the proof, we are going to use the following lemma that bounds the difference of a function evaluated at correlated normal random variables.

Lemma 9 (Correlated Differences, Lemma 6 of Kontonis et al. (2019)) *Let $r(\mathbf{x}) \in L_2(\mathbb{R}^d, \mathcal{N}^d)$ be differentiable almost everywhere and let*

$$D_\rho = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{I} & \rho \mathbf{I} \\ \rho \mathbf{I} & \mathbf{I} \end{pmatrix} \right) .$$

We call ρ-correlated a pair of random variables $(\mathbf{x}, \mathbf{y}) \sim D_\rho$. It holds

$$\frac{1}{2} \mathop{\mathbf{E}}_{(\mathbf{x}, \mathbf{z}) \sim D_\rho} [(r(\mathbf{x}) - r(\mathbf{z}))^2] \leq (1 - \rho) \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{N}^d} \left[\|\nabla r(\mathbf{x})\|_2^2 \right] \ .$$

We are now ready to prove the key technical lemma of our approach. We remark that the following dimension reduction lemma is rather general and holds for any reasonable activation function, in the sense that the error is bounded as long as its expected derivative $\mathbf{E}_{t \sim \mathcal{N}}[(\phi'(t))^2]$ is bounded.

Lemma 10 (Dimension Reduction) *Let $f_{\alpha, \mathbf{W}}(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i \phi\big(\langle \mathbf{w}^{(i)}, \mathbf{x} \rangle \big)$ with $\alpha_i > 0$, let $\mathbf{A} = \mathbf{E}_{\mathbf{x} \sim \mathcal{N}^d}[f(\mathbf{x}) \mathbf{x}\mathbf{x}^T]$ and $\mathbf{E}_{t \sim \mathcal{N}}[\phi(t)(t^2 - 1)] = C_1$. Let $\mathbf{M} \in \mathbb{R}^{d \times d}$ be a matrix such that $\|\mathbf{A} - \mathbf{M}\|_2^2 \leq \epsilon$ and let $\mathcal{V}$ be the subspace of $\mathbb{R}^d$ that is spanned by the top-k eigenvectors of $\mathbf{M}$. There exist k vectors $\mathbf{v}^{(i)} \in \mathcal{V}$ such that for the matrix $\mathbf{V} \in \mathbb{R}^{k \times d}$ constructed by the vectors $\mathbf{v}^{(i)}$, it holds $\mathbf{E}_{\mathbf{x} \sim \mathcal{N}^d}[(f_{\alpha, \mathbf{W}}(\mathbf{x}) - f_{\alpha, \mathbf{V}}(\mathbf{x}))^2] \leq 2k\epsilon \, \mathbf{E}_{\mathbf{x} \sim \mathcal{N}^d}[f_{\alpha, \mathbf{W}}(\mathbf{x})] \, \mathbf{E}_{t \sim \mathcal{N}}[(\phi'(t))^2]/C_1$.*

Proof For simplicity, let us denote $\epsilon = \|\mathbf{A} - \mathbf{M}\|_2$. Moreover, let $\mathbf{A}' = \mathbf{A} - \mathbf{E}_{t\sim\mathcal{N}}[\phi(t)] \sum_i \alpha_i \mathbf{I}$, $\mathbf{M}' = \mathbf{M} - \mathbf{E}_{t\sim\mathcal{N}}[\phi(t)] \sum_i \alpha_i \mathbf{I}$, and observe that $\|\mathbf{A} - \mathbf{M}\|_2 = \|\mathbf{A}' - \mathbf{M}'\|_2$. We note that $\mathbf{A}' = \mathbf{E}_{t\sim\mathcal{N}}[\phi(t)(t^2-1)] \sum_{i=1}^k \alpha_i \mathbf{w}^{(i)}\mathbf{w}^{(i)^T}$, from Fact 7. Let $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(k)}$ be the eigenvectors corresponding to the top-k eigenvalues of $\mathbf{M}'$ (which are also the top k eigenvectors of $\mathbf{M}$), and let $\mathcal{V} = \mathrm{span}(\mathbf{b}_1, \ldots, \mathbf{b}_k)$. Let $\mathbf{v}^{(i)} = \mathrm{proj}_{\mathcal{V}}(\mathbf{w}^{(i)})$ and $\mathbf{r}^{(i)} = \mathbf{w}^{(i)} - \mathbf{v}^{(i)}$. Let $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}$ be any k vectors in $\mathbb{R}^d$. Then we have,

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f_{\alpha,\mathbf{W}}(\mathbf{x}) - f_{\alpha,\mathbf{V}}(\mathbf{x}))^2] &\le k \sum_{i=1}^k \alpha_i^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\left(\phi\big(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle\big) - \phi\big(\langle \mathbf{v}^{(i)}, \mathbf{x}\rangle\big)\right)^2\right] \\ &\le 2k \mathop{\mathbf{E}}_{t\sim\mathcal{N}}\left[(\phi'(t))^2\right] \sum_{i=1}^k \alpha_i^2 (1 - \langle \mathbf{w}^{(i)}, \mathbf{v}^{(i)}\rangle), \end{aligned} \tag{2}$$

where for the last inequality we used Lemma 9 and the fact that the random variables $\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle$ and $\langle \mathbf{v}^{(i)}, \mathbf{x}\rangle$ are ρ_i-correlated with $\rho_i = \langle \mathbf{w}^{(i)}, \mathbf{v}^{(i)}\rangle$.

It suffices to prove that $\left\|\mathbf{r}^{(i)}\right\|_2 = \left\|\mathbf{w}^{(i)} - \mathbf{v}^{(i)}\right\|_2 \le \epsilon'$ for some sufficiently small ϵ'. Note that because $\mathbf{r}^{(i)} \in \mathcal{V}^\perp$, it holds $\mathbf{r}^{(i)^T}\mathbf{M}'\mathbf{r}^{(i)} \le \left\|\mathbf{r}^{(i)}\right\|_2^2 \max_{\mathbf{u}\in\mathcal{V}^\perp} \frac{\mathbf{u}^T\mathbf{M}'\mathbf{u}}{\|\mathbf{u}\|_2}$, because we know that the subspace $\mathcal{W}$ is spanned by the top k eigenvectors of $\mathbf{M}'$. Let $\mathbf{u} = \sum_{i=1}^d \mathbf{u}^{(i)}$, where $\mathbf{u}^{(i)} \in \ker(\mathbf{M} - \lambda_i \mathbf{I})$ for all $i \in \{k+1, \ldots, d\}$ and λ_i is the i-th greatest eigenvalue. From Weyl's inequality, we have that if A_i are the eigenvalues of $\mathbf{A}'$ in decreasing order then $\|A_i - \lambda_i\|_1 \le \epsilon$ and we know that the eigenvalues of $\mathbf{A}'$ for $i > k$ are zero, because the $\mathrm{rank}(\mathbf{A}) \le k$. Thus,

$$\max_{\mathbf{u}\in\mathcal{V}^\perp} \frac{\mathbf{u}^T\mathbf{M}'\mathbf{u}}{\|\mathbf{u}\|_2} \le \lambda_{k+1} \le \epsilon\,,$$

because the eigenvalues of the eigenvectors of $\mathbf{M}'$ in $\mathcal{V}^\perp$ are less than ϵ, which implies that $\mathbf{r}^{(i)^T}\mathbf{M}'\mathbf{r}^{(i)} \le \epsilon \left\|\mathbf{r}^{(i)}\right\|_2^2$. We also have $\mathbf{r}^{(i)^T}\mathbf{A}'\mathbf{r}^{(i)} \ge \mathbf{E}_{t\sim\mathcal{N}}[\phi(t)(t^2-1)]\alpha_i \mathbf{r}^{(i)^T}\mathbf{w}^{(i)}\mathbf{w}^{(i)^T}\mathbf{r}^{(i)} = C_1\alpha_i \cdot \left(1 - \left\|\mathbf{v}^{(i)}\right\|_2^2\right)^2 = C_1\alpha_i \left\|\mathbf{r}^{(i)}\right\|_2^4$, where the last equality follows from the Pythagorean theorem. Therefore,

$$\left\|\mathbf{r}^{(i)}\right\|_2^2 \epsilon \ge \mathbf{r}^{(i)^T}\mathbf{M}'\mathbf{r}^{(i)} \ge \mathbf{r}^{(i)^T}\mathbf{A}'\mathbf{r}^{(i)} - \epsilon\left\|\mathbf{r}^{(i)}\right\|_2^2 \ge C_1\alpha_i \left\|\mathbf{r}^{(i)}\right\|_2^4 - \epsilon\left\|\mathbf{r}^{(i)}\right\|_2^2.$$

Thus, we obtain $\alpha_i \left\|\mathbf{r}^{(i)}\right\|_2^2 \le 2\epsilon/C_1$. The bound now follows directly from (2) since $2\alpha_i(1 - \langle \mathbf{w}^{(i)}, \mathbf{v}^{(i)}\rangle) = \alpha_i \left\|\mathbf{w}^{(i)} - \mathbf{v}^{(i)}\right\|_2^2 = \alpha_i \left\|\mathbf{r}^{(i)}\right\|_2^2 \le 2\epsilon/C_1$. ■

Now we have all the ingredients to complete our proof. Since the dimension of the subspace that we have learned is at most k, we can construct a grid with $(k/\epsilon)^{O(k)}$ candidates that contains an approximate solution. Our full algorithm is summarized as Algorithm 1. The proof of Theorem 5 follows from the above discussion and can be found in Appendix A.

