Toward L_∞ Recovery of Nonlinear Functions: A Polynomial Sample Complexity Bound for Gaussian Random Fields

Kefan Dong KEFANDONG@STANFORD.EDU
Stanford University

Tengyu Ma TENGYUMA@STANFORD.EDU
Stanford University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Many machine learning applications require learning a function with a small worst-case error over the entire input domain, that is, the L_∞-error, whereas most existing theoretical works only guarantee recovery in average errors such as the L_2-error. L_∞-recovery from polynomial samples is even impossible for seemingly simple function classes such as constant-norm infinite-width two-layer neural nets. This paper makes some initial steps beyond the impossibility results by leveraging the randomness in the ground-truth functions. We prove a polynomial sample complexity bound for random ground-truth functions drawn from Gaussian random fields. Our key technical novelty is to prove that the degree-k spherical harmonics components of a function from Gaussian random field cannot be spiky in that their L_∞/L_2 ratios are upperbounded by $O(d\sqrt{\ln k})$ with high probability. In contrast, the worst-case L_∞/L_2 ratio for degree-k spherical harmonics is on the order of $\Omega(\min\{d^{k/2}, k^{d/2}\})$.

Keywords: L_∞ recovery, kernel method, Gaussian random fields, spherical harmonics

1. Introduction

Classical statistical learning theory primarily concerns with recovering functions from examples with small errors *averaged* over a distribution of inputs, e.g., the mean-squared error (that is, the L_2-error with respect to the test distribution). However, the worst-case error over the entire input domain, that is, the L_∞-error, is crucial for many applications, and also challenging to achieve. For example, an L_∞-error recovery guarantee will make the learned function more robust to adversarial examples, while standard training is vulnerable (Goodfellow et al., 2015; Madry et al., 2017). The L_∞-recovery is also necessary for many applications where the recovered models will be further used in a downstream decision making process, such as model-based bandits (Huang et al., 2021b), reinforcement learning (Huang et al., 2021a; Sutton and Barto, 2018), and physics informed neural networks (Raissi et al., 2019; Wang et al., 2022). In particular, recent theoretical works on deep reinforcement learning heavily rely on the L_∞-recovery of the Q-function to prevent the actions from misusing a small worst-case region of inputs where the error is much larger than the average error (Huang et al., 2021a). (See Section 2 for more discussions on the applications.)

This paper focuses on L_∞-error recovery of nonlinear functions from polynomial samples. For a compact domain D, we aim to learn a function that is *pointwise* close to the ground-truth over the entire domain D. Formally, given polynomial random input-output pairs from a ground-truth function f, our goal is to learn a function g with small L_∞-distance/error to f, defined by

$$\|f-g\|_\infty \triangleq \sup_{x\in D} |f(x)-g(x)|.$$

For linear function class, we can straightforwardly L_∞-recovery guarantees by relating the L_∞-error to the L_2-error on the test distribution, which in turn can be bounded standard tools such as uniform convergence (e.g., Bartlett and Mendelson (2002); Koltchinskii and Panchenko (2002); Wei and Ma (2019)). Concretely, suppose P is the training/test distribution on domain D and the covariance matrix $\Sigma = \mathbb{E}_{x\sim P}[xx^\top]$ is full-rank, we have for any linear functions f and g,

$$\|f-g\|_\infty \le (\sup_{x\in D}\|x\|_2)\cdot \lambda_{\min}(\Sigma)^{-1/2}\cdot \|f-g\|_{L_2(P)}\,. \tag{1}$$

where $\lambda_{\min}(\Sigma)$ is the minimum eigenvalue of Σ and $\|f-g\|_{L_2(P)} = \left(\mathbb{E}_{x\sim P}(f(x)-g(x))^2\right)^{1/2}$ is the squared error on the distribution P. Therefore, we can reduce the L_∞-recovery to L_2-recovery, and the inequality above is tight for most scenarios.

In contrast, L_∞-recovery of nonlinear functions is much more challenging. When the model's parameters are identifiable and can be recovered, e.g., for finite-width two-layer neural nets (without biases) (Zhong et al., 2017; Zhou et al., 2021) or low-degree polynomials (Huang et al., 2021a), L_∞-recovery of the functions follows straightforwardly from parameter recovery. Parameter recovery fundamentally requires the sample size to be larger than parameter dimension, and therefore does not apply to the over-parameterized settings that are ubiquitous in modern machine learning (Zagoruyko and Komodakis (2016); Du et al. (2018); Allen-Zhu et al. (2019); Zhang et al. (2021a) and references therein) or infinite dimensional features in the kernel method settings. Existing L_∞-recovery algorithms require quasi-polynomial or exponential in dimension samples for two-layer neural networks (Mhaskar, 2006, 2019) or general very smooth functions (with decaying higher-order derivatives) (Vybíral, 2014; Krieg, 2019) .

In fact, L_∞-recovery from polynomial samples is impossible for even seemingly simple function classes, such as two-layer single-neuron neural nets with bias (Dong et al., 2021; Li et al., 2021) or constant-norm infinite-width two-layer neural nets without bias (Theorem 8 of this paper). The fundamental challenge is that these function classes contain many *spiky* functions f such that $\|f\|_2 \ll \|f\|_\infty$, which means that inequalities analogous to Eq. (1) cannot hold. Moreover, these functions may mostly have tiny values except a spike on an exponentially-small region. Likely, none of the polynomial number of examples falls into the spiky region. As a result, the spike cannot be identified, and L_∞-recovery cannot be achieved.

Interestingly, L_∞-recovery of functions in reproducing kernel Hilbert space (RKHS) with polynomial samples is still a challenging open question, even though L_2-recovery with polynomial samples and time has been well established (Bartlett and Mendelson, 2002; Hofmann et al., 2008). Even though they are essentially linear functions with an infinite dimensional features, analysis analogous to linear models (e.g., Eq (1)) is vacuous because the covariance of the features, that is the kernel function, typically has a sequence of eigenvalues that decays to zero. In fact, L_∞-recovery of functions with constant RKHS norm for various kernels (e.g., the radial basis function (RBF) kernels kernel) requires exponential number of samples (Scarlett et al., 2017; Kuo et al., 2008). Intuitively, this is because RKHS still contains spiky functions, e.g., the k-th eigenfunctions of the RBF kernel (or any inner product kernel) on the unit sphere can be spiky for relatively large k.

Towards going beyond these intractability results and achieve polynomial sample complexity bounds, we make additional randomized and smooth assumptions on the ground-truth functions that we aim to recover. We essentially assume that the ground-truth function has decaying and random high-frequency components.

Concretely, we work with random ground-truth functions f drawn from a Gaussian random field (GRF, also known as Gaussian process) on the unit sphere (Seeger, 2004; Lang and Schwab, 2015). We assume that the covariance (or kernel) function, denoted by $K : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$, is an inner product function given by $K(x, x') = \kappa(x^\top x')$ for some function $\kappa : [-1, 1] \to \mathbb{R}$, which means that the GRF is isotropic.

All the inner product kernels $\kappa(x^\top x')$ on the unit sphere share the same eigenfunctions called spherical harmonics (Atkinson and Han, 2012). This brings opportunities for us to use the spherical harmonics tools to analyze the problem. Spherical harmonics form a complete set of basis for the square integrable functions over the sphere (in analog to the Fourier basis in $\mathbb{R}^d$). Intuitively, fast decay of the spherical harmonics components of the function implies the function is smoother. Moreover, some higher-degree spherical harmonics can be more spiky and challenging to recover in L_∞-error.

Our main result is an L_∞-recovery algorithm (Alg. 1) with polynomial sample complexity for a random ground-truth f drawn from Gaussian random fields, given that the k-th eigenvalues of the covariance function K decays at a rate $O(k^{-(1+\alpha)d})$ for any universal constant $\alpha > 0$ (Theorem 3). This decay rate is equivalent to that the degree-k spherical harmonics component of f is on the order $O(k^{-\alpha d/2})$. We note that the randomness from the Gaussian random field is the key for us to work with this decay (that is, $\alpha > 0$), because a worst-case function with $O(k^{-d/2})$ decay in the spherical harmonics components is impossible to recover with polynomial samples (Lemma 7). This suggest that the randomness property in the ground-truth function makes the L_∞-recovery much easier. Moreover, for comparison, L_2-recovery with polynomial sample is only possible when $\alpha \geq 0$.[1] In other words, the randomness assumption qualitatively makes L_∞-recovery as easy as L_2-recovery.

Our main technique is to prove a much tighter upper bound for the L_∞-norm of the high-degree components of the ground-truth f using the randomness. Lemma 5 shows that the high-degree components of f drawn from Gaussian random processes are *not* spiky: their L_∞/L_2 ratios are upper bounded by $O(d\sqrt{\ln k})$ with high probability, whereas the worst-case ratio is $\Omega(\min\{d^{k/2}, k^{d/2}\})$. This lemma might be of independent interest as it extends Burq and Lebeau (2014, Theorem 1) to the high-dimensional case with a precise bound on the dependency on d. It was not known that the dependency on d is polynomial. Our proof is also surprisingly much simpler than that in Burq and Lebeau (2014). Hence, when the eigenvalues of kernel K decays, we can truncate f at degree $\widetilde{O}(1)$ and get a low-degree polynomial approximation with small L_∞-error (Lemma 10).

The rest of this paper is organized as follows. Section 2 discusses additional related works and the applications of L_∞-recovery to bandits and reinforcement learning problems. In Section 3 we give a concise overview of the spherical harmonics, the important tools

[1] The $\alpha = 0$ case is subtle for both L_2- and L_∞-recovery and we leave it as an open question for future work.

in this paper. Section 4 states our algorithm and proves a polynomial sample complexity bound for recovering random functions from a Gaussian random field. Section 5 proves that L_∞-recovery is impossible for two-layer neural nets without bias (Theorem 8), which may be of independent interest.

Additional notations. Let $\mathbb{S}^{d-1}$ be the $(d-1)$-dimensional unit sphere. We assume that the training distribution is uniform over the $\mathbb{S}^{d-1}$. That is, the data x_i is sampled independently and uniformly from the sphere, and $y_i = f(x_i) + \mathcal{N}(0,1)$ where y is the ground-truth. With slight abuse of notations, we also use $\mathbb{S}^{d-1}$ to denote the uniform distribution over the unit sphere. For a function $g : \mathbb{S}^{d-1} \to \mathbb{R}$, let $\|g\|_p \triangleq \mathbb{E}_{x\sim\mathbb{S}^{d-1}}[g(x)^p]^{1/p}$ be its L_p-norm with respect to the uniform distribution. For two functions $g, h : \mathbb{S}^{d-1} \to \mathbb{R}$, $\langle g, h\rangle \triangleq \mathbb{E}_{x\sim\mathbb{S}^{d-1}}[h(x)g(x)]$ denotes their inner product. For a function $h : \mathbb{R} \to \mathbb{R}$, we use $h^{(k)}$ to denote its k-th derivative.

In the following, for two non-negative sequences a_k, b_k, we write $a_k = O(b_k)$ or $a_k \lesssim b_k$ if there exists an *absolute* constant c such that $a_k \le cb_k$ for every $k \ge 0$. We write $a_k = \widetilde{O}(1)$ if $a_k = O(\text{polylog}(k))$, and $a_k = \Theta(b_k)$ if $a_k \lesssim b_k$ and $b_k \lesssim a_k$.

2. Related Works

The classical uniform convergence framework (e.g., Bartlett and Mendelson (2002); Koltchinskii and Panchenko (2002); Kakade et al. (2009); Bartlett et al. (2017); Wei and Ma (2019)) does not directly solve the L_∞-recovery problem. This is because for any $p > 0$, approximating the L_p-error (defined by $\|f-g\|_p \triangleq \mathbb{E}_{x\sim D}[|f(x)-g(x)|^p]^{1/p}$) with ϵ precision requires $\text{poly}(\epsilon^{-p})$ samples. Hence, as $p \to \infty$, we cannot avhieve uniform convergence using polynomial samples, meaning that bounding the L_∞-error of the learned function requires novel analysis.

Gaussian process bandits and kernelized bandits. A closely related line of research is the Gaussian process bandits. Instead of learning a function with small L_∞-error, Gaussian process bandit algorithms aim to find a x that maximizes the function $f(x)$ when f is drawn from a Gaussian process (Grünewälder et al., 2010). Most of the existing results focuses on radial basis function kernel and Matérn kernels, and the regret is exponential in the ambient dimension d (Srinivas et al., 2009; Krause and Ong, 2011; Shekhar and Javidi, 2018; Vakili et al., 2021b). For Gaussian processes with non-isotropic kernels, Grünewälder et al. (2010) prove exponential regret upper and lower bounds. For general kernels, Lederer et al. (2019, 2021) proves a L_∞-error bounds for Gaussian process regression with no assumption on the spectrum of the covariance, and the sample complexity is also exponential.

Another line of research focuses on kernelized bandits and assumes that ground-truth f has a small RKHS norm (see Valko et al. (2013); Wang and de Freitas (2014); Chowdhury and Gopalan (2017); Vakili et al. (2021a); Zhang et al. (2021b) and references therein). For the RBF and Matérn kernels, their regret bounds are exponential in d and Scarlett et al. (2017) prove that no algorithm can achieve polynomial sample complexities. Since a small RKHS norm does not exclude spiky functions in general, the results in this setting requires a stronger assumption on ground-truth f. In fact, a function drawn from a Gaussian process has a infinite RKHS norm (defined by the same kernel) almost surely (Wahba, 1990).