Algorithm 1 Learning One-Hidden-Layer Networks with Positive Coefficients and Lipschitz Activations

procedure NNLEARNER(k, ϵ) ▷ k: number of rows of weight matrix $\mathbf{W}$, ϵ: accuracy.
  Draw $m = d\,\mathrm{poly}(k, 1/\epsilon)$ samples, $(\mathbf{x}^{(i)}, y^{(i)})$, to estimate $\widehat{\mathbf{M}}$. ▷ Lemma 8
  Find the SVD of $\widehat{\mathbf{M}}$ to obtain the k eigenvectors $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}$ that correspond to the k largest eigenvalues, and let $\mathcal{V}$ be the subspace spanned by these vectors.
  Draw $m' = O(kL^2)$ samples and compute an estimation $\hat{\mu}$ of the expectation of $f(x)$
  Let $\mathcal{G}$ be an ϵ/k-cover of a k-ball wth radius $(\hat{\mu} + c\sigma)^2$ over $\mathcal{V}$, with respect the ℓ_2-norm.
  Draw $n = \mathrm{poly}(k, 1/\epsilon)$ fresh samples $(\mathbf{x}^{(i)}, y^{(i)})$.
  For every $\mathbf{U} = (\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}) \in \mathcal{G}^k$, let $f_{\mathbf{U}} = \sum_{i=1}^{k} \left\|\mathbf{u}^{(i)}\right\|_2 \phi\big(\left\langle \mathbf{u}^{(i)}, \mathbf{x}\right\rangle / \left\|\mathbf{u}^{(i)}\right\|_2\big)$ and compute $e_{\mathbf{U}} = \frac{1}{n}\sum_{i=1}^{n}\left(f_{\mathbf{U}}(\mathbf{x}^{(i)}) - y^{(i)}\right)^2$
  Output the candidate $f_{\mathbf{U}}$ which minimizes its corresponding error $e_{\mathbf{U}}$.

4. Statistical Query Lower Bound

We start by formally defining the class of algorithms for which our lower bound applies. In the standard statistical query model, we do not have direct access to samples from the distribution, but instead can pick a function q and get an approximation to its expected value. In this work, we consider algorithms that have access to correlational statistical queries, which are more restrictive and are defined as follows. We remark that in the following definition of inner product queries we do not assume that the concept $f(\mathbf{x})$ is bounded pointwise but only in the L_2 sense. The properties that we shall need for our result hold also under this weaker assumption.

Definition 11 (Correlational/Inner Product Queries) *Let $\mathcal{D}$ be a distribution over some domain X and let $f : X \mapsto \mathbb{R}$, where $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[f^2(\mathbf{x})] \leq 1$. An inner product query is specified by some function $q : X \mapsto [-1, 1]$ and a tolerance $\tau > 0$, and returns a value u such that $u \in [\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[q(\mathbf{x})f(\mathbf{x})] - \tau, \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[q(\mathbf{x})f(\mathbf{x})] + \tau]$.*

We will prove that almost any reasonable choice of activations σ, ϕ defines a family of functions that is hard to learn. More precisely, for a pair of activations σ, ϕ, we define the following function $f_{\sigma,\phi} : \mathbb{R}^2 \to \mathbb{R}$:

$$f_{\sigma,\phi}(x, y) = \sigma\left(\sum_{m=1}^{2k}(-1)^m \phi\left(x\cos\left(\frac{\pi m}{k}\right) + y\sin\left(\frac{\pi m}{k}\right)\right)\right). \tag{3}$$

We are now ready to define the conditions on the activations σ, ϕ that are needed for our construction. We define

$$\mathcal{H} = \left\{ f_{\sigma,\phi} \;:\; \sigma \text{ is odd and } f_{\sigma,\phi} \not\equiv 0 \right\}, \tag{4}$$

where the second condition means that $f_{\sigma,\phi}(x, y)$ *as a function of* x, y is not identically zero. We can now define the class of (normalized) functions on $\mathbb{R}^d$ for which our lower bound holds. Given a set $\mathcal{W}$ of $2 \times d$ matrices, we can embed $f_{\sigma,\phi}$ into $\mathbb{R}^d$ by defining the following class of functions

$$\mathcal{F}_{\sigma,\phi}^{\mathcal{W}} = \{\mathbf{x} \mapsto f_{\sigma,\phi}(\mathbf{W}\mathbf{x}) / \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f_{\sigma,\phi}(\mathbf{W}\mathbf{x})] : \mathbf{W} \in \mathcal{W}\}. \tag{5}$$

Remark 12 *For any $f \in \mathcal{H}$, we have that $f : \mathbb{R}^2 \to \mathbb{R}$. We embedded f into $\mathbb{R}^d$ by taking $f_{\mathbf{W}}(\mathbf{x}) = f(\mathbf{W}\mathbf{x})$ for some $2 \times d$ matrix $\mathbf{W}$ with orthogonal rows. We prove correlational SQ lower bounds against learning an approximation of the embedding plane $\mathbf{W}$ from a function $f_{\mathbf{W}}$. This will imply a lower bound against learning $f_{\mathbf{W}}$ so long as the function does not vanish identically. However, this is not an entirely trivial condition. For example, if ϕ is a polynomial of degree less than k, this will happen. However, as we show in Appendix C.3, this is essentially the only way that things can go wrong. In particular, so long as ϕ is not a low degree polynomial and the parity of k is chosen appropriately, this function f will not vanish, and our lower bounds will apply.*

Theorem 13 (Correlational SQ Lower Bound) *Let σ, ϕ be activations such that $f_{\sigma,\phi} \in \mathcal{H}$ (see Eq. (4)). There exists a set $\mathcal{W}$ of matrices $\mathbf{W} \in \mathbb{R}^{2\times d}$ such that for all $f \in \mathcal{F}^{\mathcal{W}}_{\sigma,\phi}$ (see Eq. (5)) $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})] = 1$ and the following holds: Any correlational SQ learning algorithm that for every concept $f \in \mathcal{F}^{\mathcal{W}}_{\sigma,\phi}$ learns a hypothesis h such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[(f(\mathbf{x}) - h(\mathbf{x}))^2] \leq \epsilon$, where $\epsilon > 0$ is some sufficiently small constant, requires either $2^{d^{\Omega(1)}}$ inner product queries or at least one query with tolerance $d^{-\Omega(k)} + 2^{-d^{\Omega(1)}}$.*

To prove our lower bound we will use an appropriate notion of SQ dimension. Specifically, we define the Correlational SQ Dimension that captures the difficulty of learning a class $\mathcal{C}$.

Definition 14 (Correlational Statistical Query Dimension) *Let $\rho > 0$, let $\mathcal{D}$ be a probability distribution over some domain X, and let $\mathcal{C}$ be a family of functions $f : X \mapsto \mathbb{R}$. We denote by $\rho(\mathcal{C})$ the average pairwise correlation of any two functions in $\mathcal{C}$, that is $\rho(\mathcal{C}) = \frac{1}{|\mathcal{C}|^2}\sum_{g,r\in\mathcal{C}} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[g(\mathbf{x}) \cdot r(\mathbf{x})]$. The correlational statistical dimension of $\mathcal{C}$ relative to $\mathcal{D}$ with average correlation, denoted by $\mathrm{SDA}(\mathcal{C}, \mathcal{D}, \rho)$, is defined to be the largest integer m such that for every subset $\mathcal{C}' \subseteq \mathcal{C}$ of size at least $|\mathcal{C}'| \geq |\mathcal{C}|/m$, we have $\rho(\mathcal{C}') \leq \rho$.*

The following lemma relates the Correlational Statistical Query Dimension of a concept class with the number of correlational statistical queries needed to learn it. The difficulty lies in creating a large family of functions with small average correlation. We will use the following result that translates correlational statistical dimension to a lower bound on the number of inner product queries needed to learn the function $f \in \mathcal{C}$. We note that in this paper we consider inner-product queries of the form $g(x)y$ where y is not necessarily bounded. In fact, the proof of the following lemma does not require $g(x)y$ to be pointwise bounded (bounded L_2 norm is sufficient) as it can be seen from the arguments in Szörényi (2009), Goel et al. (2020a), Vempala and Wilmes (2019).

Lemma 15 *Let $\mathcal{D}$ be a distribution on a domain X and let $\mathcal{C}$ be a family of functions $f : X \mapsto \mathbb{R}$. Suppose for some $m, \tau > 0$, we have $\mathrm{SDA}(\mathcal{C}, \mathcal{D}, \tau) \geq m$ and assume that for all $f \in \mathcal{C}$, $1 \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[f^2(\mathbf{x})] > \eta^2$. Any SQ learning algorithm that is allowed to make only inner product queries and for any $f \in \mathcal{C}$ outputs some hypothesis h such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[(h(\mathbf{x}) - f(\mathbf{x}))^2] \leq c\,\eta^2$, where $c > 0$ is a sufficiently small constant, requires at least $\Omega(m)$ queries of tolerance $\sqrt{\tau}$.*

We will require the following technical lemma, whose proof relies on Hermite polynomials, and can be found in Appendix C.

Lemma 16 *Let $p(\mathbf{x}) : \mathbb{R}^2 \mapsto \mathbb{R}$ be a function and let $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{2\times d}$ be linear maps such that $\mathbf{U}\mathbf{U}^T = \mathbf{V}\mathbf{V}^T = \mathbf{I} \in \mathbb{R}^{2\times 2}$. Then, $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[p(\mathbf{U}\mathbf{x})p(\mathbf{V}\mathbf{x})] \leq \sum_{m=0}^{\infty} \left\|\mathbf{U}\mathbf{V}^T\right\|_2^m \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[(p^{[m]}(\mathbf{x}))^2]$.*

In the following simple lemma, we show that two random 2-dimensional subspaces in high dimensions are nearly orthogonal. In particular, we can have an exponentially large family of almost orthogonal planes. For the proof see Appendix C.

Lemma 17 *For any $0 < c < 1/2$, there exists a set S of at least $2^{\Omega(d^c)}$ matrices in $\mathbb{R}^{2\times d}$ such that for each pair $\mathbf{A}, \mathbf{B} \in S$, it holds $\left\|\mathbf{A}\mathbf{B}^T\right\|_2 \leq O(d^{c-1/2})$.*

The following lemma shows that the correlation of any function f of $\mathcal{H}$ with any low-degree polynomial is zero. For the proof see Appendix C.

Lemma 18 *Let $f_{\sigma,\phi} \in \mathcal{H}$. For every polynomial $p(\mathbf{x})$ of degree at most k, it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[f_{\sigma,\phi}(\mathbf{x}) \cdot p(\mathbf{x})] = 0$.*

We are now ready to prove our main result.

Proof [Proof of Theorem 13] Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ from Lemma 18. Let $c > 0$ and fix a set $\mathcal{W}$ of matrices in $\mathbb{R}^{2\times d}$ satisfying the properties of Lemma 17. We consider the class of functions $F^{\mathcal{W}}_{\sigma,\phi}$ (see Eq. (5)). In particular, for all $\mathbf{A}_i, \mathbf{A}_j \in \mathcal{W}$, let functions $G_i(\mathbf{x}) = f(\mathbf{A}_i\mathbf{x})/\sqrt{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^2}[f^2(\mathbf{x})]}$ and $G_j(\mathbf{x}) = f(\mathbf{A}_j\mathbf{x})/\sqrt{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^2}[f^2(\mathbf{x})]}$. Notice that since $\mathbf{A}_i\mathbf{A}_i^T = \mathbf{I}$ we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[G_i^2(\mathbf{x})] = 1$ for all i. The pairwise correlation of G_i and G_j is

$$\rho(G_i, G_j) = \frac{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[G_i(\mathbf{x})G_j(\mathbf{x})]}{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^2}[f^2(\mathbf{x})]}\,, \tag{6}$$

where in the second equality we used that Gaussian distributions are invariant under rotations and in the last that the expectation of $p(\mathbf{x})$ is zero. Then, using Lemma 16, it holds

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[G_i(\mathbf{x})G_j(\mathbf{x})] &= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{A}_i\mathbf{x})p(\mathbf{A}_j\mathbf{x})] \leq \sum_{m>k} \left\|\mathbf{A}_i\mathbf{A}_j^T\right\|_2^m \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}[(f^{[m]}(\mathbf{x}))^2] \\
&\leq \left\|\mathbf{A}_i\mathbf{A}_j^T\right\|_2^{k+1} \sum_{m>k} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}[(f^{[m]}(\mathbf{x}))^2] \leq \left\|\mathbf{A}_i\mathbf{A}_j^T\right\|_2^{k+1} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}[(f(\mathbf{x}))^2] \\
&\leq O(d^{k(c-1/2)}) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^2}[(f(\mathbf{x}))^2]\,,
\end{aligned} \tag{7}$$

where in the first inequality we used that the first k moments are zero, in the second the fact that the spectral norm of these two matrices is less than one, and in the third inequality we used Parseval's theorem. Thus, using Equation (7) into Equation (6), we get that the pairwise correlation is less than $\tau = O(d^{k(c-1/2)})$. Thus, from a straighforward calculation, the average correlation of the set $\mathcal{F}^{\mathcal{W}}_{\sigma,\phi}$ is less $\tau + \frac{1-\tau}{|\mathcal{F}^{\mathcal{W}}_{\sigma,\phi}|} \leq \tau + |\mathcal{F}^{\mathcal{W}}_{\sigma,\phi}|^{-1} \leq \tau + 2^{-\Omega(d^c)}$. Moreover, for $\tau' = d^{O(k(c-1/2))} + 2^{-\Omega(d^c)}$, the $\mathrm{SDA}(\mathcal{F}^{\mathcal{W}}_{\sigma,\phi}, \mathcal{D}, \tau') = 2^{\Omega(d^c)}$ and the result follows from Lemma 15. ■

References

A. Bakshi, R. Jayaram, and D. P. Woodruff. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory, COLT 2019*, pages 195–268, 2019.

A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the Twenty-Sixth Annual Symposium on Theory of Computing*, pages 253–262, 1994.

A. Daniely, R. Frostig, and Y. Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 2253–2261, 2016.

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

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

V. Feldman. A general characterization of the statistical query complexity. *CoRR*, abs/1608.02198, 2016.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of STOC'13*, pages 655–664, 2013.

V. Feldman, C. Guzman, and S. Vempala. Statistical query algorithms for stochastic convex optimization. *CoRR*, abs/1512.09170, 2015a.

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

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *J. ACM*, 64(2):8:1–8:37, 2017.

R. Ge, J. D. Lee, and T. Ma. Learning one-hidden-layer neural networks with landscape design. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

R. Ge, R. Kuditipudi, Z. Li, and X. Wang. Learning two-layer neural networks with symmetric inputs. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

S. Goel and A. R. Klivans. Learning neural networks with two nonlinear layers in polynomial time. In *Conference on Learning Theory, COLT 2019*, pages 1470–1499, 2019.

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

S. Goel, A. Gollakota, Z. Jin, S. Karmalkar, and A. Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent, 2020a.

S. Goel, A. Klivans, P. Manurangsi, and D. Reichman. Tight hardness results for learning depth-2 relu networks, 2020b. Personal communication.

M. Janzamin, H. Sedghi, and A. Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods, 2015.

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

A. Klivans. Talk at stoc'17 workshop on new challenges in machine learning – robustness and nonconvexity, 2017.

V. Kontonis, C. Tzamos, and M. Zampetakis. Efficient truncated statistics with unknown truncation. In *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019*, pages 1578–1595, 2019.

P. Manurangsi and D. Reichman. The computational complexity of training relu(s), 2018.

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

H. Sedghi, M. Janzamin, and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, pages 1223–1231, 2016.

O. Shamir. Distribution-specific hardness of learning neural networks. *J. Mach. Learn. Res.*, 19: 32:1–32:29, 2018. URL http://jmlr.org/papers/v19/17-537.html.

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

B. Szörényi. Characterizing statistical query learning: simplified notions and proofs. In *International Conference on Algorithmic Learning Theory*, pages 186–200. Springer, 2009.

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

S. Vempala. Learning convex concepts from gaussian distributions with PCA. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 124–130, 2010.

S. Vempala and J. Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and SQ lower bounds. In *Conference on Learning Theory, COLT 2019*, pages 3115–3117, 2019. Full version available at https://arxiv.org/abs/1805.02677.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Y. Zhang, J. D. Lee, and M. I. Jordan. L1-regularized neural networks are improperly learnable in polynomial time. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, pages 993–1001, 2016.

K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 4140–4149, 2017.

Appendix A. Omitted Proofs from Section 3

In the following simple fact, we compute the degree-1 and degree-2 Chow parameters of a one-layer network.

Fact 19 (Low-degree Chow parameters) *Let* $f(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i \phi\big(\langle \mathbf{w}^{(i)}, \mathbf{x} \rangle\big)$. *Then*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})] = B_1 \sum_{i=1}^{k} \alpha_i \qquad \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})\mathbf{x}] = C \sum_{i=1}^{k} \alpha_i \mathbf{w}^{(i)}$$

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})(\mathbf{x}\mathbf{x}^T - \mathbf{I})] = B \sum_{i=1}^{k} \alpha_i \mathbf{w}^{(i)} \mathbf{w}^{(i)^T}$$

where $B_1 = \mathbf{E}_{t\sim\mathcal{N}}[\phi(t)]$,$C = \mathbf{E}_{t\sim\mathcal{N}}[\phi(t)t]$ *and* $B = \mathbf{E}_{t\sim\mathcal{N}}[\phi(t)(t^2-1)]$.