Neural nets recovery. The parameters of finite-width two-layer neural networks can be recoverd with additional assumptions on the correlation between neurons (Zhong et al., 2017; Fu et al., 2020; Zhou et al., 2021), or the condition number of the first-layer weights (Zhang et al., 2019). For two-layer neural networks with unbiased ReLU activation, Bakshi et al. (2019) design algorithms whose sample complexity scales exponentially in the number of hidden neurons. In addition, Milli et al. (2019) proves a recovery guarantee of two-layer ReLU neural networks when the algorithm can query the gradient of the ground-truth. These methods cannot be applied to infinite-width neural networks, and L_∞-recovery of a infinite-width neural network requires exponential samples (Theorem 8).

Applications to bandits, reinforcement learning, and PINN. The best-arm identification problem in nonlinear bandits can be reduced to an L_∞-recovery problem. If we can learn a function g that approximates the true reward f with $\|f-g\|_\infty \le \epsilon/2$, the action $\hat{x} \triangleq \operatorname{argmax}_{x\in D} g(x)$ is ϵ-optimal. Similarly, if the Q-function can be learned with a small L_∞-error for finite horizon reinforcement learning, we can guarantee the optimality of the learned policy (Huang et al., 2021a).

For physics informed neural networks (Raissi et al., 2019), minimizing the L_2 loss may not be satisfactory (Wang et al., 2022; Krishnapriyan et al., 2021; Wang et al., 2021). When learning the Hamilton-Jacobi-Bellman equations (an analog of Bellman equations for continuous time), Wang et al. (2022) proves that a small L_∞-error can guarantee a good final performance, while a small L_2-error cannot.

L_∞-recovery for other nonlinear functions. Several other related works study the L_∞-recovery with different assumptions on the ground-truth function. Bertin (2004b); Korostelev (1994); Tsybakov (1998); Golubev et al. (2000) study the minimax rate for L_∞-recovering for one-dimensional smooth functions (e.g., functions in Hölder, Sobolev, or Besov classes). For general functions with bounded or decaying high-order derivatives, Vybíral (2014); Krieg (2019) design estimators with quasi-polynomial or exponential in dimension samples. Ibragimov and Khas' minskii (1984); Stone (1982); Nyssbaum (1987); Bertin (2004a) determine the asymptotically optimal rate for L_p-recovery of Hölder smooth functions for general $p \in [1,\infty]$ in high dimensions.

When the ground-truth function lies in the reproducing kernel Hilbert space, Kuo et al. (2009) prove some sufficient conditions for L_∞-recovery with polynomial sample complexity. In general, L_∞-recovery with polynomial samples is impossible unless the eigenvalues of the kernel decay very fast (Long and Han, 2023; Kuo et al., 2008). We refer the readers to Ebert and Pillichshammer (2021) for a comprehensive survey in this direction.

Another line of research focuses on learning a nonlinear function with respect to the Sobolev norm (Fischer and Steinwart, 2020; Steinwart et al., 2009). While their analysis can lead to L_∞-recovery bounds, they require stronger smoothness assumptions to exclude the worst-case hard instances shown in Lemma 7. In contrast, our algorithms achieve L_∞-recovery in the average case using much weaker smoothness assumptions.

3. Preliminaries on Spherical Harmonics

Now we give a brief overview of spherical harmonics, the essential tools in this paper, based on Atkinson and Han (2012, Section 2). Spherical harmonics are the eigenfunctions of the

Laplacian operator on the sphere. The eigenfunctions corresponding to the k-th eigenvalue are degree-k polynomials, and form a Hilbert space denoted by $\mathbb{Y}_{k,d}$. The dimension of $\mathbb{Y}_{k,d}$ is $N_{k,d} \triangleq \binom{d+k-1}{d-1} - \binom{d+k-3}{d-1}$. When $k \to \infty$, $N_{k,d} = \Theta(d^k)$ and when $d \to \infty$, $N_{k,d} = \Theta(k^d)$. Spherical harmonics with different degrees are orthogonal to each other, and their linear combinations can represent all square integrable functions over the sphere.

We use Π_k to denote the projection operator to the degree-k spherical harmonics space $\mathbb{Y}_{k,d}$. We use $\mathbb{Y}_{\le k,d}$ to denote the space of spherical harmonics up to degree k, and $\Pi_{\le k} \triangleq \sum_{l=0}^{k} \Pi_l$ the projection operator to $\mathbb{Y}_{\le k,d}$.

Spherical harmonics are closely related to Legendre polynomials. The degree-k Legendre polynomial $P_{k,d} : \mathbb{R} \to \mathbb{R}$ is defined by the following recursive relationship

$$P_{0,d}(t) = 1, \quad P_{1,d}(t) = t, \tag{2}$$

$$P_{k,d}(t) = \frac{2k+d-4}{k+d-3} t P_{k-1,d}(t) - \frac{k-1}{k+d-3} P_{k-2,d}(t), \quad \forall k \ge 2. \tag{3}$$

Let $\bar{P}_{k,d}(t) \triangleq \sqrt{N_{k,d}} P_{k,d}(t)$ be the normalized Legendre polynomial. Normalized Legendre polynomial is a set of complete orthonormal basis for square-integrable functions over $[-1,1]$ with respect to the measure $\mu_d(t) \triangleq (1-t^2)^{\frac{d-3}{2}} \frac{\Gamma(d/2)}{\Gamma((d-1)/2)} \frac{1}{\sqrt{\pi}}$, which equals to the density of x_1 when $x = (x_1, \cdots, x_d)$ is uniformly drawn from sphere $\mathbb{S}^{d-1}$. In other words, $\left\langle \bar{P}_{k,d}, \bar{P}_{k',d} \right\rangle_{\mu_d} \triangleq \int_{-1}^{1} \bar{P}_{k,d}(t) \bar{P}_{k',d}(t) \mu_d(t) \mathrm{d}t = \mathbb{I}\left[k = k'\right]$.

Properties of spherical harmonics and Legendre polynomials. Our proof heavily replies on the following properties of spherical harmonics and Legendre polynomials.

Let $\{Y_{k,j}\}_{j=1}^{N_{k,d}}$ be an orthonormal basis of $\mathbb{Y}_{k,d}$. Then for any function $f : \mathbb{S}^{d-1} \to \mathbb{R}$ with $\|f\|_2 < \infty$, there is a unique decomposition $f(\cdot) = \sum_{k\ge 0} \sum_{j=1}^{N_{k,d}} a_{k,j} Y_{k,j}(\cdot)$ with coefficients $\{a_{k,j}\}_{k \ge 0, 1 \le j \le N_{k,d}}$ that satisfies $\|f\|_2^2 = \sum_{k \ge 0} \sum_{j=1}^{N_{k,d}} a_{k,j}^2$.

Spherical harmonics are the eigenfunctions of any inner-product kernels on the sphere, summarized by the following theorem (Atkinson and Han, 2012, Theorem 2.22).

Theorem 1 (Funk-Hecke formula) *Let $\sigma : [-1,1] \to \mathbb{R}$ be any one-dimensional function with $\int_{-1}^{1} |\sigma(t)| \mu_d(t) \mathrm{d}t < \infty$, and $\lambda_k = N_{k,d}^{-1/2} \left\langle \sigma, \bar{P}_{k,d} \right\rangle_{\mu_d}$. Then for any function $Y_k \in \mathbb{Y}_{k,d}$,*

$$\forall x \in \mathbb{S}^{d-1}, \quad \mathbb{E}_{z \sim \mathbb{S}^{d-1}}[\sigma(x^\top z) Y_k(z)] = \lambda_k Y_k(x). \tag{4}$$

In other words, $\mathbb{Y}_{k,d}$ is the space of eigenfunctions of the inner product kernel $K(x,z) \triangleq \sigma(x^\top z)$ corresponding to the eigenvalue λ_k.

We can construct spherical harmonics using Legendre polynomials. For any degree $k \ge 0$, let $g_u : \mathbb{S}^{d-1} \to \mathbb{R}$ be the function $g_u(x) = \bar{P}_{k,d}(\langle x, u \rangle)$. Then for any $u \in \mathbb{S}^{d-1}$, $g_u \in \mathbb{Y}_{k,d}$ and $\|g_u\|_2 = 1$.

In the worst-case, high-order spherical harmonics can be very spiky because their L_∞ / L_2 ratio is very large:

Fact 2 *For every fixed $k \ge 0, g \in \mathbb{Y}_{k,d}$ we have $\|g\|_\infty \le \sqrt{N_{k,d}} \|g\|_2$, and the equality is achieved by $g_u(\cdot) = \bar{P}_{k,d}(\langle \cdot, u \rangle)$ for any $u \in \mathbb{S}^{d-1}$.*

4. Main Results

In this section, we will first design a L_∞-recovery algorithm that achieves polynomial sample complexity when the ground-truth function f satisfies two conditions (Conditions 1 and 2). We then establish these two conditions when f is drawn from an isotropic Gaussian random fields (Lemma 5).

The first condition states that the spherical harmonics decomposition of the ground-truth f decays at a proper rate.

Condition 1 *The ground-truth function f satisfies $\|\Pi_k f\|_2 \le c_1 N_{k,d}^{-\alpha/2}, \forall k \ge 0$ for some $c_1 > 0$ and $\alpha > 0$.*

We treat α as a constant that doesn't depend on the ambient dimension d. The parameter $\alpha > 0$ is intuitively a notion of smoothness of the function f. This is because the derivatives of higher-degree spherical harmonics are larger. Hence, qualitatively speaking, functions with a faster decay (larger α) is smoother.

Condition 1 holds for a wide range of functions. For example, any function of the form $g(\cdot) = h(\langle \cdot, u\rangle)$, where $u \in \mathbb{S}^{d-1}$ and $h : [-1,1] \to \mathbb{R}$ with $\sup_{t\in[-1,1]} |h^{(k)}(t)| \le 1, \forall k \ge 0$, satisfies Condition 1 with parameter $\alpha = 1$ and $c_1 = 1$ (Proposition 13). In addition, if two functions g, h satisfy Condition 1, so do their convex combinations $\theta g + (1-\theta)h, \forall \theta \in [0,1]$. Hence Condition 1 holds for *any* two-layer NNs with bounded L_1 norm and infinitely smooth activation h (e.g., exponential activation).

The following condition states that f is not spiky when projected to the degree-k spherical harmonics space. This condition is central to our analysis because it excludes the hard instances in the lower bounds (e.g., spiky functions constructed in Lemma 7).

Condition 2 *The ground-truth function f satisfies $\|\Pi_k f\|_\infty \le c_2 \sqrt{\ln(k+1)} \|\Pi_k f\|_2, \forall k \ge 0$ for some $c_2 > 0$.*

Condition 2 requires that the L_∞/L_2 ratio of $\Pi_k f$ is bounded by $c_2\sqrt{\ln(k+1)}$, whereas the worst case ratio is $\sqrt{N_{k,d}}$ (Fact 2). As we will show later, Condition 2 holds with high probability for random functions drawn from degree-k spherical harmonics space $\mathbb{Y}_{k,d}$ (Lemma 6) and functions drawn from isotropic Gaussian random fields (Lemma 5).

With Conditions 1 and 2, our main theorem states that there exists an algorithm (described later in Alg. 1) that achieves L_∞-recovery using only polynomial samples drawn from the uniform distribution over the sphere $\mathbb{S}^{d-1}$.

Theorem 3 *Suppose the ground-truth function f satisfies Conditions 1 and 2 for some fixed $\alpha \in (0,1]$ and $c_1, c_2 > 0$. If $d \ge 10\alpha^{-1}2^{5/\alpha} + 2$, then for any $\epsilon > 0, \delta > 0$, with probability at least $1-\delta$ over the randomness of the data, Alg. 1 outputs a function g such that $\|f - g\|_\infty \le \epsilon$ using $O(\mathrm{poly}(c_1c_2, d, 1/\epsilon, \ln 1/\delta)^{1/\alpha})$ samples.*

In comparison, the classical kernel methods assume f has a bounded RKHS norm, which is equivalent to assuming that $\|\Pi_k f\|_2$ decays at a rate determined by the choice of kernel. For any function f with decay parameter $\alpha \in (0, 1/2)$, Proposition 15 shows that the RKHS norm of f is infinite with respect to *any* bounded inner product kernel (e.g., the RBF kernel), and thus violates the assumption of kernel methods. In contrast, Theorem 3 still implies polynomial sample complexity for $\alpha \in (0, 1/2)$ thanks to the additional randomness condition (Condition 2)

Our algorithm is stated in Alg. 1. On a high level, given any desired error $\epsilon > 0$, the algorithm selects a truncation threshold $k \geq 0$ (Line 1), and uses empirical risk minimization to find the best degree-k polynomial approximation to the ground-truth f.

Instead of directly learning a degree-k polynomial, we can also use two-layer neural networks with polynomial activation to approximate the function f. The algorithm and discussion are deferred to Appendix B.

```
Algorithm 1 $L_\infty$-learning via Low-degree Polynomial Approximation
    Parameters: $\alpha, c_1, c_2 > 0$, desired error $\epsilon > 0$, and failure probability $\delta > 0$.
    Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \sim \mathbb{S}^{d-1}$ are i.i.d. samples from the unit sphere, and $y_i = f(x_i) + \mathcal{N}(0, 1)$.
1: Set the truncation threshold $k \leftarrow \inf_{l \geq 0}\{2c_1c_2(l+1)^{3/2}(N_{l+1,d})^{-\alpha/2} \leq \epsilon/2\}$.
2: Define the function class
       $$\mathcal{F}_k \leftarrow \{g \in \mathbb{Y}_{\leq k,d} : \|\Pi_l g\|_2 \leq c_1, \forall l \in [0, k]\}. \quad (5)$$
3: Run empirical risk minimization: $g \leftarrow \operatorname{argmin}_{h \in \mathcal{F}_k} \sum_{i=1}^n (h(x_i) - y_i)^2$.
4: Return $g$.
```

We present a proof sketch of Theorem 3 in Section 4.2, and defer the full proof to Appendix A.1.

4.1. Instantiation of Theorem 3 on Gaussian Random Fields

In this section, we instantiate Theorem 3 on isotropic Gaussian random fields.

Given any positive semi-definite covariance function $K : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$, the mean-zero Gaussian random field is a collection of random variables $\{h(x)\}_{x \in \mathbb{S}^{d-1}}$ such that the distribution of any finite subset $(h(x_1), \cdots, h(x_n))$ is a Gaussian vector with covariance $\Sigma_{ij} = K(x_i, x_j)$. When the distribution is rotationally invariant, i.e., the distribution of $h(x_1), \cdots, h(x_n)$ equals to the distribution of $h(Rx_1), \cdots, h(Rx_n)$ for any rotation matrix $R \in \mathbb{R}^{d \times d}$, the covariance $K(x, x')$ only depends on the inner product $x^\top x'$ and can be written as $K(x, x') = \kappa(x^\top x')$ for some $\kappa : [-1, 1] \to \mathbb{R}$. The corresponding GRF is called isotropic.

We focus on the case where the eigenvalues of the covariance (or equivalently, the Legendre polynomial decomposition of κ, by the Funk-Hecke formula) decays with a proper rate. Concretely, we assume κ has the decomposition $\kappa(t) = \sum_{k \geq 0} \hat{\kappa}_k \bar{P}_{k,d}(t)$ where $\hat{\kappa}_k \leq O(N_{k,d}^{-1/2-\alpha})$ for some $\alpha > 0$. Later we will show that a function f drawn from GRF with covariance $K(x, x') = \kappa(x^\top x')$ satisfies $\|\Pi_k f\|_2 \leq O(N_{k,d}^{-\alpha/2})$ and this inequality is tight. The decay rate $O(N_{k,d}^{-1/2-\alpha})$ is slightly faster than the decay of RBF kernels (given by $\kappa(t) = \exp(t)$), which is $\approx N_{k,d}^{-1/2}$ when k is small (Minh et al., 2006).

The following theorem proves that Alg. 1 can achieve L_∞-recovery for function drawn from Gaussian random fields.

Theorem 4 *Let $f : \mathbb{S}^{d-1} \to \mathbb{R}$ be a function drawn from a Gaussian random field with covariance $K(x, x') = \kappa(x^\top x)$. Suppose for all $k \geq 0$, $\langle \kappa, \bar{P}_{k,d} \rangle_{\mu_d} \leq c^2 N_{k,d}^{-1/2-\alpha}$ for some $c > 0, \alpha > 0$. Given any $\epsilon > 0, \delta > 0$, with probability at least $1 - \delta$ over the randomness of f and the dataset, Alg. 1 outputs a function $g : \mathbb{S}^{d-1} \to \mathbb{R}$ such that $\|g - f\|_\infty \leq \epsilon$ using $O(\mathrm{poly}(c, \epsilon^{-1}, d, \ln 1/\delta)^{1/\alpha})$ samples.*

To the best of our knowledge, Theorem 4 is the first result that achieves a L_∞-error guarantee for isotropic Gaussian processes using only polynomial samples drawn uniformly from the unit sphere.

A closely related line of research to Theorem 4 is the Gaussian process bandit problem, where the algorithm can adaptively query any data point and the goal is to maximize the function f drawn from a Gaussian random field (Srinivas et al., 2010). We can modify the GP-UCB algorithm in Srinivas et al. (2010) to a L_∞-recovery algorithm with *adaptive* samples, and this modification, together with the analysis in Vakili et al. (2021a), lead to a polynomial sample complexity with the same condition as Theorem 4.[2] In comparison, our algorithm only requires samples from the uniform distribution while GP-UCB must be adaptive. In addition, Theorem 3 holds for general functions with Condition 1 and 2 while the analysis of Srinivas et al. (2010) is specialized to Gaussian processes.

We prove Theorem 4 by establishing Conditions 1 and 2 for functions drawn from isotropic Gaussian random fields using the following lemma, and then directly invoking Theorem 3.

Lemma 5 *In the setting of Theorem 4, with probability at least $1 - \delta$ we have*

$$\forall k \geq 0, \quad \|\Pi_k f\|_\infty \leq 5\sqrt{2\ln(6/\delta) + 2(d^2+1)\ln(k+1)}\|\Pi_k f\|_2, \tag{6}$$

and

$$\forall k \geq 0, \quad \|\Pi_k f\|_2 \leq 3c\sqrt{\ln(2/\delta)} N_{k,d}^{-\alpha/2}. \tag{7}$$

We the proof of Lemma 5 is deferred to Section 4.3.

4.2. Proof Sketch of Theorem 3

In this section, we present the proof sketches of Theorem 3. On a high level, we prove that (a) the ground-truth f can be approximated by a low-degree polynomial with a small L_∞-error, and (b) learning a low-degree polynomial in L_∞-error only requires polynomial samples.

[2] On a high level, at every iteration $t \geq 1$ the original GP-UCB algorithm selects the query x_t that maximizes the upper confidence bound of f (Srinivas et al., 2010). Srinivas et al. (2010) construct the upper confidence bound by analytically compute the posterior mean and variance of f given any data points, assuming that the ground-truth f is drawn from a Gaussian process prior. To get an algorithm for L_∞-recovery, we can choose x_t that maximizes the posterior variance of $f(\cdot)$. In this case, the analysis in (Srinivas et al., 2010) implies that with high probability after n iterations, the L_∞-error of the posterior mean is upper bounded by the maximum information gain, denoted by $\sqrt{\gamma_n/n}$. Combining with the refined analysis in Vakili et al. (2021a), we can upper bound the information gain γ_n using the spectrum decay of κ, which leads to a polynomial sample complexity in our setting.

Proof sketch of Theorem 3 For better exposition, in the following we present the proof sketch for the case $\alpha = 1/2$, and the general case is proved similarly.

For any fixed threshold $k \geq 0$, we first upper bound the L_∞-distance between the ground-truth f and its low-degree components $\Pi_{\leq k} f$. Concretely,

$$\|f - \Pi_{\leq k} f\|_\infty = \| \textstyle\sum_{l>k} \Pi_l f\|_\infty \leq \sum_{l>k} \|\Pi_l f\|_\infty. \tag{8}$$

Under Conditions 1 and 2, the term $\|\Pi_l f\|_\infty$ decays at rate $\sqrt{\ln(l+1)} N_{l,d}^{-1/2}$. Since $N_{l,d}^{-1/2} \approx \min\{l^d, d^l\}^{-1/2}$ decays very fast, we get

$$\|f - \Pi_{\leq k} f\|_\infty \leq \textstyle\sum_{l>k} \sqrt{\ln(l+1)} N_{l,d}^{-1/2} \lesssim \sqrt{\ln(k+1)} N_{k,d}^{-1/2}. \tag{9}$$

Next we show that the low-degree components $\Pi_{\leq k} f$ can be learned w.r.t. L_∞-error using polynomial samples because the L_∞-error of a low-degree polynomial is upper bounded by its L_2-error. Indeed, Fact 2 states that $\|h\|_\infty \leq \sqrt{N_{k,d}} \|h\|_2, \forall h \in \mathbb{Y}_{k,d}$. Then for any low-degree polynomial $g \in \mathbb{Y}_{\leq k,d}$,

$$\|g - \Pi_{\leq k} f\|_\infty = \|\Pi_{\leq k}(g-f)\|_\infty \leq \textstyle\sum_{l=0}^k \|\Pi_l(g-f)\|_\infty \leq \sum_{l=0}^k \|\Pi_l(g-f)\|_2 N_{l,d}^{1/2}. \tag{10}$$

When $g \in \mathbb{Y}_{\leq k,d}$, we have $\|g - \Pi_{\leq k} f\|_2^2 = \|\Pi_{\leq k}(g-f)\|_2^2 = \sum_{l=0}^k \|\Pi_l(g-f)\|_2^2$. Continuing Eq. (10) by applying Cauchy-Schwarz, we get

$$\|g - \Pi_{\leq k} f\|_\infty \leq N_{k,d}^{1/2} \sqrt{k+1} \|\Pi_{\leq k}(g-f)\|_2 = N_{k,d}^{1/2} \sqrt{k+1} \|g - \Pi_{\leq k} f\|_2. \tag{11}$$

Now we can choose an threshold $k \geq 0$ to balance the two terms in Eq. (9) and Eq. (11). For any desired error level $\epsilon > 0$, we can choose an k such that $\sqrt{\ln(k+1)} N_{k,d}^{-1/2} = \Theta(\epsilon/2)$ and get

$$\|g - f\|_\infty \leq \|g - \Pi_{\leq k} f\|_\infty + \|f - \Pi_{\leq k} f\|_\infty \lesssim \mathrm{poly}(1/\epsilon) \|g - \Pi_{\leq k} f\|_2 + \epsilon/2. \tag{12}$$

Finally, for any truncation threshold $k > 0$, $\Pi_{\leq k} f$ is a low-degree polynomial and belongs to the family $\mathcal{F}_k$ defined in Eq. (5). Therefore classic statistical learning theory implies that empirical risk minimization outputs a function g with $\|g - \Pi_{\leq k} f\|_2 \leq \mathrm{poly}(\epsilon)$ using only $\mathrm{poly}(1/\epsilon)$ samples (Lemma 9), which completes the proof. ■

4.3. Proof of Lemma 5

To prove Lemma 5, we first characterize an isotropic Gaussian random field in the spherical harmonics expansion.

Let $f : \mathbb{S}^{d-1} \to \mathbb{R}$ be a function drawn from an isotropic Gaussian random field with covariance $\kappa : [0,1] \to \mathbb{R}$, and $\{Y_{k,j}\}_{k \geq 0, 1 \leq j \leq N_{k,d}}$ a set of orthonormal spherical harmonics basis. We will show that the projection of f to the degree-k spherical harmonics space is isotropic. In other words, $\{\langle f, Y_{k,j} \rangle\}_{1 \leq j \leq N_{k,d}}$ are i.i.d. random variables.

Indeed, by Lang and Schwab (2015, Theorem 5.5), f admits the following spherical harmonics decomposition

$$f(\cdot) \overset{d}{=} \textstyle\sum_{k \geq 0} \left(\hat{\kappa}_k^{1/2} N_{k,d}^{-1/4} \sum_{j=1}^{N_{k,d}} a_{k,j} Y_{k,j}(\cdot) \right), \tag{13}$$

where $a_{k,j}$ are i.i.d. unit Gaussian random variables. Hence, to prove Lemma 5 we only need to examine the property of a random function drawn from the spherical harmonics space $\mathbb{Y}_{k,d}$, which is a $N_{k,d}$-dimensional Hilbert space.

The following lemma shows that a random spherical harmonics is not spiky because its L_∞/L_2 ratio is upperbounded by $O(d\sqrt{\ln k})$ with high probability, whereas the worst case ratio is $\sqrt{N_{k,d}} = \Omega(\min\{d^{k/2}, k^{d/2}\})$.

Lemma 6 *For any fixed $k \geq 0$, let $\{Y_{k,j}\}_{j=1}^{N_{k,d}}$ be any set of orthonormal basis for degree-k spherical harmonics $\mathbb{Y}_{k,d}$. Let $g = \sum_{j=1}^{N_{k,d}} a_j Y_{k,j}$ be a random spherical harmonics where $a_j \sim \mathcal{N}(0,1)$ are independent unit Gaussian random variables. For any $\delta > 0$ we have, with probability at least $1-\delta$,*

$$\|g\|_\infty \leq 5\sqrt{\ln(3/\delta) + 2d^2\ln(k+1)}\|g\|_2. \tag{14}$$

Lemma 6 is a high-dimensional version of Burq and Lebeau (2014, Theorem 2). The proof of Burq and Lebeau (2014) relies on the Sobolev embedding theorem, which treats the dimension d as a constant. In contrast, we compute the exact dependency on the dimension d by instantiating the Riesz representation theorem on the space of spherical harmonics and then applying a uniform convergence argument.

In the following, we present a proof sketch of Lemma 6. The full proof is deferred to Appendix D.

Proof Sketch of Lemma 6 To prove the L_∞/L_2 norm ratio of g, we first invoke Lemma 22 which states that

$$\forall x \in \mathbb{S}^{d-1}, \quad g(x) = \sqrt{N_{k,d}}\left\langle g, \bar{P}_{k,d}(\langle x, \cdot\rangle)\right\rangle. \tag{15}$$

Since $\bar{P}_{k,d}(\langle x,\cdot\rangle) \in \mathbb{Y}_{k,d}$, Lemma 22 is an instantiation of the Riesz representation theorem on the space $\mathbb{Y}_{k,d}$. The Riesz representation theorem states that for a Hilbert space, every continuous linear functional (in this case, the evaluation functional $\mathrm{ev}_x : g \to g(x)$) can be represented by the inner product with an element in the space (in this case, $\sqrt{N_{k,d}}\bar{P}_{k,d}(\langle x,\cdot\rangle)$).