Proof

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})] &= \sum_{i=1}^{k} \alpha_i \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\phi\left(\left\langle \mathbf{w}^{(i)}, \mathbf{x} \right\rangle\right)\right] \\
&= \sum_{i=1}^{k} \alpha_i \int_{\mathbb{R}^d} \phi\left\langle \mathbf{x}, \mathbf{w}^{(i)} \right\rangle \mathcal{N}(\mathbf{x})\mathrm{d}\mathbf{x} = B_1 \sum_{i=1}^{k} \alpha_i \,,
\end{aligned}$$

where in the third equality we used the fact that normal distribution is invariant under rotations. For the second equality, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})\mathbf{x}] &= \sum_{i=1}^{k} \alpha_i \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\phi\left(\left\langle \mathbf{w}^{(i)}, \mathbf{x} \right\rangle\right)\mathbf{x}\right] \\
&= \sum_{i=1}^{k} \alpha_i \int_{\mathbb{R}^d} \max\left(\left\langle \mathbf{x}, \mathbf{w}^{(i)} \right\rangle, 0\right)\mathbf{x}\mathcal{N}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \sum_{i=1}^{k} \alpha_i \mathbf{R}_i^{-1} \int_{\mathbb{R}^d} \max\left(\langle \mathbf{x}, \mathbf{e}_1 \rangle, 0\right)\mathbf{x}\mathcal{N}(\mathbf{x})\det(J(\mathbf{R}_i))\mathrm{d}\mathbf{x} \\
&= \sum_{i=1}^{k} \alpha_i \mathbf{w}^{(i)}/2 \,,
\end{aligned}$$

where $\mathbf{R}_i$ is some rotation matrix that maps $\mathbf{w}^{(i)}$ to $\mathbf{e}_1$, and J is the Jacobian of this rotation which has always determinant of 1. The Chow parameters of degree-2 are given by

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})(\mathbf{x}\mathbf{x}^T-\mathbf{I})] &= \sum_{i=1}^k \alpha_i \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\phi\left(\left\langle \mathbf{w}^{(i)},\mathbf{x}\right\rangle\right)(\mathbf{x}\mathbf{x}^T-\mathbf{I})\right] \\
&= \sum_{i=1}^k \alpha_i \int_{\mathbb{R}^d} \phi\left(\left\langle \mathbf{x},\mathbf{w}^{(i)}\right\rangle\right)(\mathbf{x}\mathbf{x}^T-\mathbf{I})\mathcal{N}(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \sum_{i=1}^k \alpha_i \mathbf{R}_i^{-1} \int_{\mathbb{R}^d} \phi\left(\langle \mathbf{x},\mathbf{e}_1\rangle\right)(\mathbf{x}\mathbf{x}^T-\mathbf{I})\mathcal{N}(\mathbf{x})\det(J(\mathbf{R}_i))\mathrm{d}\mathbf{x}\mathbf{R}_i^{-1^T} \\
&= \sum_{i=1}^k \alpha_i \mathbf{R}_i^{-1} \int_{\mathbb{R}^d} \sum_{k,l=1}^d \phi\left(\mathbf{x}_1\right)(\mathbf{x}_k\mathbf{x}_l-\delta_{k,l})\mathbf{e}_k\mathbf{e}_l^T\mathcal{N}(\mathbf{x})\mathrm{d}\mathbf{x}\mathbf{R}_i^{-1^T}
\end{aligned}$$

There are 4 cases. First case is when $k\neq l\neq 1$. Then, by independence we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[\phi(\mathbf{x}_1)(\mathbf{x}_k\mathbf{x}_l-\delta_{k,l})] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[\phi(\mathbf{x}_1)]\,\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[(\mathbf{x}_k\mathbf{x}_l)] = 0$, where we used the independence of the random variables $\mathbf{x}_k,\mathbf{x}_l$. Similarly, if $k\neq l$ and $k=1$ we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[\phi(\mathbf{x}_1)(\mathbf{x}_1\mathbf{x}_l)]=0$. If $k=l\neq 1$, then $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[\phi(\mathbf{x}_1)(\mathbf{x}_l^2-1)=0$, because $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[\mathbf{x}_l^2]=1$. Thus, the only non zero case is when $k=l=1$. Then,

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})(\mathbf{x}\mathbf{x}^T-\mathbf{I})] &= \sum_{i=1}^k \alpha_i \mathbf{R}_i^{-1} \int_{\mathbb{R}^d} \phi\left(\mathbf{x}_1\right)(\mathbf{x}_1^2-1)\mathbf{e}_1\mathbf{e}_1^T\mathcal{N}(\mathbf{x})\mathrm{d}\mathbf{x}\mathbf{R}_i^{-1^T} \\
&= B\sum_{i=1}^k \alpha_i \mathbf{w}^{(i)}\mathbf{w}^{(i)^T}\,,
\end{aligned}$$

■

Lemma 20 *Let $f_{\alpha,\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^k \alpha_i\phi(\langle \mathbf{w}^{(i)},\mathbf{x}\rangle)$ and $f_{\beta,\mathbf{V}}(\mathbf{x}) = \sum_{i=1}^k \beta_i\phi(\langle \mathbf{v}^{(i)},\mathbf{x}\rangle)$ with $\alpha_i,\beta_i > 0$, then it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[(f_{\alpha,\mathbf{W}}(\mathbf{x})-f_{\beta,\mathbf{V}}(\mathbf{x}))^2] \leq 2k\,\mathbf{E}_{t\sim\mathcal{N}}[(\phi'(t))^2]\sum_{i=1}^k \alpha_i^2\left\|\mathbf{v}^{(i)}-\mathbf{w}^{(i)}\right\|_2 + k\,\mathbf{E}_{t\sim\mathcal{N}}[\phi(t)^2]\sum_{i=1}^k(\alpha_i-\beta_i)^2$.*

Proof We have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f_{\alpha,\mathbf{W}}(\mathbf{x})-f_{\beta,\mathbf{V}}(\mathbf{x}))^2] &\leq k \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\sum_{i=1}^k\left(\alpha_i\phi\Big(\left\langle \mathbf{w}^{(i)},\mathbf{x}\right\rangle\Big)-\beta_i\phi\Big(\left\langle \mathbf{v}^{(i)},\mathbf{x}\right\rangle\Big)\right)^2\right] \\
\leq k \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\sum_{i=1}^k \alpha_i\left(\phi\Big(\left\langle \mathbf{w}^{(i)},\mathbf{x}\right\rangle\Big)-\phi\Big(\left\langle \mathbf{v}^{(i)},\mathbf{x}\right\rangle\Big)\right)^2\right] &+ \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\sum_{i=1}^k \phi\Big(\left\langle \mathbf{v}^{(i)},\mathbf{x}\right\rangle\Big)^2(\alpha_i-\beta_i)^2\right] \\
&\leq 2k \mathop{\mathbf{E}}_{t\sim\mathcal{N}}[(\phi'(t))^2]\sum_{i=1}^k \alpha_i^2\left\|\mathbf{v}^{(i)}-\mathbf{w}^{(i)}\right\|_2 + k\mathop{\mathbf{E}}_{t\sim\mathcal{N}}[\phi(t)^2]\sum_{i=1}^k(\alpha_i-\beta_i)^2\,,
\end{aligned}$$

where we used Lemma 10. ■

Fact 21 *Let $f(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i \phi(\langle \mathbf{w}_i, \mathbf{x} \rangle)$ and $y = f(\mathbf{x}) + \xi$ where ξ is zero mean subgaussian with variance σ^2. Let $B_2 = \mathbf{E}_{t\sim\mathcal{N}}[\phi^2(t)]$ and $c > 0$ a constant, then using $O(kB_2)$ samples we can find $\hat{\mu}$ such as*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})] \leq 2\hat{\mu} + 2c\sqrt{\frac{\sigma^2}{k}} \quad \text{and} \quad \hat{\mu} \leq \frac{3}{2} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})] + c\sqrt{\frac{\sigma^2}{k}}$$

with probability at least 3/4.

Proof Let $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}$, then from Chebyshev's inequality, we have

$$\mathbf{Pr}[|\hat{\mu} - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})]| \geq 2\sqrt{\mathbf{Var}[y]/m}] \leq 1/4 \,.$$

Thus with probability $3/4$, it holds

$$\begin{aligned} |\hat{\mu} - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})]| \leq 2\sqrt{\mathbf{Var}[y]/m} &\leq 2\sqrt{\frac{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})]}{m}} + 2\sqrt{\frac{\sigma^2}{m}} \\ &\leq 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})]\sqrt{\frac{kB_2}{m}} + 2\sqrt{\frac{\sigma^2}{m}} \,, \end{aligned}$$

to get last inequality we used CauchySchwarz. Taking $m = O(kB_2)$ we get $\frac{1}{2}\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})] \leq \hat{\mu} + c\sqrt{\frac{\sigma^2}{k}}$ and $\hat{\mu} \leq \frac{3}{2}\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})] + c\sqrt{\frac{\sigma^2}{k}}$. ■

Lemma 22 *Let $f(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i \phi(\langle \mathbf{w}_i, \mathbf{x} \rangle)$, $B_2 = \mathbf{E}_{t\sim\mathcal{N}}[\phi^2(t)]$ and $B_4 = \mathbf{E}_{t\sim\mathcal{N}}[\phi^4(t)]$. Then $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f^4(\mathbf{x})] \leq \frac{B_4}{B_2^2} k^2 \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})^2]^2$.*

Proof To bound $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f^4(x)]$, using Cauchy-Schwartz, it holds that

$$\begin{aligned} \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f^4(\mathbf{x})]} \leq \sum_{i=1}^{k} \alpha_i^2 \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}\Big[\Big(\sum_{i=1}^{k} \phi^2(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle)\Big)^2\Big]} &\leq \sum_{i=1}^{k} \alpha_i^2 \sqrt{k\,\mathbf{E}\Big[\sum_{i=1}^{k} \phi^4(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle)\Big]} \\ &\leq k B_4^{1/2} \sum_{i=1}^{k} \alpha_i^2 \leq k \frac{B_4^{1/2}}{B_2} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})^2] \,, \end{aligned}$$

where in the last inequality we used that $\sum_{i=1}^{k} \alpha_i^2 B_2 \leq \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})]$. ■

Lemma 23 *Let $f(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i \phi(\langle \mathbf{w}_i, \mathbf{x} \rangle)$ and $y = f(\mathbf{x}) + \xi$ where ξ is zero mean subgaussian with variance σ^2. Moreover, let $B_2 = \mathbf{E}_{t\sim\mathcal{N}}[\phi^2(t)]$ and $B_4 = \mathbf{E}_{t\sim\mathcal{N}}[\phi^4(t)]$. Then, if $Y_u = \frac{1}{m}\sum_{i=1}^{m}(f_u(\mathbf{x}^{(i)}) - y^{(i)}))^2$,we can find $\hat{Y}_u$ such that*

$$|\hat{Y}_u - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[Y_u]| \leq c\epsilon^2 k^2 \frac{B_4^{1/2}}{B_2} \left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f_u^2(\mathbf{x})] + \sigma^2 \right)$$

with probability $1 - \delta$ with $O(\frac{1}{\epsilon^4}\log(1/\delta))$ samples, where c is a universal constant.