For any fixed $x \in \mathbb{S}^{d-1}$, because $g = \Pi_k f$ is a Gaussian vector in the $N_{k,d}$-dimensional space $\mathbb{Y}_{k,d}$ and $\bar{P}_{k,d}(\langle x,\cdot\rangle) \in \mathbb{Y}_{k,d}$ is a fixed vector, the function value $g(x)$ has a Gaussian distribution. Formally speaking, we can write $\bar{P}_{k,d}(\langle x,\cdot\rangle) = \sum_{j=1}^{N_{k,d}} u_{k,j}Y_{k,j}(\cdot)$ for some fixed parameters $u_{k,j}$. Let $\boldsymbol{a}_k = [a_{k,j}]_{1\leq j\leq N_{k,d}}$ and $\boldsymbol{u}_k = [u_{k,j}]_{1\leq j\leq N_{k,d}}$, then we get

$$g(x) = \sqrt{N_{k,d}}\left\langle g, \bar{P}_{k,d}(\langle x,\cdot\rangle)\right\rangle = \sqrt{N_{k,d}}\left\langle \boldsymbol{a}_k, \boldsymbol{u}_k\right\rangle \sim \sqrt{N_{k,d}}\mathcal{N}(0, \|\boldsymbol{u}_k\|_2^2). \tag{16}$$

Since $\|\boldsymbol{u}_k\|_2 = \|\bar{P}_{k,d}(\langle x,\cdot\rangle)\|_2 = 1$ and $\|g\|_2 = \|\boldsymbol{a}_k\|_2 \approx \sqrt{N_{k,d}}$, by concentration inequality of Gaussian vectors (Lemma 27) we get for any fixed $x \in \mathbb{S}^{d-1}$, with high probability $|g(x)| \lesssim \|\boldsymbol{a}_k\|_2 = \|g\|_2$. Finally, we can use a covering number argument to prove a uniform convergence of all $x \in \mathbb{S}^{d-1}$. Hence, we prove that with high probability, $\forall x \in \mathbb{S}^{d-1}$, $|g(x)| \leq \widetilde{O}(d\sqrt{\ln k})\|g\|_2$, which implies Eq. (14). ■

With Lemma 6, we can now prove Lemma 5.

Proof of Lemma 5 Recall that Lang and Schwab (2015, Theorem 5.5) gives the following spherical harmonics decomposition

$$f(x) \overset{d}{=} \textstyle\sum_{k\geq 0} \left(\hat{\kappa}_k^{1/2} N_{k,d}^{-1/4} \sum_{j=1}^{N_{k,d}} a_{k,j} Y_{k,j}(x)\right) \tag{17}$$

where $a_{k,j} \sim \mathcal{N}(0,1)$ are independent Gaussian random variables. By Lemma 6, for any fixed $k \geq 0$, with probability at least $1-\delta/(2(k+1)^2)$ we have

$$\begin{aligned} \|\Pi_k f\|_\infty &\leq 5\sqrt{\ln(6(k+1)^2/\delta) + 2d^2 \ln(k+1)}\|\Pi_k f\|_2 &(18)\\ &\leq 5\sqrt{2\ln(6/\delta) + 2(d^2+1)\ln(k+1)}\|\Pi_k f\|_2. &(19)\end{aligned}$$

By union bound over k, with probability at least $1-\delta$ we get

$$\forall k \geq 0, \quad \|\Pi_k f\|_\infty \leq 5\sqrt{2\ln(6/\delta) + 2(d^2+1)\ln(k+1)}\|\Pi_k f\|_2, \tag{20}$$

which proves Eq. (6).

Now we prove the second part of lemma. Since $\{Y_{k,j}\}_{j=1}^{N_{k,d}}$ forms an orthonormal basis of $\mathbb{Y}_{k,d}$, we get

$$\|\Pi_k f\|_2^2 = \hat{\kappa}_k N_{k,d}^{-1/2} \textstyle\sum_{j=1}^{N_{k,d}} a_{k,j}^2 \leq c^2 N_{k,d}^{-1-\alpha} \sum_{j=1}^{N_{k,d}} a_{k,j}^2. \tag{21}$$

For any fixed $k \geq 0$, since $a_{k,j}$ are i.i.d. unit Gaussian random variables, by the concentration of the norm of Gaussian vectors (Laurent and Massart, 2000, Lemma 1), we have

$$\forall t > 0, \quad \Pr\left(\textstyle\sum_{j=1}^{N_{k,d}} a_{k,j}^2 \geq N_{k,d} + 2\sqrt{N_{k,d}}\sqrt{t} + 2t\right) \leq \exp(-t). \tag{22}$$

Take $t = \ln(2(k+1)^2/\delta)$. Note that $N_{k,d} \geq k \geq \ln(k+1)$. As a result, for all $k \geq 0$ we get

$$N_{k,d} + 2\sqrt{N_{k,d}}\sqrt{t} + 2t \leq 9N_{k,d}\ln(2/\delta). \tag{23}$$

Consequently,

$$\Pr\left(\textstyle\sum_{j=1}^{N_{k,d}} a_{k,j}^2 \geq 9N_{k,d}\ln(2/\delta)\right) \leq (k+1)^{-2}\delta/2. \tag{24}$$

Combining with Eq. (21) and union bound over k, with probability at least $1-\delta$ we get

$$\forall k \geq 0, \quad \|\Pi_k f\|_2 \leq c N_{k,d}^{-1/2-\alpha/2} \left(\textstyle\sum_{j=1}^{N_{k,d}} a_{k,j}^2\right)^{1/2} \leq 3c\sqrt{\ln(2/\delta)} N_{k,d}^{-\alpha/2}, \tag{25}$$

which proves Eq. (7). ■

5. Lower Bounds

In this section, we present two lower bounds to motivate our Condition 2. Both lower bounds hold for any algorithm that can *adaptively* choose its data point x_i and observes a noisy signal $f(x_i) + \mathcal{N}(0,1)$, where f denotes the ground-truth function. Our lower bounds may be of independent interest.

Lower bounds for functions with decay rate $N_{k,d}^{-1/2}$. The following lemma proves that, in the worst case, L_∞-recovery is hard even when the function's spherical harmonics decomposition decays at a rate of $N_{k,d}^{-1/2}$.

Lemma 7 *For a fixed integer $k \geq 4$ and $\beta_k \in (0,1)$, define $\mathcal{F}_k = \{\beta_k P_{k,d}(\langle \cdot, u\rangle) : u \in \mathbb{S}^{d-1}\}$ be the hypothesis class. For any fixed algorithm, let $E_{f,n}$ be the probability that the algorithm outputs $\hat{f}$ such that $\|\hat{f} - f\|_\infty \leq \beta_k/4$ using n samples when the ground-truth function is f. Then if $n < N_{k,d}\beta_k^{-2}$, $\min_{f\in\mathcal{F}_k} E_{f,n} \leq 1/2$.*

Since $\|P_{k,d}(\langle \cdot, u\rangle)\|_2 = N_{k,d}^{-1/2}$, the function class $\mathcal{F}_k$ (when $\beta_k = 1$) is a subset of functions that satisfies Condition 1 with $\alpha = 1$. Therefore, no algorithm can achieve polynomial sample complexity for L_∞-recovery with only the smoothness condition (Condition 1).

Lemma 7 is proved by showing that no algorithm can distinguish all the functions $f \in \mathcal{F}_k$ using $o(N_{k,d})$ samples because the average signal-to-noise ratio of any data point is roughly $N_{k,d}^{-1/2}$. Hence, the worst-case sample complexity is at least $\Omega(N_{k,d})$. The proof is deferred to Appendix A.5.

Lower bounds for two-layer ReLU neural networks. We first formally define the class of two-layer neural networks used in this paper. Let NN-ReLU(L_p) be the family of two layer neural networks (NNs) with L_p-norm bounds. Formally speaking,

$$\text{NN-ReLU}(L_p) = \{g(x) \triangleq \mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[\sigma(x^\top \xi)c(\xi)] : \|c\|_p \leq 1\}, \tag{26}$$

where σ is the ReLU activation and $c : \mathbb{S}^{d-1} \to \mathbb{R}$ is the weight of the NN. Classical finite width neural networks belong to NN-ReLU(L_1) because their weights c can be represented by the mixtures of Dirac measures.

The following theorem shows that learning two-layer neural networks with ReLU activation is statistically hard even when the NN has a constant norm. The lower bound holds for NN-ReLU(L_2), which is a subset of NN-ReLU(L_1).

Theorem 8 *Given the hypothesis class NN-ReLU(L_2). If an algorithm, when running on every possible instance $f \in$ NN-ReLU(L_2), takes in n data points uniformly sampled from the sphere $\mathbb{S}^{d-1}$ and outputs a function g such that $\|f - g\|_\infty \leq \epsilon$ with probability at least $1/2$, then $n \geq \Omega\left(\left(0.002\epsilon^{-1}d^{-7/4}\right)^{d/2}\right)$. As a corollary, the minimax sample complexity of learning NN-ReLU(L_2) with L_∞-error $\epsilon = O(d^{-7/4})$ requires at least 2^d samples.*

Theorem 8 does not contradict with existing results on the recovery of two-layer neural networks (Zhong et al., 2017; Zhou et al., 2021) because they focus on the finite-width case while our lower bound holds for infinite-width neural networks. Compared with the lower bound in Dong et al. (2021), Theorem 8 does not rely on the bias term in the ReLU activation to kill the signal. Instead, we invoke the Funk-Hecke formula (Theorem 1) to show that two-layer ReLU NNs can represent spiky functions with constant norm.

Theorem 8 is proved by showing that $\mathcal{F}_k$ defined in Lemma 7 is a subset of NN-ReLU(L_2) if we take $\beta_k \approx k^{-2}$. The proof is deferred to Appendix A.6.

6. Conclusion

In this paper, we make some initial steps toward L_∞-recovery for nonlinear models by proving a polynomial sample complexity bound for random function drawn from Gaussian random fields. We also prove a $\exp(d)$ sample complexity lower bound for recovering the worst-case infinite-width two-layer neural nets with unbiased ReLU activation, which may be of independent interest.

For future works, we raise the following open questions:

1. To instantiate Condition 2, this paper focuses on functions f drawn from Gaussian random fields because they have *independent* components in the spherical harmonics space. However, Condition 2 also holds when f has correlated components. For example, when $\Pi_k f = \sum_{j=1}^{N_{k,d}} a_{k,j} Y_{k,j}$ where $[a_{k,j}]_{j=1}^{N_{k,d}}$ lies on the $(N_{k,d}-1)$-dimensional sphere. Is it possible to prove Condition 2 for functions drawn from other distribution?

2. A two-layer single-neuron neural nets with exponential activation, i.e., functions of the form $g(\cdot) = \exp(\langle \cdot, u \rangle)$ for some $u \in \mathbb{S}^{d-1}$, does not satisfy Condition 2. In fact, $\Pi_k g$ is the most spiky function in $\mathbb{Y}_{k,d}$ because $\Pi_k g = \lambda_k \bar{P}_{k,d}(\langle \cdot, u \rangle)$. Can we find a natural (random) subset of two-layer neural networks that satisfy Condition 2?

Acknowledgment

The authors would like to thank Ruixiang Zhang, Yakun Xi, Yuhao Zhou, Jason D. Lee for helpful discussions. The authors would also like to thank anonymous reviewers for the references to additional related works. The authors would like to thank the support of NSF CIF 2212263.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

Kendall Atkinson and Weimin Han. *Spherical harmonics and approximations on the unit sphere: an introduction*, volume 2044. Springer Science & Business Media, 2012.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory*, pages 195–268. PMLR, 2019.

Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Karine Bertin. Asymptotically exact minimax estimation in sup-norm for anisotropic hölder classes. *Bernoulli*, 10(5):873–888, 2004a.

Karine Bertin. Minimax exact constant in sup-norm for nonparametric regression with random design. *Journal of statistical planning and inference*, 123(2):225–242, 2004b.

Jean Bourgain and Joram Lindenstrauss. Projection bodies. In *Geometric Aspects of Functional Analysis*, pages 250–270. Springer, 1988.

Nicolas Burq and Gilles Lebeau. Probabilistic sobolev embeddings, applications to eigenfunctions estimates. *Geometric and spectral analysis*, 630:307–318, 2014.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

Kefan Dong, Jiaqi Yang, and Tengyu Ma. Provable model-based nonlinear bandit and reinforcement learning: Shelve optimism, embrace virtual curvature. *Advances in Neural Information Processing Systems*, 34, 2021.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, pages 1675–1685, November 2018.

Adrian Ebert and Friedrich Pillichshammer. Tractability of approximation in the weighted korobov space in the worst-case setting—a complete picture. *Journal of Complexity*, 67: 101571, 2021.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

Haoyu Fu, Yuejie Chi, and Yingbin Liang. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE transactions on signal processing*, 68:3225–3235, 2020.