Proof Let $Y = (f_u(\mathbf{x}) - y)^2 = (f_u(\mathbf{x}) - f(\mathbf{x}))^2 + y^2 - 2y(f_u(\mathbf{x}) - f(\mathbf{x}))$. Then the variance of each term is

$$\begin{aligned}\mathbf{Var}[(f_u(\mathbf{x}) - f(\mathbf{x}))^2] &\le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f_u(\mathbf{x}) - f(\mathbf{x}))^4] \le 4 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f_u^2(\mathbf{x}) + f^2(\mathbf{x}))^2] \\ &\le 8 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f_u^4(\mathbf{x}) + f^4(\mathbf{x})] \le 8\frac{B_4}{B_2^2}k^2\big(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})]^2 + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f_u^2(\mathbf{x})]^2\big)\,,\end{aligned}$$

where in the third and in the fourth inequality we used that $(a \pm b)^2 \le 2a^2 + 2b^2$ and in the last one we used Lemma 22. Thus,

$$\begin{aligned}\mathbf{Var}[Y] &\le 8\frac{B_4}{B_2^2}k^2\big(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})]^2 + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f_u^2(\mathbf{x})]^2\big) + 16e^2\sigma^4 + 2\sigma^2\big(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f_u^2(\mathbf{x})]\big) \\ &\le ck^4\frac{B_4}{B_2^2}\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})] + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f_u^2(\mathbf{x})] + \sigma^2\right)^2\,,\end{aligned}$$

where c is a universal constant. From Chebyshev's inequality, we have that we need $m = O(\frac{1}{\epsilon^4})$ for an error at most $\sqrt{c}\epsilon^2 k^2(\frac{B_4^{1/2}}{B_2}\left(\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})] + \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f_u^2(\mathbf{x})] + \sigma^2\right)$. Then using the median trick, we can boost the confidence to $1-\delta$ with $m\log(1/\delta)$ samples. ■

Since the dimension of the subspace, that we have learned, is at most k, the following standard lemma gives us that a grid with $(k/\epsilon)^{O(k)}$ candidates suffices.

Lemma 24 (Corollary 4.2.13 of Vershynin (2018)) *There exists be an ϵ-cover of the unit ball in $\mathbb{R}^k$, with respect the ℓ_2 norm, of size at most $(1 + 2/\epsilon)^k$.*

We now restate and prove our main theorem, Theorem 5, which we restate for convenience.

Theorem 25 (Learning sum of Lipschitz Activations) *Let $f(\mathbf{x}) = \sum_{i=1}^k \alpha_i\phi(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle)$ with $\alpha_i > 0$ for all $i \in [k]$, where $\phi(t)$ is an L-Lipschitz, non-negative activation function such that $\mathbf{E}_{t\sim\mathcal{N}}[\phi(t)] \ge C$, $\mathbf{E}_{t\sim\mathcal{N}}[\phi(t)(t^2-1)] \ge C$, where $C > 0$ and $\mathbf{E}_{t\sim\mathcal{N}}[\phi^2(t)]$ is finite. There exists an algorithm that given $k \in \mathbb{N}, \epsilon > 0$, and sample access to a noisy set of samples from $f : \mathbb{R}^d \to \mathbb{R}_+$, draws $m = d\cdot\mathrm{poly}(k, 1/\epsilon)\cdot\mathrm{poly}(L/C)$ samples, runs in time $\mathrm{poly}(m) + \widetilde{O}((1/\epsilon)^{k^2})$, and outputs a proper hypothesis h that, with probability at least $9/10$, satisfies*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f(\mathbf{x}) - h(\mathbf{x}))^2] \le \epsilon^2\mathrm{poly}(L/C)\left(\sigma^2 + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})^2]\right)\,.$$

Proof Denote $M_f = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})]$ and $M_{f^2} = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})^2]$. Using Lemma 8, we get that with $m = \widetilde{O}(d\,k^3/\epsilon^2)$ samples with high constant probability it holds that $\left\|\widehat{\mathbf{M}} - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})\mathbf{x}\mathbf{x}^T]\right\|_2 \le \frac{\epsilon}{k}(M_f\frac{L}{C} + \sigma)$. From Lemma 10, we obtain that there exists a matrix $\mathbf{V} \in \mathbb{R}^{k\times d}$ whose rows are vectors of the subspace $\mathcal{V}$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[(f_{\mathbf{V}}(\mathbf{x}) - f(\mathbf{x}))^2] \le 2\epsilon^2\frac{L^2}{C}(M_f^2\frac{L}{C} + M_f\sigma)$. From Fact 21, let $\hat{\mu}$ be an upper bound to M_f (that is $\hat{\mu} \le 2\mu + 2c\sqrt{\sigma^2/k}$ where μ is the estimated value), then using Fact 26, with the value $\hat{\mu}/k$, we get an approximation of each a_i with error $\epsilon\hat{\mu}/k$.

Using Lemma 24, the size of a cover is $|\mathcal{G}| \leq \left((1+4k/\epsilon)^k \log(k\epsilon)/\epsilon\right)^k$, because we need vectors with norm from ϵM_f to M_f, our cover is created using the upper bound on M_f. We have that there exists $\mathbf{U}$ whose rows are vectors in the cover $\mathcal{G}$ such that

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f_{\mathbf{U}}(\mathbf{x}) - f_{\mathbf{V}}(\mathbf{x}))^2] &\leq c(\epsilon^2 M_f^2 L^2 + L^2\epsilon^2\hat{\mu}^2) \\ &\leq c\epsilon^2 L^2(M_f^2 + M_f\sigma + \sigma^2) \\ &\leq c\epsilon^2 L^2(M_f + \sigma)^2 \,, \end{aligned} \tag{8}$$

where in the first inequality we used Lemma 20 and in the second one Fact 21. The error of the best hypothesis(i.e., the one that minimizes the error) in the cover, will be

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f_{\mathbf{U}}(\mathbf{x}) - f(\mathbf{x}))^2] &\leq 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f_{\mathbf{U}}(\mathbf{x}) - f_{\mathbf{V}}(\mathbf{x}))^2] + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f_{\mathbf{V}}(\mathbf{x}) - f(\mathbf{x}))^2] \\ &\leq 2c\epsilon^2 L^2(M_f + \sigma)^2 + 4\epsilon^2\frac{L^2}{C}(M_f^2\frac{L}{C} + M_f\sigma) \\ &\leq \epsilon^2 \mathrm{poly}(L/C)(M_f + \sigma)^2 \,. \end{aligned} \tag{9}$$

Finally, using the estimator from Line 7, Lemma 23, we conclude that $m'' = O(\frac{k^4}{\epsilon^4}\log(|\mathcal{G}|))$ samples are sufficient to test all the vectors of the cover $\mathcal{G}$ and find the one that minimizes the error with high probability. For each element $i \in \mathcal{G}$, let $e_i = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[(f(\mathbf{x}) - f_i(\mathbf{x}))^2] + \sigma^2$, which is the square error of the i-th hypothesis in $\mathcal{G}$ and let $\hat{e}_i$ be the estimated value. We have with high probability that

$$|\hat{e}_i - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[\hat{e}_i]| \leq \epsilon^2\mathrm{poly}(L/C)(\sigma^2 + M_{f^2}) \,, \tag{10}$$

where we used $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f_{\mathbf{U}}(\mathbf{x})] \leq k\hat{\mu} \leq kM_f + 2c'\sigma\sqrt{k}$. Set $h(\mathbf{x}) = \mathrm{argmin}_{i\in\mathcal{G}}\,|\hat{e}_i - \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[\hat{e}_i]|$, using Equations (9) and (10), then

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f(\mathbf{x}) - h(\mathbf{x}))^2] &\leq \epsilon^2\mathrm{poly}(L/C)(\sigma^2 + M_{f^2}) + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(f_{\mathbf{U}}(\mathbf{x}) - f(\mathbf{x}))^2] \\ &\leq \epsilon^2\mathrm{poly}(L/C)(\sigma^2 + M_{f^2}) + \epsilon^2\mathrm{poly}(L/C)\,(\sigma + M_f)^2 \\ &\leq \epsilon^2\mathrm{poly}(L/C)\left(\sigma^2 + M_{f^2}\right) \,, \end{aligned}$$

where the last inequality follows from Jensen's inequality. ■

Fact 26 *Let $\mathcal{G}$ be a set of unit vectors of size m. Then, we can construct a new set $\mathcal{G}'$ of size $m\log(1/\epsilon)/\epsilon$ with the property: For every $\alpha \in [0, B]$ and every vector $\mathbf{v} \in \mathcal{G}$, there exists a $\mathbf{w} \in \mathcal{G}'$ such that $\|\alpha\mathbf{v} - \mathbf{w}\|_2^2 \leq \epsilon^2 B^2$ and $\left\|\mathbf{v} - \frac{\mathbf{w}}{\|\mathbf{w}\|_2}\right\|_2^2 = 0$.*

Proof For each vector $\mathbf{v} \in \mathcal{G}$, add the vectors $(1-\epsilon)^i B\mathbf{v}$ for $i = 0, \ldots, \log(1/\epsilon)/\epsilon$ to $\mathcal{G}'$. Then, for all $\alpha \in [0, B]$ and for every vector $\mathbf{v} \in \mathcal{G}$ there exists a $\mathbf{w} \in G'$ such that $\|\alpha\mathbf{v} - \mathbf{w}\|_2^2 \leq \left\|(1-\epsilon)^{t+1}\mathbf{v} - \mathbf{v}(1-\epsilon)^t B\right\|_2^2 \leq \epsilon^2 B^2$, for a value t such that $\alpha \in [(1-\epsilon)^{t+1}B, (1-\epsilon)^t B]$. ■

Appendix B. Empirical Estimates of Chow Parameters

In this section we show that roughly $O(dk/\epsilon^2)$ samples are sufficient to estimate the Degree-2 Chow parameters in spectral norm.