F Golubev, O Lepski, and B Levit. On adaptive estimation using the sup-norm losses. 2000.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Steffen Grünewälder, Jean-Yves Audibert, Manfred Opper, and John Shawe-Taylor. Regret bounds for gaussian process bandit problems. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 273–280. JMLR Workshop and Conference Proceedings, 2010.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. 2008.

Baihe Huang, Kaixuan Huang, Sham Kakade, Jason D Lee, Qi Lei, Runzhe Wang, and Jiaqi Yang. Going beyond linear rl: Sample efficient neural function approximation. *Advances in Neural Information Processing Systems*, 34:8968–8983, 2021a.

Baihe Huang, Kaixuan Huang, Sham Kakade, Jason D Lee, Qi Lei, Runzhe Wang, and Jiaqi Yang. Optimal gradient-based algorithms for non-concave bandit optimization. *Advances in Neural Information Processing Systems*, 34:29101–29115, 2021b.

IA Ibragimov and RZ Khas' minskii. Asymptotic bounds on the quality of the nonparametric regression estimation in. *Journal of Soviet Mathematics*, 24:540–550, 1984.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Alexander P Korostelev. An asymptotically minimax regression estimator in the uniform norm up to exact constant. *Theory of Probability & Its Applications*, 38(4):737–743, 1994.

Andreas Krause and Cheng Ong. Contextual gaussian process bandit optimization. *Advances in neural information processing systems*, 24, 2011.

David Krieg. Uniform recovery of high-dimensional cr-functions. *Journal of Complexity*, 50:116–126, 2019.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34:26548–26560, 2021.

Frances Y Kuo, Grzegorz W Wasilkowski, and Henryk Woźniakowski. Multivariate l_∞ approximation in the worst case setting over reproducing kernel hilbert spaces. *Journal of approximation theory*, 152(2):135–160, 2008.

Frances Y Kuo, Grzegorz W Wasilkowski, and Henryk Woźniakowski. On the power of standard information for multivariate approximation in the worst case setting. *Journal of Approximation Theory*, 158(1):97–125, 2009.

Annika Lang and Christoph Schwab. Isotropic gaussian random fields on the sphere: regularity, fast simulation and stochastic partial differential equations. *The Annals of Applied Probability*, 25(6):3047–3094, 2015.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Uniform error bounds for gaussian process regression with application to safe control. *Advances in Neural Information Processing Systems*, 32, 2019.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Uniform error and posterior variance bounds for gaussian process regression with application to safe control. *arXiv preprint arXiv:2101.05328*, 2021.

Gene Li, Pritish Kamath, Dylan J. Foster, and Nathan Srebro. Eluder dimension and generalized rank, 2021.

Jihao Long and Jiequn Han. Reinforcement learning with function approximation: From linear to nonlinear. *arXiv preprint arXiv:2302.09703*, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks (published at ICLR 2018). *arXiv*, 2017.

Hrushikesh N Mhaskar. Function approximation with zonal function networks with activation functions analogous to the rectified linear unit functions. *Journal of Complexity*, 51: 1–19, 2019.

Hrushikesh Narhar Mhaskar. Weighted quadrature formulas and approximation by zonal function networks on the sphere. *Journal of Complexity*, 22(3):348–370, 2006.

Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

Ha Quang Minh, Partha Niyogi, and Yuan Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pages 154–168. Springer, 2006.

M Nyssbaum. Nonparametric estimation of a regression function that is smooth in a domain in rˆk. *Theory of Probability & Its Applications*, 31(1):108–115, 1987.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. In *Conference on Learning Theory*, pages 1723–1742. PMLR, 2017.

Rolf Schneider. Zu einem problem von shephard über die projektionen konvexer körper. *Mathematische Zeitschrift*, 101(1):71–82, 1967.

Matthias Seeger. Gaussian processes for machine learning. *International journal of neural systems*, 14(02):69–106, 2004.

Shubhanshu Shekhar and Tara Javidi. Gaussian process bandits with adaptive discretization. 2018.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, number CONF, pages 1015–1022. Omnipress, 2010.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Alexandre B Tsybakov. Pointwise and sup-norm sharp adaptive estimation of functions on the sobolev classes. *The Annals of Statistics*, 26(6):2420–2469, 1998.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021a.

Sattar Vakili, Henry Moss, Artem Artemev, Vincent Dutordoir, and Victor Picheny. Scalable thompson sampling using sparse gaussian process models. *Advances in neural information processing systems*, 34:5631–5643, 2021b.

Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 654–663, 2013.

Jan Vybíral. Weak and quasi-polynomial tractability of approximation of infinitely differentiable functions. *Journal of Complexity*, 30(2):48–55, 2014.

Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

Chuwei Wang, Shanda Li, Di He, and Liwei Wang. Is l2 physics-informed loss always suitable for training physics-informed neural network? *arXiv preprint arXiv:2206.02016*, 2022.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.

Ziyu Wang and Nando de Freitas. Theoretical analysis of bayesian optimisation with unknown gaussian process hyper-parameters. *arXiv preprint arXiv:1406.7758*, 2014.

Colin Wei and Tengyu Ma. Improved sample complexities for deep networks and robust classification via an all-layer margin. *arXiv preprint arXiv:1910.04284*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *NIN*, 8:35–67, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021a.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling. In *International Conference on Learning Representation (ICLR)*, 2021b.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics*, pages 1524–1534. PMLR, 2019.

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International conference on machine learning*, pages 4140–4149. PMLR, 2017.

Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In *Conference on Learning Theory*, pages 4577–4632. PMLR, 2021.

List of Appendices

Appendix A. Missing Proofs

A.1. Proof of Theorem 3

In the following, we first state two lemmas that are critical to the proof of Theorem 3.

The next lemma proves that the empirical risk minimization step used in Alg. 1 outputs a function with small L_2 loss, whose proof is deferred to Appendix A.2.

Lemma 9 *Suppose the function $f : \mathbb{S}^{d-1} \to \mathbb{R}$ satisfies Condition 1 for some fixed $\alpha \in (0,1], c_1, c_2 > 0$. For any $\epsilon > 0$, let $k = \inf_{l \geq 0}\{2c_1c_2(l+1)^{3/2}(N_{l+1,d})^{-\alpha/2} \leq \epsilon/2\}$.*

Let $\mathcal{F}_k \leftarrow \{g \in \mathbb{Y}_{\leq k,d} : \|\Pi_l g\|_2 \leq c_1, \forall l \in [0,k]\}$ be the function class defined in Alg. 1. For a given dataset $\{(x_i,y_i)\}_{i=1}^n$, let $\hat{\mathcal{L}}(h) \triangleq \frac{1}{n}\sum_{i=1}^n (h(x_i)-y_i)^2$ be the empirical L_2 loss, and $g = \operatorname{argmin}_{h \in \mathcal{F}_k} \hat{\mathcal{L}}(h)$.

For any $\delta > 0, \epsilon_1 > 0$, when $d \geq \max\{2e, 4/\alpha\}$ and the number of samples $n \geq \Omega(\mathrm{poly}(c_1c_2, 1/\epsilon)^{1/\alpha}\mathrm{poly}(1/\epsilon_1, \ln(1/\delta)))$, with probability at least $1-\delta$,

$$\|\Pi_{\leq k}(f-g)\|_2 = \|\Pi_{\leq k} f - g\|_2 \leq \epsilon_1. \tag{27}$$

The following lemma proves that with Conditions 1 and 2, $\|f - \Pi_{\leq k} g\|_\infty$ can be upper bounded by $\|\Pi_{\leq k}(f-g)\|_2$ for properly chosen k.

Lemma 10 *Suppose the function $f : \mathbb{S}^{d-1} \to \mathbb{R}$ satisfies Conditions 1 and 2 for some fixed $\alpha \in (0,1], c_1, c_2 > 0$, and $d \geq 10\alpha^{-1}2^{5/\alpha} + 2$. For any $\epsilon > 0$, define $k = \inf_{l\geq 0}\{2c_1c_2(l+1)^{3/2}(N_{l+1,d})^{-\alpha/2} \leq \epsilon/2\}$. Then for any function $g : \mathbb{S}^{d-1} \to \mathbb{R}$ with $\|\Pi_{\leq k}(f-g)\|_2 \leq \frac{1}{4}\epsilon^{3/\alpha+1}(4c_1c_2)^{-3/\alpha}d^{-4/\alpha}$, we have $\|f - \Pi_{\leq k} g\|_\infty \leq \epsilon$.*

Proof of Lemma 10 is deferred to Appendix A.4.

Now we are ready to prove Theorem 3.

Proof of Theorem 3 Let $\epsilon_1 = \frac{1}{4}\epsilon^{3/\alpha+1}(4c_1c_2)^{-3/\alpha}d^{-4/\alpha}$. We prove Theorem 3 in the following two steps.

Step 1: upper bound the population L_2 loss. In this step, we use classic statistical learning tools to show that the ERM step (i.e., $g = \operatorname{argmin}_{h\in\mathcal{F}_k}\sum_{i=1}^n (h(x_i)-y_i)^2$) returns a function g with small L_2 loss. In particular, by Lemma 9 we get $\|\Pi_{\leq k}(f-g)\|_2 \leq \epsilon_1$.

Step 2: upper bound the L_∞-error via truncation. In this step we show that with Conditions 1 and 2 on the ground-truth function, any function g with a small L_2-error will also have a small L_∞-error when projected to the low-degree spherical harmonics space. Formally speaking, invoking Lemma 10 we get $\|\Pi_{\leq k}(f-g)\|_2 \leq \epsilon_1 \implies \|f - \Pi_k g\|_\infty \leq \epsilon$.

Finally, since $g \in \mathcal{F}_k \subset \mathbb{Y}_{\leq k,d}$, we get $g = \Pi_k g$. Hence, combining these two steps we prove the desired result. ■

A.2. Proof of Lemma 9

In the following we prove Lemma 9.

Proof of Lemma 9 We prove Lemma 9 in two steps.

Step 1: expressivity. In this step, we prove that $\Pi_{\leq k} f \in \mathcal{F}_k$. Indeed, by Condition 1 we get

$$\|\Pi_k f\|_2 \leq c_1 N_{k,d}^{-\alpha/2} \leq c_1, \tag{28}$$

meaning that $\Pi_{\leq k} f \in \mathcal{F}_k$.

Consequently, using the definition $g = \operatorname{argmin}_{h \in \mathcal{F}_k} \hat{\mathcal{L}}(g)$ we have $\hat{\mathcal{L}}(g) \leq \hat{\mathcal{L}}(\Pi_{\leq k} f)$.

Step 2: uniform convergence. In this step, we prove that using

$$n = \Omega(\text{poly}(c_1, N_{k,d}, \ln(1/\delta), 1/\epsilon_1))$$

samples, Alg. 1 outputs a function $g \in \mathcal{F}$ such that

$$\|g - \Pi_{\leq k} f\|_2 \leq \epsilon_1. \tag{29}$$

To this end, by the uniform convergence of $\mathcal{F}_k$ (Lemma 11), when

$$n = \Omega(\text{poly}(c_1, N_{k,d}, \ln(1/\delta), 1/\epsilon_1)),$$

with probability at least $1 - \delta$,

$$\|g - f\|_2^2 \leq \hat{\mathcal{L}}(g) + \epsilon_1^2/2 \leq \hat{\mathcal{L}}(\Pi_{\leq k} f) + \epsilon_1^2/2 \leq \|\Pi_{\leq k} f - f\|_2^2 + \epsilon_1^2. \tag{30}$$

Since $\mathcal{F}_k \subseteq \mathbb{Y}_{\leq k,d}$, by the Parseval's identity we get

$$\begin{aligned}\forall h \in \mathcal{F}_k, \quad \|h - f\|_2^2 &= \|\Pi_{\leq k}(h - f)\|_2^2 + \|\Pi_{>k}(h - f)\|_2^2 && (31)\\ &= \|\Pi_{\leq k}(h - f)\|_2^2 + \|\Pi_{>k} f\|_2^2 = \|h - \Pi_{\leq k} f\|_2^2 + \|f - \Pi_{\leq k} f\|_2^2. && (32)\end{aligned}$$

Note that $g \in \mathcal{F}_k$ and $\Pi_{\leq k} f \in \mathcal{F}_k$. Combining with Eq. (30) we get

$$\|g - \Pi_{\leq k} f\|_2^2 \leq \epsilon_1^2. \tag{33}$$

Finally, by the choice of k and Proposition 29, $N_{k,d} = \text{poly}(c_1 c_2, 1/\epsilon)^{1/\alpha}$, which means that

$$n = \Omega(\text{poly}(c_1, N_{k,d}, \ln(1/\delta), 1/\epsilon_1)) = \Omega(\text{poly}(c_1 c_2, 1/\epsilon)^{1/\alpha} \text{poly}(\ln(1/\delta), 1/\epsilon_1)). \tag{34}$$

■

The following lemma proves uniform convergence results for the function class $\mathcal{F}_k$.

Lemma 11 *In the setting of Lemma 9, for any $\delta > 0, \epsilon_1 > 0$ and $n \geq \Omega(\text{poly}(c_1, N_{k,d}, \ln(1/\delta), 1/\epsilon_1))$, with probability at least $1 - \delta$ we have*

$$\sup_{g \in \mathcal{F}_k} \left| \|g - f\|_2^2 - \hat{\mathcal{L}}(g) \right| \leq \epsilon_1. \tag{35}$$

Proof We prove this lemma using the Rademacher complexity of kernel methods (Bartlett and Mendelson, 2002). First we upper bound the Rademacher complexity of $\mathcal{F}_k$. Let $x_1, \cdots, x_n$ be a set of data points and $\hat{R}_n(\mathcal{F}_k)$ the empirical Rademacher complexity of $\mathcal{F}_k$, defined by

$$\hat{R}_n(\mathcal{F}_k) = \frac{1}{n}\mathbb{E}_{\sigma_1,\cdots,\sigma_n\sim\{-1,1\}^n}\left[\sup_{g\in\mathcal{F}_k}\left|\sum_{i=1}^n \sigma_i g(x_i)\right|\right]. \tag{36}$$

Recall that $\{Y_{k,j}\}_{j=1}^{N_{k,d}}$ is an orthonormal basis of $\mathbb{Y}_{k,d}$, and any function $g \in \mathcal{F}_k$ can be written as $g(x) = \sum_{l=0}^{k}\sum_{j=1}^{N_{l,d}} a_{l,j}Y_{l,j}(x)$ where $\sum_{j=1}^{N_{k,d}} a_{l,j}^2 \le c_1^2, \forall l \in [0,k]$. Hence, after defining $\phi_k(x) \triangleq [Y_{l,j}(x)]_{l\in[0,k],j\in[N_{l,d}]}$ as the feature vector, and $\boldsymbol{a} \triangleq [a_{l,j}]_{l\in[0,k],j\in[N_{l,d}]}$, we have $g(x) = \langle \phi_k(x), \boldsymbol{a}\rangle$ and $\|\boldsymbol{a}\|_2 \le \sqrt{k+1}c_1$.

Let $k(x,x') = \langle \phi_k(x), \phi_k(x')\rangle$ be the kernel function. Then by the fact that $\sum_{j=1}^{N_{l,d}} Y_{l,j}(x)^2 = N_{l,d}, \forall l \ge 0$ (Atkinson and Han, 2012, Theorem 2.9), we have

$$k(x,x) = \sum_{l=0}^{k}\sum_{j=1}^{N_{l,d}} Y_{l,j}(x)^2 = \sum_{l=0}^{k} N_{k,d} \le (k+1)N_{k,d}. \tag{37}$$

By Bartlett and Mendelson (2002, Lemma 22) we get $\hat{R}_n(\mathcal{F}_k) \le \frac{2(k+1)\sqrt{N_{k,d}}}{\sqrt{n}}$.

Since for any x, we get $g(x) \le \|\phi_k(x)\|_2\|a\|_2 = c_1(k+1)\sqrt{N_{k,d}}$, the L_2 loss is $(2c_1(k+1)\sqrt{N_{k,d}})$-Lipschitz. As a result, Kakade et al. (2008, Theorem 3) implies that with probability at least $1-\delta$, $\forall g \in \mathcal{F}_k$

$$|\|g - f\|_2^2 - \hat{\mathcal{L}}(g)| = |\mathbb{E}[\hat{\mathcal{L}}(g)] - \hat{\mathcal{L}}(g)| \lesssim \frac{c_1(k+1)^2 N_{k,d}}{\sqrt{n}} + c_1(k+1)^2 N_{k,d}\sqrt{\frac{\ln(1/\delta)}{n}}. \tag{38}$$

Note that $N_{k,d} \ge k$. As a result, when $n \ge \Omega(\text{poly}(c_1, N_{k,d}, \ln(1/\delta), 1/\epsilon_1))$, we get

$$\forall g \in \mathcal{F}_k, \quad |\|g - f\|_2^2 - \hat{\mathcal{L}}(g)| \le \epsilon_1. \tag{39}$$

which proves the desired result. ■

A.3. Proof of Lemma 12

In the following we present and prove Lemma 12, which is used to prove Lemma 10.

Lemma 12 *Suppose the function $f : \mathbb{S}^{d-1} \to \mathbb{R}$ satisfies Conditions 1 and 2 for some fixed $\alpha \in (0,1]$ and $c_1, c_2 > 0$. When $d \ge 10\alpha^{-1}2^{5/\alpha} + 2$, we have*

$$\|f - \Pi_{\le k-1}f\|_\infty \le 2c_1c_2k^{3/2}(N_{k,d})^{-\alpha/2}, \quad \forall k \ge 1. \tag{40}$$

Proof of Lemma 12 Let $c = c_1c_2$. By basic algebra we get

$$\|f - \Pi_{\le k-1}f\|_\infty = \left\|\sum_{l\ge0}\Pi_l f - \Pi_{\le k-1}f\right\|_\infty = \left\|\sum_{l\ge k}\Pi_l f\right\|_\infty \le \sum_{l\ge k} c\sqrt{\ln(l+1)}(N_{l,d})^{-\alpha/2}. \tag{41}$$

Therefore we only need to prove

$$\sum_{l\geq k} c\sqrt{\ln(l+1)}(N_{l,d})^{-\alpha/2} \leq 2ck^{3/2}(N_{k,d})^{-\alpha/2}. \tag{42}$$

Recall that $N_{l,d} = \frac{2l+d-2}{l+d-2}\frac{\Gamma(l+d-1)}{\Gamma(l+1)\Gamma(d-1)}$. It follows that

$$\sum_{l\geq k} c\sqrt{\ln(l+1)}(N_{l,d})^{-\alpha/2} \leq \sum_{l\geq k} c\sqrt{l}\left(\frac{\Gamma(l+d-1)}{\Gamma(l+1)\Gamma(d-1)}\right)^{-\alpha/2}. \tag{43}$$

Let $a_l \triangleq \left(\frac{\Gamma(l+d-1)}{\Gamma(l+1)\Gamma(d-1)}\right)^{-\alpha/2}\sqrt{l}$. We first prove that when $d \geq \frac{10}{\alpha}2^{5/\alpha}+2$, $\frac{a_{l+1}}{a_l} \leq \left(\frac{l}{l+1}\right)^2, \forall l \geq 1$. By basic algebra we get

$$\frac{a_{l+1}}{a_l} = \sqrt{\frac{l+1}{l}}\left(\frac{l+1}{l+d-1}\right)^{\alpha/2}. \tag{44}$$

Let $\kappa = 2^{5/\alpha+1}$. We first focus on the case when $l \geq \frac{d}{\kappa-1}$. Since $\alpha(d-2)/5 \geq \kappa$, we have

$$\left(\frac{l+1}{l+d-1}\right)^{\alpha/5} = \left(1-\frac{d-2}{l+d-1}\right)^{\alpha/5} \leq 1-\frac{\alpha(d-2)/5}{l+d-1} \leq 1-\frac{\kappa}{l+d-1}. \tag{45}$$

When $l \geq \frac{d}{\kappa-1}$ we have $\frac{\kappa}{l+d-1} \geq \frac{1}{l+1}$. As a result, $\left(\frac{l+1}{l+d-1}\right)^{\alpha/5} \leq 1-\frac{1}{l+1} = \frac{l}{l+1}$. Equivalently, we get

$$\sqrt{\frac{l+1}{l}}\left(\frac{l+1}{l+d-1}\right)^{\alpha/2} \leq \left(\frac{l}{l+1}\right)^2. \tag{46}$$

Now we focus on the case when $l < \frac{d}{\kappa-1}$. In this case we have

$$\left(\frac{l+1}{l+d-1}\right)^{\alpha/2} < \left(\frac{\frac{d}{\kappa-1}+1}{\frac{d}{\kappa-1}+d-1}\right)^{\alpha/2} \leq \left(\frac{\frac{d}{\kappa-1}+2}{\frac{d}{\kappa-1}+d}\right)^{\alpha/2}. \tag{47}$$

Since $\frac{d}{\kappa-1} \geq 2$, we have

$$\left(\frac{\frac{d}{\kappa-1}+2}{\frac{d}{\kappa-1}+d}\right)^{\alpha/2} \leq \left(\frac{2\frac{d}{\kappa-1}}{\frac{d}{\kappa-1}+d}\right)^{\alpha/2} = \left(\frac{2}{\kappa}\right)^{\alpha/2} \leq 2^{5/2} \leq \left(\frac{l}{l+1}\right)^{5/2}. \tag{48}$$

Consequently,

$$\left(\frac{l+1}{l+d-1}\right)^{\alpha/2}\left(\frac{l+1}{l}\right)^{1/2} \leq \left(\frac{l}{l+1}\right)^2. \tag{49}$$

Combining Eq. (46) and Eq. (49), in both cases we have

$$\sqrt{\frac{l+1}{l}}\left(\frac{l+1}{l+d-1}\right)^{\alpha/2} \leq \left(\frac{l}{l+1}\right)^2. \tag{50}$$

Now continue Eq. (43) we get,

$$\sum_{l\ge k} c\sqrt{l}\left(\frac{\Gamma(l+d-1)}{\Gamma(l+1)\Gamma(d-1)}\right)^{-\alpha/2} = ca_k\sum_{l\ge k}\frac{a_l}{a_k} = ca_k\sum_{l\ge k}\prod_{l'=k}^{l-1}\frac{a_{l'+1}}{a_{l'}} \tag{51}$$

$$\le ca_k\sum_{l\ge k}\frac{k^2}{l^2} \le cka_k = ck^{3/2}\left(\frac{\Gamma(k+d-1)}{\Gamma(k+1)\Gamma(d-1)}\right)^{-\alpha/2} \le ck^{3/2}2^{\alpha/2}(N_{k,d})^{-\alpha/2} \tag{52}$$

$$\le 2ck^{3/2}(N_{k,d})^{-\alpha/2}. \tag{53}$$

■

A.4. Proof of Lemma 10

In this section we prove Lemma 10.

Proof of Lemma 10 Let $c = c_1c_2$. Recall that $k = \inf_{l\ge 0}\{2c(l+1)^{3/2}(N_{l+1,d})^{-\alpha/2} \le \epsilon/2\}$. By Lemma 12 we get

$$\|f - \Pi_{\le k}f\|_\infty \le \epsilon/2. \tag{54}$$

Hence, we only need to prove $\|\Pi_{\le k}g - \Pi_{\le k}f\|_\infty \le \epsilon/2$ and the desired result follows directly from triangle inequality.

Since $\Pi_{\le k}(g-f)$ has degree at most k, applying Fact 2 we get

$$\|\Pi_{\le k}g - \Pi_{\le k}f\|_\infty \le \sum_{l=0}^{k}\|\Pi_l(g-f)\|_\infty \le \sqrt{N_{k,d}}\sum_{l=0}^{k}\|\Pi_l(g-f)\|_2. \tag{55}$$

By Cauchy-Schwarz and Parseval's theorem we have

$$\sum_{l=0}^{k}\|\Pi_l(g-f)\|_2 \le \left((k+1)\sum_{l=0}^{k}\|\Pi_l(g-f)\|_2^2\right)^{1/2} \le \sqrt{k+1}\|\Pi_{\le k}(g-f)\|_2. \tag{56}$$

As a result,

$$\|\Pi_{\le k}g - \Pi_{\le k}f\|_\infty \le \sqrt{k+1}\sqrt{N_{k,d}}\|\Pi_{\le k}(g-f)\|_2. \tag{57}$$

In the following, we show that

$$\sqrt{k+1}\sqrt{N_{k,d}} \le 2(4c/\epsilon)^{3/\alpha}d^{4/\alpha}. \tag{58}$$

By the definition of k we have $2ck^{3/2}(N_{k,d})^{-\alpha/2} > \epsilon/2$. Hence,

$$\sqrt{N_{k,d}} \le \left(\frac{4c}{\epsilon}k^{3/2}\right)^{1/\alpha}. \tag{59}$$

To upper bound k, note that $N_{k,d} \ge (k/d)^{d-2}$. Therefore,

$$\epsilon < 4ck^{3/2}(N_{k,d})^{-\alpha/2} \le 4ck^{3/2}\left(\frac{d}{k}\right)^{-(d-2)\alpha/2}. \tag{60}$$

Solving for k we get $k \le (4c/\epsilon)^{\frac{2}{d\alpha-5}} d^{\frac{d\alpha-2}{d\alpha-5}}$. Combining with Eq. (59) and using the assumption that $d \ge \frac{10}{\alpha} 2^{5/\alpha} + 2$, we get

$$\sqrt{k+1}\sqrt{N_{k,d}} \le 2\,(4c/\epsilon)^{\frac{1}{\alpha}}\, k^{\frac{3}{2\alpha}+\frac{1}{2}} \tag{61}$$

$$\le 2\,(4c/\epsilon)^{\frac{1}{\alpha}}\,(4c/\epsilon)^{\frac{2}{d\alpha-5}\left(\frac{3}{2\alpha}+\frac{1}{2}\right)} d^{\frac{d\alpha-2}{d\alpha-5}\left(\frac{3}{2\alpha}+\frac{1}{2}\right)} \le 2(4c/\epsilon)^{3/\alpha} d^{4/\alpha}. \tag{62}$$

Finally, combining Eq. (62), Eq. (57) and the assumption $\|\Pi_{\le k}(g-f)\|_2 \le \frac{1}{4}\epsilon(4c/\epsilon)^{-3/\alpha} d^{-4/\alpha}$ we get

$$\|\Pi_{\le k} g - \Pi_{\le k} f\|_\infty \le \epsilon/2. \tag{63}$$

By triangle inequality and Eq. (54), we prove the desired result:

$$\|f - \Pi_{\le k} g\|_\infty \le \|f - \Pi_{\le k} f\|_\infty + \|\Pi_{\le k} g - \Pi_{\le k} f\|_\infty \le \epsilon. \tag{64}$$

■

A.5. Proof of Lemma 7

In this section we prove Lemma 7.