Lemma 27 (Estimation of Degree-2 Chow parameters) *Let $f_{\alpha,\mathbf{W}}(\mathbf{x}) = \sum_{i=1}^{k} \alpha_i \phi(\langle \mathbf{w}^{(i)}, \mathbf{x}\rangle)$, where $\phi(x)$ is an L-Lipschitz, positive activation function such that $\mathbf{E}_{x\sim\mathcal{N}}[\phi(x)] \geq C$. Let $\mathbf{\Sigma} = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f_{\alpha,\mathbf{W}}(\mathbf{x})\mathbf{x}\otimes\mathbf{x}]$ be the degree-2 Chow parameters of $f_{\alpha,\mathbf{W}}$. Then, for some $N = \widetilde{O}(dk/\epsilon^2)$ samples $(\mathbf{x}^{(i)}, y^{(i)})$, where $y^{(i)} = f_{\alpha,\mathbf{W}}(\mathbf{x}^{(i)}) + \xi_i$ and ξ_i is a zero-mean, subgaussian noise with variance σ^2, it holds with probability at least 99% that*

$$\left\| \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)} y^{(i)} - \mathbf{\Sigma} \right\|_2 \leq \epsilon \left(\sigma + \frac{L}{C} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f_{\alpha,\mathbf{W}}(\mathbf{x})] \right) .$$

We are going to use the following lemma from Vershynin (2010) about concentration of matrices with heavy-tailed independent rows.

Lemma 28 (Theorem 5.48 of Vershynin (2010)) *Let $\mathbf{A}$ be an $N \times d$ matrix whose rows $\mathbf{A}_i$ are independent random vectors in $\mathbb{R}^d$ with the common second moment matrix $\mathbf{\Sigma} = \mathbf{E}[\mathbf{A}_i \mathbf{A}_i^T]$. Let $m = \mathbf{E}[\max_{i\leq N} \|\mathbf{A}_i\|_2^2]$. Then*

$$\mathbf{E}\left[\left\| \frac{1}{N}\mathbf{A}^T\mathbf{A} - \mathbf{\Sigma} \right\|_2 \right] \leq \max(\|\mathbf{\Sigma}\|_2^{1/2}\delta, \delta^2) , \qquad \text{where } \delta = C\sqrt{\frac{m\log(\min(N,d))}{N}} .$$

We are also going to use the following concentration result on sums of random matrices.

Lemma 29 (Rudelson's Inequality Corollary 5.28 in Vershynin (2010)) *Let $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ be fixed vectors in $\mathbb{R}^d$. Let $\xi^{(1)}, \ldots, \xi^{(N)}$ be zero mean sub-Gaussian with variance σ^2 random variables. Then*

$$\mathbf{E}\left[\left\| \sum_{i=1}^{N} \xi^{(i)} \mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)} \right\|_2 \right] \leq C\sigma\sqrt{\log d} \cdot \max_{i\leq N} \left\| \mathbf{x}^{(i)} \right\|_2 \left\| \sum_{i=1}^{N} \mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)} \right\|_2^{1/2} .$$

We are going to also use the following well-known result on concentration of polynomials of independent Gaussian random variables. See for example O'Donnell (2014).

Lemma 30 (Gaussian Hypercontractivity) *Let $p(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ be a degree m polynomial. Then*

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\left| p(\mathbf{x}) - \mathop{\mathbf{E}}_{\mathbf{y}\sim\mathcal{N}^d}[p(\mathbf{y})] \right| > t \right] \leq e^2 \exp\left(- \left(\frac{t^2}{C\, \mathbf{Var}_{\mathbf{x}\sim\mathcal{N}^d}[p(\mathbf{x})]} \right)^{1/m} \right),$$

where $C > 0$ is an absolute constant.

Proof [Proof of Lemma 8:] We have

$$\begin{aligned} \left\| \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)} f_{\alpha,\mathbf{W}}(\mathbf{x}^{(i)}) + \xi_i) - \mathbf{\Sigma} \right\|_2 &\leq \left\| \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)} f_{\alpha,\mathbf{W}}(\mathbf{x}^{(i)}) - \mathbf{\Sigma} \right\|_2 \\ &\quad + \left\| \frac{1}{N}\sum_{i=1}^{N} \xi^{(i)} \mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)} \right\|_2 . \end{aligned}$$

We next bound the probability that $\|\mathbf{x}\|_2^2 f_{\alpha,\mathbf{W}}(\mathbf{x})$ is large. We have

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}^d}[\,\|\mathbf{x}\|_2^2 f_{\alpha,\mathbf{W}} \geq t] &= \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\sum_{j=1}^k \alpha_j \|\mathbf{x}\|_2^2 \phi\left(\left\langle \mathbf{x}, \mathbf{w}^{(j)}\right\rangle\right) \geq t\right] \\ &\leq \sum_{j=1}^k \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\|\mathbf{x}\|_2^2 \phi\left(\left\langle \mathbf{x}, \mathbf{w}^{(j)}\right\rangle\right) \geq \frac{t}{k\sum_{j=1}^k \alpha_j}\right] \leq k \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\|\mathbf{x}\|_2^2 |\mathbf{x}_1| \geq \frac{t}{Lk\sum_{j=1}^k \alpha_j}\right], \end{aligned} \tag{11}$$

where for the second inequality we used the union bound and for the last one we used the rotation invariance of the normal distribution and the Euclidean norm to set $\mathbf{w}^{(j)} = \mathbf{e}_1$. Moreover, we used the fact that $\phi(\mathbf{x}_1) \leq L|x_1|$ since $\phi(\cdot)$ is L-Lipschitz.

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\|\mathbf{x}\|_2^2 |\mathbf{x}_1| \geq t\right] = \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}^d}\left[\left|\|\mathbf{x}\|_2^2 \mathbf{x}_1 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[\|\mathbf{x}\|_2^2 \mathbf{x}_1]\right| \geq t\right] \leq \exp(2 - (t^2/(C'd^2))^{1/3}) . \tag{12}$$

where we used Lemma 30 and the fact that $\mathbf{Var}_{\mathbf{x}\sim\mathcal{N}^d}[\|\mathbf{x}\|_2^2 x_1] = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[\|\mathbf{x}\|_2^4 x_1^2] = d^2 + 4d + 10 \leq 15d^2$ for all $d \geq 1$. Note that $C' = 15C$, where C is the absolute constant of Lemma 30. Combining Equation (11), Equation (12) and the fact that $\sum_{j=1}^k \alpha_j = \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f_{\alpha,\mathbf{W}}(\mathbf{x})]/\,\mathbf{E}_{t\sim\mathcal{N}}[\phi(t)] := B$ we obtain

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{N}^d}[\|\mathbf{x}\|_2^2 f_{\alpha,\mathbf{W}}(\mathbf{x}) \geq t] \leq k\exp(2 - (t^2/(4C'L^2B^2k^2d^2))^{1/3}) .$$

Define the random variables $y_i = \left\|\mathbf{x}^{(i)}\right\|_2^2 f_{\alpha,\mathbf{w}}(\mathbf{x}^{(i)})$. Set $S = O(kdBL)$ and $Q = O(S\log k \log^3 N)$ we have

$$\begin{aligned}\mathbf{E}\left[\max_{i\leq N} y_i\right] - Q &= \int_0^\infty \mathbf{Pr}\left[\max_{i\leq N} y_i \geq t + Q\right] \mathrm{d}t \leq Nk\int_0^\infty \mathbf{Pr}\left[y_1 \geq t + Q\right] \mathrm{d}t \\ &\leq Nk\int_0^\infty \exp(-(t/S)^{2/3})\mathrm{d}t = \widetilde{O}(dkBL)\end{aligned}$$

Now, that we have a bound on the expected maximum deviation we can apply Lemma 28 with $\mathbf{A} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)} f_{\alpha,\mathbf{W}}(\mathbf{x}^{(i)}) \in \mathbb{R}^{N\times d}$ and $m = \widetilde{O}(dkBL)$. Since $\|\mathbf{\Sigma}\|_2 \leq (1 + 1/\sqrt{2\pi})B$ we obtain that for $N = \widetilde{O}(dk/\epsilon^2)$ it holds $\mathbf{E}\left\|(1/N)\mathbf{A}^T\mathbf{A} - \mathbf{\Sigma}\right\|_2 \leq BL\epsilon$.

To finish the proof it remains to bound the norm of the sum $\frac{1}{N}\sum_{i=1}^N \xi^{(i)}\mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)}$. From Lemma 29 we obtain that it is bounded above by

$$C\sigma\sqrt{\log d} \mathop{\mathbf{E}}_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)}}\left[\max_{i\leq N}\left\|\mathbf{x}^{(i)}\right\|_2 \left\|\sum_{i=1}^N \mathbf{x}^{(i)} \otimes \mathbf{x}^{(i)}\right\|_2^{1/2}\right] .$$

We now use Cauchy-Schwarz for the above expectation and observe that

$$\mathop{\mathbf{E}}_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)}}\left[\max_{i\leq N}\left\|\mathbf{x}^{(i)}\right\|_2^2\right] \leq \widetilde{O}(d\log N) ,$$

which follows from Lemma 30 similarly as our previous bound. Moreover, from Lemma 28 we obtain that

$$\mathop{\mathbf{E}}_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)}}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}\otimes\mathbf{x}^{(i)}\right\|_2\right]\leq\widetilde{O}(\sqrt{d/N})\,.$$

Putting everything together we obtain that with $N=\widetilde{O}(d/\epsilon^2)$ samples, the expected norm of $\frac{1}{N}\sum_{i=1}^{N}\xi^{(i)}\mathbf{x}^{(i)}\otimes\mathbf{x}^{(i)}$ is at most $O(\sigma\epsilon)$. The result now follows from Markov's inequality. ■

Appendix C. Details of the Lower Bound

C.1. Preliminaries: Multilinear Algebra

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

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

C.2. Preliminaries: Hermite Polynomials

We are also going to use the Hermite polynomials that form a orthonormal system with respect to the Gaussian measure. We denote by $L^2(\mathbb{R}^d,\mathcal{N})$ the vector space of all functions $f:\mathbb{R}^d\to\mathbb{R}$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})]<\infty$. The usual inner product for this space is $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})g(\mathbf{x})]$. The L_2 norm of a function f is then defined as $\|f\|_2=\sqrt{\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f^2(\mathbf{x})]}$. While, usually one considers the probabilists's or physicists' Hermite polynomials, in this work we define the *normalized* Hermite polynomial of degree i to be $H_0(x)=1,H_1(x)=x,H_2(x)=\frac{x^2-1}{\sqrt{2}},\ldots,H_i(x)=\frac{He_i(x)}{\sqrt{i!}},\ldots$ where by $He_i(x)$ we denote the probabilists' Hermite polynomial of degree i. These normalized Hermite polynomials form a complete orthonormal basis for the single dimensional version of the inner product space defined above. To get an orthonormal basis for $L^2(\mathbb{R}^d,\mathcal{N}^d)$, we use a multi-index $J\in\mathbb{N}^d$ to define the d-variate normalized Hermite polynomial as $H_J(\mathbf{x})=\prod_{i=1}^{d}H_{v_i}(\mathbf{x}_i)$. The total degree of H_J is $|J|=\sum_{v_i\in J}v_i$. Given a function $f\in L^2$ we compute its Hermite coefficients as $\hat{f}(J)=\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})H_J(\mathbf{x})]$ and express it uniquely as $\sum_{J\in\mathbb{N}^d}\hat{f}(J)H_J(\mathbf{x})$. For more details on the Gaussian space and Hermite Analysis (especially from the theoretical computer

science perspective), we refer the reader to O'Donnell (2014). Most of the facts about Hermite polynomials that we use in this work are well known properties and can be found, for example, in Szegö (1967).

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

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

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

$$\nabla H_M(\mathbf{x}) = \sqrt{m_i} H_{M-E_i}(\mathbf{x}) \in \mathbb{R}^d, \tag{13}$$

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

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[\langle \nabla H_M(\mathbf{x}), \nabla H_L(\mathbf{x})\rangle] = |M|\delta_{M,L}. \tag{14}$$

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

Fact 31 *Let p, q be a harmonic polynomials of degree k. Then*

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

In particular,

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

Proof Write $p(\mathbf{x}) = \sum_{M:|M|=k} b_M H_M(\mathbf{x})$ and $q(\mathbf{x}) = \sum_{M:|M|=k} c_M H_M(\mathbf{x})$. Since the Hermite polynomials are orthonormal we obtain $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}}[p(\mathbf{x})q(\mathbf{x})] = \sum_{M:|M|=k} c_M b_M$. Now, using Equation 13 iteratively we obtain

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

Using this equality we obtain

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

■

Observe that for every harmonic polynomial $p(x)$ of degree k we have that $\nabla^k p(\mathbf{x})$ is a symmetric tensor of order k. Since the degree of the polynomial is k and we differentiate k times this tensor no longer depends on $\mathbf{x}$. Using Fact 31 we observe that this operation (modulo a division by $\sqrt{k!}$) preserves the L_2 norm of the harmonic polynomial p, that is $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[p^2(\mathbf{x})] = \left\|\nabla^k p(\mathbf{x})\right\|_2^2 / k!$.

Lemma 32 *Let $p(\mathbf{x}) : \mathbb{R}^2 \mapsto \mathbb{R}$ be a function and let $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{2\times d}$ be linear maps such that $\mathbf{U}\mathbf{U}^T = \mathbf{V}\mathbf{V}^T = \mathbf{I} \in \mathbb{R}^{2\times 2}$. Then, $\mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[p(\mathbf{U}\mathbf{x})p(\mathbf{V}\mathbf{x})] \leq \sum_{m=0}^{\infty} \left\|\mathbf{U}\mathbf{V}^T\right\|_2^m \mathbf{E}_{\mathbf{x}\sim\mathcal{N}^d}[(p^{[m]}(\mathbf{x}))^2]$.*

Proof To simplify notation write $f(\mathbf{x}) = p(\mathbf{U}\mathbf{x})$ and $g(\mathbf{x}) = p(\mathbf{V}\mathbf{x})$. The (total) degree of f is the same as the degree of p. Write $f(\mathbf{x}) = \sum_{m=0}^{\infty} f^{[m]}(\mathbf{x})$ and $g(\mathbf{x}) = \sum_{m=0}^{\infty} g^{[m]}(\mathbf{x})$. Then using Fact 31 we obtain

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f(\mathbf{x})g(\mathbf{x})] &= \sum_{m=0}^{\infty} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[f^{[m]}(\mathbf{x})g^{[m]}(\mathbf{x})] = \sum_{m=0}^{\infty} \frac{1}{m!}\left\langle \nabla^m f^{[m]}(\mathbf{x}), \nabla^m g^{[m]}(\mathbf{x})\right\rangle \\ &= \sum_{m=0}^{\infty} \frac{1}{m!}\left\langle \nabla^m p^{[m]}(\mathbf{U}\mathbf{x}), \nabla^m p^{[m]}(\mathbf{V}\mathbf{x})\right\rangle . \end{aligned} \tag{15}$$

Denote by $\mathcal{U} \subseteq \mathbb{R}^d$ the image of the linear map $\mathbf{U}^T$. Now observe that, using the chain rule, for any function $h(\mathbf{U}\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$ it holds $\nabla h(\mathbf{U}\mathbf{x}) = \partial_i h(\mathbf{U}\mathbf{x})\mathbf{U}_{ij} \ \in \mathcal{U}$, where we used Einstein's summation notation for repeated indices. Applying the above rule m-times we have that

$$\nabla h(\mathbf{U}\mathbf{x}) = \partial_{i_m} \dots \partial_{i_1} h(\mathbf{U}\mathbf{x})\mathbf{U}_{i_1 j_1} \dots \mathbf{U}_{i_m j_m} \ \in \mathcal{U}^{\otimes m} .$$

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

$$\langle \mathbf{S}, \mathbf{T}\rangle = \left\langle (\mathbf{U}^T)^{\otimes m}\mathbf{R}, (\mathbf{V}^T)^{\otimes m}\mathbf{R}\right\rangle = \left\langle \mathbf{R}, \mathbf{M}^{\otimes m}\mathbf{R}\right\rangle \leq \left\|\mathbf{M}^{\otimes m}\right\|_2 \|\mathbf{R}\|_2^2 = m!\,\|\mathbf{M}\|_2^m \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{N}^d}[(p^{[m]}(\mathbf{x}))^2] ,$$

where to get the last equality we used again Fact 31. To finish the proof we combine this inequality with Equation (15). ■

In the following simple lemma we prove that random 2-dimensional subspaces in high dimensions are roughly orthogonal.

Lemma 33 *For any $0 < c < 1/2$, there exists a set S of at least $2^{\Omega(d^c)}$ matrices in $\mathbb{R}^{2\times d}$ such that for each pair $\mathbf{A}, \mathbf{B} \in S$, it holds $\left\|\mathbf{A}\mathbf{B}^T\right\|_2 \leq O(d^{c-1/2})$.*

Proof We are going to use the following lemma.