Proof of Lemma 7 In the following, we prove that for any $T < N_{k,d}$, there exists $f \in \mathcal{F}_k$ such that $\Pr_{f,n}(\|\hat{f}_T - f\|_\infty < \beta_k/4) < 1/2$, and the desired result follows directly.

Suppose at round i the algorithm query $x_i \in \mathbb{S}^{d-1}$ and receive $y_i = f(x_i) + \mathcal{N}(0,1)$ where f is the ground-truth. At round T, the algorithm outputs $\hat{f}_T$. Let $\Pr_{u,n}(\cdot)$ be the probability space of $(x_1, y_1, \cdots, x_T, y_T)$ when the ground-truth is $f = \beta_k P_{k,d}(\langle \cdot, u\rangle)$, and $\Pr_{0,n}(\cdot)$ the space when the ground-truth is $f = 0$. We use $\mathbb{E}_{u,n}$ and $\mathbb{E}_{0,n}$ to denote the corresponding expectation, respectively. Let $\mathcal{H}_i$ be the σ-field of random variable $(x_1, y_1, \cdots, x_{i-1}, y_{i-1}, x_i)$.

For every $u \in \mathbb{S}^{d-1}$, let $E_{u,n} \triangleq \mathbb{I}\left[\|\hat{f}_T - \beta_k P_{k,d}(\langle\cdot,u\rangle)\|_\infty < \beta_k/4\right]$ be the event that $\hat{f}_T$ is close to $\beta_k P_{k,d}(\langle \cdot, u\rangle)$. By Pinsker's inequality and chain rule of KL divergence, we have

$$\mathbb{E}_{u,n}[E_{u,n}] \le \mathbb{E}_{0,n}[E_{u,n}] + D_{\mathrm{TV}}(\Pr_{0,n} \,\|\, \Pr_{u,n}) \tag{65}$$

$$\le \mathbb{E}_{0,n}[E_{u,n}] + \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\Pr_{0,n} \,\|\, \Pr_{u,n})} \tag{66}$$

$$= \mathbb{E}_{0,n}[E_{u,n}] + \sqrt{\frac{1}{2}\mathbb{E}_{0,n}\left[\sum_{i=1}^n D_{\mathrm{KL}}(\Pr_{0,n}(y_i \mid \mathcal{H}_i) \| \Pr_{u,n}(y_i \mid \mathcal{H}_i))\right]} \tag{67}$$

$$= \mathbb{E}_{0,n}[E_{u,n}] + \sqrt{\frac{\beta_k^2}{4}\mathbb{E}_{0,n}\left[\sum_{i=1}^n P_{k,d}(x_i^\top u)^2\right]}. \tag{68}$$

Consequently,

$$\mathbb{E}_{u\sim\mathbb{S}^{d-1}}[\mathbb{E}_{u,n}[E_{u,n}]] \le \mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{E}_{0,n}[E_{u,n}] + \sqrt{\frac{\beta_k^2}{4}\mathbb{E}_{0,n}\left[\sum_{i=1}^n P_{k,d}(x_i^\top u)^2\right]}\right] \tag{69}$$

$$\leq \mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{E}_{0,n}[E_{u,n}]\right] + \sqrt{\frac{\beta_k^2}{4}\mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{E}_{0,n}\left[\sum_{i=1}^{n} P_{k,d}(x_i^\top u)^2\right]\right]} \tag{70}$$

$$= \mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{E}_{0,n}[E_{u,n}]\right] + \sqrt{\frac{\beta_k^2}{4}\mathbb{E}_{0,n}\left[\sum_{i=1}^{n} \mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[P_{k,d}(x_i^\top u)^2\right]\right]} \tag{71}$$

$$= \mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{E}_{0,n}[E_{u,n}]\right] + \sqrt{\frac{\beta_k^2}{4}\mathbb{E}_{0,n}\left[\frac{n}{N_{k,d}}\right]} \tag{72}$$

$$= \mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{E}_{0,n}[E_{u,n}]\right] + \sqrt{\frac{\beta_k^2}{4}\frac{n}{N_{k,d}}}. \tag{73}$$

Now we upper bound the first term in Eq. (73). Let $\hat{u}_T = \min_{u\in\mathbb{S}^{d-1}} \|\hat{f}_T - P_{k,d}(\langle\cdot,u\rangle)\|_\infty$. Consider the event $E'_{u,n} = \mathbb{I}\left[\|P_{k,d}(\langle\cdot,u\rangle) - P_{k,d}(\langle\cdot,\hat{u}_n\rangle)\|_\infty < \beta_k/2\right]$. In the following we prove that $\neg E'_{u,n} \implies \neg E_{u,n}$. Indeed, when $\|P_{k,d}(\langle\cdot,u\rangle) - P_{k,d}(\langle\cdot,\hat{u}_n\rangle)\|_\infty \geq \beta_k/2$ we get

$$\|\hat{f}_T - P_{k,d}(\langle\cdot,u\rangle)\|_\infty \geq \frac{1}{2}\left(\|\hat{f}_T - P_{k,d}(\langle\cdot,u\rangle)\|_\infty + \|\hat{f}_T - P_{k,d}(\langle\cdot,\hat{u}_n\rangle)\|_\infty\right) \tag{By the optimality of $\hat{u}_T$}$$

$$\geq \frac{1}{2}\|P_{k,d}(\langle\cdot,u\rangle) - P_{k,d}(\langle\cdot,\hat{u}_n\rangle)\|_\infty \geq \beta_k/4. \tag{Triangle inequality}$$

Therefore, we get

$$\mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{E}_{0,n}[E_{u,n}]\right] \leq \mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{E}_{0,n}[E'_{u,n}]\right] \tag{74}$$

$$= \mathbb{E}_{0,n}\left[\mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{I}\left[\|P_{k,d}(\langle\cdot,u\rangle) - P_{k,d}(\langle\cdot,\hat{u}_n\rangle)\|_\infty \leq \beta_k/4\right]\right]\right] \tag{75}$$

$$\leq \mathbb{E}_{0,n}\left[\mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{I}\left[|P_{k,d}(\langle u,u\rangle) - P_{k,d}(\langle u,\hat{u}_n\rangle)| \leq \beta_k/4\right]\right]\right] \tag{76}$$

$$= \mathbb{E}_{0,n}\left[\mathbb{E}_{u\sim\mathbb{S}^{d-1}}\left[\mathbb{I}\left[|1 - P_{k,d}(\langle u,\hat{u}_n\rangle)| \leq \beta_k/4\right]\right]\right] \tag{77}$$

$$\leq \mathbb{E}_{0,n}\left[\Pr_{u\sim\mathbb{S}^{d-1}}\left(P_{k,d}(\langle u,\hat{u}_n\rangle) \geq 1 - \beta_k/4\right)\right] \tag{78}$$

$$\leq \frac{16}{9N_{k,d}} \leq \frac{1}{4}. \tag{Proposition 25}$$

Finally, when $T < N_{k,d}\beta_k^{-2}$ we have

$$\min_{f\in\mathcal{F}_k} \Pr_{f,n}(\|\hat{f}_T - f\|_\infty < \beta_k/4) = \min_{u\in\mathbb{S}^{d-1}} \mathbb{E}_{u,n}[E_{u,n}] \leq \mathbb{E}_{u\in\mathbb{S}^{d-1}}\mathbb{E}_{u,n}[E_{u,n}] < \frac{1}{2}. \tag{79}$$

■

A.6. Proof of Theorem 8

In this section we present the proof of Theorem 8.

Proof of Theorem 8 When $\epsilon > d^{-7/4}$ the lower bound is trivial. Hence we focus on the regime $\epsilon < d^{-7/4}$.

Let k be the largest even number smaller than $\frac{d^{1/8}}{\sqrt{480\epsilon}}$ and $\tau_k = \left\langle \mathrm{ReLU}, \bar{P}_{k,d}\right\rangle_{\mu_d}$. First we prove that the set $\mathcal{F}_k \triangleq \{\tau_k P_{k,d}(\langle\cdot,u\rangle) : u \in \mathbb{S}^{d-1}\}$ belongs to NN-ReLU(L_2).

To this end, we prove that for every $f \in \mathcal{F}_k$, we can construct $c : \mathbb{S}^{d-1} \to \mathbb{R}$ such that $\|c\|_2 \leq 1$ and $f(x) = \mathbb{E}_{\xi \in \mathbb{S}^{d-1}}[\mathrm{ReLU}(\xi^\top x)c(\xi)]$ for every $x \in \mathbb{S}^{d-1}$. For every $f = \tau_k P_{k,d}(\langle \cdot, u\rangle) \in \mathcal{F}_k$, by Funk-Hecke formula (Theorem 1) we have

$$f(x) = \tau_k P_{k,d}(\langle \cdot, u\rangle) = \sqrt{N_{k,d}}\mathbb{E}_{\xi \in \mathbb{S}^{d-1}}[\mathrm{ReLU}(\xi^\top x)P_{k,d}(\langle \cdot, u\rangle)] = \mathbb{E}_{\xi \in \mathbb{S}^{d-1}}[\mathrm{ReLU}(\xi^\top x)\bar{P}_{k,d}(\langle \cdot, u\rangle)]. \tag{80}$$

Since $\|\bar{P}_{k,d}(\langle \cdot, u\rangle)\|_2 = 1$, we get $f \in \text{NN-ReLU}(L_2)$.

In the following we prove the desired result by invoking Lemma 7 with the hypothesis $\mathcal{F}_k$. First of all, by Lemma 21 and the definition of k we get

$$\tau_k/4 > \frac{d^{1/4}}{480k^{5/4}(k+d)^{3/4}} \geq \frac{d^{1/4}}{480k^2} \geq \epsilon. \tag{81}$$

Therefore, Lemma 7 implies that the minimax sample complexity is at least $N_{k,d}\tau_k^{-2}$. By basic Lemma 21 and algebra we have

$$N_{k,d}\tau_k^{-2} \geq \binom{k+d-2}{d-2}\frac{k^{5/2}(k+d)^{3/2}}{1200d^{1/2}} \gtrsim \left(\frac{k}{d-2}+1\right)^{d-2}\frac{k^{5/2}(k+d)^{3/2}}{d^{1/2}} \tag{82}$$

$$\geq \left(\frac{k}{d}\right)^d = (0.002\epsilon^{-1}d^{-7/4})^{d/2}, \tag{83}$$

which proves the desired result. ■

A.7. Missing Propositions

In this section we state and prove the missing propositions in Section 4.

Proposition 13 *Let $h : [-1,1] \to \mathbb{R}$ be a one-dimensional function satisfies $\sup_{t\in[-1,1]} |h^{(k)}(t)| \leq 1, \forall k \geq 0$. Then*

$$\|\Pi_k h(\langle \cdot, u\rangle)\|_2 \leq 2N_{k,d}^{-1/2}. \tag{84}$$

Proof For a fixed $u \in \mathbb{S}^{d-1}$, by the completeness of the Legendre polynomial basis, we have

$$h(\langle \cdot, u\rangle) = \sum_{k\geq 0} \tau_k \bar{P}_{k,d}(\langle \cdot, u\rangle), \tag{85}$$

where $\tau_k \triangleq \left\langle h, \bar{P}_{k,d}\right\rangle_{\mu_d}$. Since $\bar{P}_{k,d}(\langle \cdot, u\rangle) \in \mathbb{Y}_{k,d}$, it follows that

$$\|\Pi_k \exp(\langle \cdot, u\rangle)\|_2 = \tau_k\|\bar{P}_{k,d}(\langle \cdot, u\rangle)\|_2 = \tau_k. \tag{86}$$

As a result, we only need to prove that

$$\tau_k \leq N_{k,d}^{-1/2}, \quad \forall k \geq 0. \tag{87}$$

By Rodrigues formula Atkinson and Han (2012, Proposition 2.26) we get

$$\tau_k = \int_{-1}^{1} h(t)\bar{P}_{k,d}(t)\mu_d(t)\mathrm{d}t = \frac{\sqrt{N_{k,d}}\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}\int_{-1}^{1} h(t)P_{k,d}(t)(1-t^2)^{\frac{d-3}{2}}\mathrm{d}t \tag{88}$$

$$= \frac{\sqrt{N_{k,d}}\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}\frac{\Gamma\left(\frac{d-1}{2}\right)}{2^k\Gamma\left(k+\frac{d-1}{2}\right)}\int_{-1}^{1} h^{(k)}(t)(1-t^2)^{k+\frac{d-3}{2}}\mathrm{d}t \tag{89}$$

$$\leq \frac{\sqrt{N_{k,d}}\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}2^k\Gamma\left(k+\frac{d-1}{2}\right)}\int_{-1}^{1} |h^{(k)}(t)|(1-t^2)^{k+\frac{d-3}{2}}\mathrm{d}t \tag{90}$$

$$\leq \frac{\sqrt{N_{k,d}}\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}2^k\Gamma\left(k+\frac{d-1}{2}\right)}\int_{-1}^{1} (1-t^2)^{k+\frac{d-3}{2}}\mathrm{d}t \tag{91}$$

$$\leq \frac{\sqrt{N_{k,d}}\Gamma\left(\frac{d}{2}\right)}{2^k\Gamma\left(k+\frac{d-1}{2}\right)}\frac{\Gamma\left(k+\frac{d-1}{2}\right)}{\Gamma\left(k+\frac{d}{2}\right)} = \frac{\sqrt{N_{k,d}}\Gamma\left(\frac{d}{2}\right)}{2^k\Gamma\left(k+\frac{d}{2}\right)}. \tag{92}$$

As a result, we only need to prove $\frac{\Gamma\left(\frac{d}{2}\right)}{2^k\Gamma\left(k+\frac{d}{2}\right)} \leq 2N_{k,d}^{-1}$ and then Eq. (87) follows directly.