Lemma 34 (Lemma 3.7 of Diakonikolas et al. (2017)) *For any $0 < c < 1/2$, there is a set S of at least $2^{\Omega(d^c)}$ unit vectors in $\mathbb{R}^d$ such that for each pair of distinct $\mathbf{u}, \mathbf{v} \in S$, it hold $|\langle \mathbf{u}, \mathbf{v}\rangle| \leq O(d^{c-1/2})$.*

Let matrices $\mathbf{A}_1, \dots \mathbf{A}_j$ in $\mathbb{R}^{2\times d}$, where $\mathbf{A}_i = \left(\mathbf{u}_{i,1}^T, \mathbf{u}_{i,2}^T\right)$, for some unit vectors $\mathbf{u}_{i,j}$ in $\mathbb{R}^d$. Then

$$\left\|\mathbf{A}_j\mathbf{A}_i^T\right\|_2 = \left\|\mathbf{A}_j^T\mathbf{A}_i\right\|_2 = \sqrt{\sum_{x,y=1}^{2}(\mathbf{u}_{i,x}^T\mathbf{u}_{j,y})^2} \leq 2\max_{\mathbf{u}_{i,x},\mathbf{u}_{j,y}} |\cos\theta(\mathbf{u}_{i,x},\mathbf{u}_{j,y})| \ .$$

From Lemma 34, it holds that there exists a set of $2^{\Omega(d^c)}$ of unit vectors such that $|\cos\theta(\mathbf{u},\mathbf{v})| \leq O(d^{c-1/2})$, taking this vectors as columns in each matrix the result follows. ■

Lemma 35 *Let $f_{\sigma,\phi} \in \mathcal{H}$. For every polynomial $p(\mathbf{x})$ of degree at most k, it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[f_{\sigma,\phi}(\mathbf{x}) \cdot p(\mathbf{x})] = 0$.*

Proof Let $\mathbf{w}^{(m)} = (\cos\frac{2\pi m}{2k}, \sin\frac{2\pi m}{2k})$ and $\alpha_m = (-1)^m$, for $m = 1, \dots, 2k$. Let $R_{\pi/k}$ be an operator over functions that rotates the coordinates by π/k (i.e., $(x,y) \mapsto (x\cos\frac{\pi}{k} + y\sin\frac{\pi}{k}, -x\sin\frac{\pi}{k} + y\cos\frac{\pi}{k})$). Then

$$\begin{aligned} R_{\pi/k}[f](x,y) &= f\big((x\cos\frac{\pi}{k} + y\sin\frac{\pi}{k}, -x\sin\frac{\pi}{k} + y\cos\frac{\pi}{k})\big) \\ &= \sigma\left(\sum_{m=1}^{2k-1}\alpha_m\phi\left(\left\langle\mathbf{x},\mathbf{w}^{(m+1)}\right\rangle\right) + \alpha_{2k}\phi\left(\left\langle\mathbf{x},\mathbf{w}^{(1)}\right\rangle\right)\right) \\ &= \sigma\left(\sum_{m=1}^{2k} -\alpha_m\phi\left(\left\langle\mathbf{x},\mathbf{w}^{(m)}\right\rangle\right)\right) = -f(x,y) \ , \end{aligned} \tag{16}$$

where to get the second equality we used that $\alpha_i\phi\left(\left\langle(x\cos\frac{\pi}{k} + y\sin\frac{\pi}{k}, -x\sin\frac{\pi}{k} + y\cos\frac{\pi}{k}), \mathbf{w}^{(i)}\right\rangle\right) = \alpha_i\phi\left(\left\langle(x,y),\mathbf{w}^{(i+1)}\right\rangle\right)$ from basic trigonometric identities and in the last one we used that σ is an odd function. Let $p(x,y) = (x+iy)^a(x-iy)^b$, where i is the imaginary unit, then we are going to prove that $\mathbf{E}_{\mathbf{x}\sim D}[f(\mathbf{x})p(\mathbf{x})] = 0$ as long as $a - b \not\equiv k \mod 2k$. We have

$$\begin{aligned} R_{\pi/k}[p](x,y) &= R_{\pi/k}[(x+iy)^a(x-iy)^b] = R_{\pi/k}[(x^2+y^2)^{a+b}e^{-i\theta(a-b)}] \\ &= (x^2+y^2)^{a+b}e^{i(\theta+\pi/k)(a-b)} = e^{i(\pi/k)(a-b)}p(x,y) \ , \end{aligned} \tag{17}$$

where θ is the argument (or the "phase") of $x+iy$. This means that $p(x,y)$ is an eigenfunction of $R_{\pi/k}$ and $e^{i(\pi/k)(a-b)}$ the corresponding eigenvalue. Thus, it holds

$$e^{i(\pi/k)(a-b)}\mathop{\mathbf{E}}_{\mathbf{x}\sim D}[f(\mathbf{x})p(\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim D}[f(\mathbf{x})R_{\pi/k}[p](\mathbf{x})] = \mathop{\mathbf{E}}_{\mathbf{x}\sim D}[R_{-\pi/k}[f](\mathbf{x})p(\mathbf{x})] = -\mathop{\mathbf{E}}_{\mathbf{x}\sim D}[f(\mathbf{x})p(\mathbf{x})] \ ,$$

where we used that $R_{\pi/k}$ is an adjoint operator in the inner product space of continuous functions along with Equations (16), (17). Thus, $\mathbf{E}_{\mathbf{x}\sim D}[f(\mathbf{x})p(\mathbf{x})] = 0$ when $e^{i(\pi/k)(a-b)} \neq -1$, which happens when $a - b \not\equiv k \mod 2k$. To conclude the proof, note that every polynomial at most degree k is a linear combination of the polynomials $p(x,y) = (x+iy)^a(x-iy)^b$ where $a, b \leq k$. This can be seen by setting $x = \frac{z+\bar{z}}{2}$ and $y = \frac{z-\bar{z}}{2i}$, where $z = x+iy$ and $\bar{z} = x-iy$. ■

C.3. Interpretation of the class $\mathcal{H}$

In order for the lower bound construction of Section 4 to produce useful lower bounds, it will be necessary that the function given in Lemma 18 is non-vanishing. It turns out that this is the case under fairly weak conditions. In order to state our final result, we will first introduce some terminology:

Definition 36 *For an integer k the k-parity-part of a function $\phi : \mathbb{R} \to \mathbb{R}$ is the odd part of ϕ if k is odd and the even part of ϕ if k is even.*

Definition 37 *For functions f on $\mathbb{R}^2$ define the operators R_k to be the rotation by π/k and define $S_k(f) = \sum_{s=1}^{2k}(-1)^s R_k^s(f)$.*

Given these we have the following result implying that $S_k\phi(x) \neq 0$ for a number of functions of interest. For example, if $\phi(x) = \max(0, x)$, is a ReLU, then the even part of ϕ is the absolute value function, so $S_k\phi \neq 0$ for any even k. Similarly, if ϕ is a sigmoid, $S_k\phi \neq 0$ for any odd k.

Proposition 38 *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a function with $\mathbf{E}[\phi^2(x)] < \infty$. Then $S_k\phi(x) = 0$ if and only if the k-parity-part of ϕ is a polynomial of degree less than k.*

Proof We begin by noting that $S_k\phi(x) \neq 0$ if and only if $(S_k\phi(x))^{[m]} \neq 0$ for some m. We note that as a rotation R_k preserves the degree-m Hermite parts of a function, and therefore so does S_k. In particular, $(S_k\phi(x))^{[m]} = (S_k\phi(x)^{[m]})$. In order to analyze this, we consider the variables $z = x + iy$ and $\bar{z} = x - iy$. We note that if $\phi(x)$ in one variable is given by the Hermite expansion $\phi(x) = \sum_{t=0}^{\infty} a_t h_t(x)$, that the two-variable version is given by $\sum_{t=0}^{\infty} a_t h_m((z+\bar{z})/2)$. Furthermore, we have that $(\phi(x))^{[m]} = a_m h_m((z+\bar{z})/2)$.

Now if $a_m = 0$, then $(\phi(x))^{[m]} = 0$ and therefore $S_k(\phi(x))^{[m]} = 0$. Otherwise, $a_m h_m(x)$ has non-vanishing x^t coefficients for all $t \leq m$ with $t \equiv m \pmod 2$. Therefore, in this case $(\phi(x))^{[m]}$ will have a non-vanishing $z^a\bar{z}^b$ coefficient for all $a, b \geq 0$ with $a+b \leq m$ and $a+b \equiv m \pmod 2$. Next, we need to understand what S_k does to $z^a\bar{z}^b$.

For this we note that $Rz = e^{\pi i/k}z$ and $R\bar{z} = e^{-\pi i/k}\bar{z}$. Thus $R(z^a\bar{z}^b) = e^{\pi i(a-b)/k}z^a\bar{z}^b$. Therefore,

$$S_k(z^a\bar{z}^b) = z^a\bar{z}^b \sum_{t=1}^{2k} e^{2\pi i(a-b+k)/(2k)} = \begin{cases} 2kz^a\bar{z}^b & \text{if } a-b \equiv k \pmod{2k} \\ 0 & \text{else} \end{cases}$$

Thus, $S_k(\phi(x))^{[m]}$ will be non-vanishing if and only if $a_m \neq 0$ and there are some $a, b \geq 0$ with $a+b \leq m, a+b \equiv m \pmod 2$ and $a-b \equiv k \pmod{2k}$. We claim that such a, b exist if and only if $m \equiv k \pmod 2$ and $m \geq k$. The only if part of this condition is clear. For the if part, we note that if these conditions are satisfied, we may take $a = \frac{m+k}{2}$ and $b = \frac{m-k}{2}$.

Therefore, we have that $S_k\phi(x) \neq 0$ if and only if there is some $m \equiv k \pmod 2$ with $a_m \neq 0$ and $m \geq k$. Note that the k-parity-part of ϕ has the same Hermite coefficients as ϕ for $m \equiv k \pmod 2$ and 0 coefficient for $m \not\equiv k \pmod 2$. Thus, ϕ has a non-vanishing coefficient for some $m \geq k, m \equiv k \pmod 2$ if and only if the k-parity-part of ϕ has some non-vanishing coefficient of degree $m \geq k$. Of course this happens if and only if the k-parity-part of ϕ is not a polynomial with degree less than k. This completes our proof. ■