By the recursive formula of Γ function we get

$$\frac{2^k\Gamma\left(k+\frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} = 2^k\prod_{l=1}^{k}\left(k+\frac{d}{2}-l\right) = \prod_{l=1}^{k}(2k+d-2l)\,. \tag{93}$$

By the definition of $N_{k,d}$ we have

$$N_{k,d} = \frac{2k+d-2}{k+d-2}\binom{k+d-2}{k} \leq \frac{2\prod_{l=1}^{k}(k+d-1-l)}{k!}. \tag{94}$$

Observe that for any $l \in [1,k]$, $k+d-1-l \leq 2k+d-2l$. Consequently,

$$N_{k,d} \leq \frac{2\prod_{l=1}^{k}(k+d-1-l)}{k!} \leq 2\prod_{l=1}^{k}(2k+d-2l) = 2\frac{2^k\Gamma\left(k+\frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}. \tag{95}$$

Equivalently,

$$\frac{\Gamma\left(\frac{d}{2}\right)}{2^k\Gamma\left(k+\frac{d}{2}\right)} \leq 2N_{k,d}^{-1}. \tag{96}$$

■

Proposition 14 *Let* NN-Exp(L_p) *be the family of two layer NNs with activation* $\exp(\cdot)$ *and L_p norm bounds. Then any function $f \in$* NN-Exp(L_1) *satisfies* $\|\Pi_k f\|_2 \leq 2eN_{k,d}^{-1/2}$.

Proof Recall that if two functions f, g satisfies $\|\Pi_k f\|_2 \leq N_{k,d}^{-1/2}$ and $\|\Pi_k g\|_2 \leq N_{k,d}^{-1/2}$, their convex combinations $h = \theta f + (1-\theta)g$ also satisfies $\|\Pi_k g\|_2 \leq N_{k,d}^{-1/2}$. Since any function in NN-Exp(L_1) can be written as a convex combination of functions $\{\pm\exp(\langle\cdot,u\rangle) : u \in \mathbb{S}^{d-1}\}$, we only need to prove that $\|\Pi_k \exp(\langle\cdot,u\rangle)\|_2 \leq N_{k,d}^{-1/2}$ for every $u \in \mathbb{S}^{d-1}$.

Let $h(t) = e^{-1}\exp(t)$. Then we have $\sup_{t\in[-1,1]}|h^{(k)}(t)| \le 1$. Invoking Proposition 13 we get

$$\|\Pi_k \exp(\langle\cdot,u\rangle)\|_2 = e\|\Pi_k h(\langle\cdot,u\rangle)\|_2 \le 2eN_{k,d}^{-1/2}. \tag{97}$$

■

Proposition 15 *Suppose the function f satisfies $\|\Pi_k f\|_2 = \Omega(1)N_{k,d}^{-\alpha/2}, \forall k \ge 0$ for some constant $\alpha > 0$. For any inner product kernel $K(x,x')$ on the sphere where $\sup_{x,x'\in\mathbb{S}^{d-1}}|K(x,x')| \le 1$, f has a infinite RKHS norm induced by K when $\alpha < 1/2$.*

Proof Since $K(x,x')$ is a bounded inner product kernel, we can write $K(x,x') = h(\langle x,x'\rangle)$ for some one-dimensional function $h : [-1,1] \to [-1,1]$. Let λ_k be the eigenvalues of kernel K. By the Funk-Hecke formula (Theorem 1) we get

$$\lambda_k = N_{k,d}^{-1/2}\left\langle h, \bar{P}_{k,d}\right\rangle_{\mu_d} \le N_{k,d}^{-1/2}\|h\|_{\mu_d}\|\bar{P}_{k,d}\|_{\mu_d} \le N_{k,d}^{-1/2}\|h\|_\infty\|\bar{P}_{k,d}\|_{\mu_d} \le N_{k,d}^{-1/2}. \tag{98}$$

Since $\mathbb{Y}_{k,d}$ is the space of eigenfunctions of kernel K corresponding to the eigenvalue λ_k, the RKHS norm of f is defined by

$$\|f\|_K^2 = \sum_{k\ge0}\frac{\|\Pi_k f\|_2^2}{\lambda_k} \ge \sum_{k\ge0}\|\Pi_k f\|_2^2 N_{k,d}^{1/2}. \tag{99}$$

As a result, when $\alpha < 1/2$ we get

$$\|f\|_K^2 \gtrsim \sum_{k\ge0} N_{k,d}^{1/2-\alpha} = \infty. \tag{100}$$

■

Appendix B. Learning with Two-layer Finite-width Neural Networks

In this section, we show that using a finite-width two-layer neural network with polynomial activation can also achieve a small L_∞-error bound.

Our algorithm is presented as Alg. 2. On a high level, given any desired error level $\epsilon > 0$, the algorithm selects a truncation threshold $k \ge 0$ (Line 1), and use empirical risk minimization to find two-layer neural network g with polynomial activation that fits the ground-truth f the best. The activation $\sigma_k : [-1,1] \to \mathbb{R}$ is the degree-k approximation of the ReLU activation in the Legendre polynomial space, given by

$$\sigma_k(t) \triangleq \textstyle\sum_{l=0}^{k}\left\langle \mathrm{ReLU}, \bar{P}_{l,d}\right\rangle_{\mu_d}\bar{P}_{l,d}(t). \tag{101}$$

Since $\bar{P}_{k,d}(\langle w,\cdot\rangle) \in \mathbb{Y}_{k,d}$ for every $k \ge 0, w \in \mathbb{S}^{d-1}$, any two-layer NN with activation σ_k is a degree-k polynomial (more precisely, it is the projection of a two-layer ReLU network with the same parameters to the space $\mathbb{Y}_{\le k,d}$). Hence, our algorithm essentially aims to find the best low-degree approximation of the ground-truth f using noisy data.

The following theorem states the sample complexity of Alg. 2

Theorem 16 *Suppose the ground-truth function satisfies Conditions 1 and 2 for some fixed $\alpha \in (0,1]$ and $c_1, c_2 > 0$. If $d \ge 10\alpha^{-1}2^{5/\alpha} + 2$, then for any $\epsilon > 0, \delta > 0$, with probability at least $1-\delta$ over the randomness of the data, Alg. 2 outputs a function g such that $\|f - g\|_\infty \le \epsilon$ using $O(\mathrm{poly}(c_1c_2, d, 1/\epsilon, \ln 1/\delta)^{1/\alpha})$ samples.*

Algorithm 2 L_∞-learning via Two-layer NNs with Polynomial Activation

```
Input: parameters $\alpha, c_1, c_2 > 0$, desired error level $\epsilon > 0$, and failure probability $\delta > 0$.
Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \sim \mathbb{S}^{d-1}$ are independent and uniformly sampled from the unit sphere $\mathbb{S}^{d-1}$, and $y_i = f(x_i) + \mathcal{N}(0,1)$.
Set the truncation threshold $k \leftarrow \inf_{l \geq 0}\{2c_1c_2(l+1)^{3/2}(N_{l+1,d})^{-\alpha/2} \leq \epsilon/2\}$.
Set the parameters for the neural network: norm bound $B = 35c_1\sqrt{d}\left(\frac{4c_1c_2}{\epsilon}\right)^{3+4/\alpha}$, and width $m \leftarrow 256B^2\epsilon^{-6/\alpha-2}(4c_1c_2)^{6/\alpha}d^{8/\alpha}$.
Define the family of two-layer NNs with polynomial activation $\sigma_k$ defined in Eq. (101):
    $\mathcal{F}_k = \left\{g(x) = \sum_{j=1}^m a_j\sigma_k(w_j^\top x) : w_j \in \mathbb{S}^{d-1}, \sum_{j=1}^m |a_j| \leq B\right\}.$
Run empirical risk minimization and get $g = \operatorname{argmin}_{f \in \mathcal{F}_k} \sum_{i=1}^n (f(x_i) - y_i)^2$.
Return $g$.
```

B.1. Proof of Theorem 16

Proof of Theorem 16 Let $\epsilon_1 = \frac{1}{4}\epsilon^{3/\alpha+1}(4c_1c_2)^{-3/\alpha}d^{-4/\alpha}$. We prove Theorem 3 in the following two steps.

Step 1: upper bound the population L_2 loss. In this step, we use classic statistical learning tools to show that the ERM step (i.e., $g = \operatorname{argmin}_{h \in \mathcal{F}_k} \sum_{i=1}^n (h(x_i) - y_i)^2$) returns a function g with small L_2 loss. In particular, by Lemma 17 we get $\|\Pi_{\leq k}(f-g)\|_2 \leq \epsilon_1$.

Step 2: upper bound the L_∞-error via truncation. This step is exactly the same as in the proof of Theorem 3.

Combining these two steps we prove the desired result. ■

Lemma 17 *Suppose the function $f : \mathbb{S}^{d-1} \to \mathbb{R}$ satisfies Conditions 1 and 2 for some fixed $\alpha \in (0,1], c_1, c_2 > 0$. For any $\epsilon > 0$, let $k = \inf_{l \geq 0}\{2c_1c_2(l+1)^{3/2}(N_{l+1,d})^{-\alpha/2} \leq \epsilon/2\}$.*

For any $\epsilon_1 > 0$, let $B = 35c_1\sqrt{d}\left(\frac{4c_1c_2}{\epsilon}\right)^{3+4/\alpha}$, $\sigma_k(t) = \sum_{l=0}^k \left\langle \mathrm{ReLU}, \bar{P}_{l,d}\right\rangle_{\mu_d} \bar{P}_{l,d}(t)$, and $m = 16B^2/\epsilon_1^2$, define the function class

$$\mathcal{F}_k = \left\{h(x) = \sum_{j=1}^m a_j\sigma_k(w_j^\top x) : w_j \in \mathbb{S}^{d-1}, \sum_{j=1}^m |a_j| \leq B\right\}. \tag{102}$$

For a given dataset $\{(x_i, y_i)\}_{i=1}^n$, let $\hat{\mathcal{L}}(h) \triangleq \frac{1}{N}\sum_{i=1}^n (h(x_i) - y_i)^2$ be the empirical L_2 loss, and $g = \operatorname{argmin}_{h \in \mathcal{F}} \hat{\mathcal{L}}(h)$.

For any $\delta > 0$, when $d \geq \max\{2e, 4/\alpha\}$ and $n \geq \Omega(\mathrm{poly}(d, (c_1c_2)^{1/\alpha}, \epsilon^{-1/\alpha}, \ln(1/\delta), 1/\epsilon_1))$, with probability at least $1-\delta$,

$$\|\Pi_{\leq k}(f-g)\|_2 = \|\Pi_{\leq k}f - g\|_2 \leq \epsilon_1. \tag{103}$$

B.2. Proof of Lemma 17

In this section, we prove Lemma 17.

Proof of Lemma 17 First we prove that there exists $\hat{f} \in \mathcal{F}$ such that the population loss is small. Since $\mathcal{F} \subseteq \mathbb{Y}_{\leq k,d}$, we get

$$\forall h \in \mathcal{F}, \quad \|h - f\|_2^2 = \|h - \Pi_{\leq k} f\|_2^2 + \|f - \Pi_{\leq k} f\|_2^2. \tag{104}$$

By Lemma 19, $\Pi_{\leq k} f$ can be represented by a infinite-width two-layer ReLU neural network with weight c such that $\|c\|_1 \leq 35 c_1 \sqrt{d} \left(\frac{4c_1c_2}{\epsilon}\right)^{3+4/\alpha} = B$. By Lemma 28, when $m > 16B^2/\epsilon_1^2$ there exists a finite-width approximation $\hat{f} \in \mathcal{F}$ such that $\|\hat{f} - \Pi_{\leq k} f\|_2 \leq \epsilon_1/2$.

In the following we show that ERM outputs a function $g \in \mathcal{F}$ such that

$$\|g - f\|_2^2 \leq \|\hat{f} - f\|_2^2 + \epsilon_1^2/2. \tag{105}$$

By the uniform convergence of two-layer neural networks (Lemma 18), when

$$n \geq \Omega(\mathrm{poly}(d, (c_1c_2)^{1/\alpha}, \epsilon^{-1/\alpha}, \ln(1/\delta), 1/\epsilon_1))$$

we have

$$\|g - f\|_2^2 \leq \hat{\mathcal{L}}(g) + \epsilon_1^2/4 \leq \hat{\mathcal{L}}(\hat{f}) + \epsilon_1^2/4 \leq \|\hat{f} - f\|_2^2 + \epsilon_1^2/2. \tag{106}$$

Combining with Eq. (104) we get

$$\|g - \Pi_{\leq k} f\|_2^2 \leq \|\hat{f} - \Pi_{\leq k} f\|_2^2 + \epsilon_1^2/2 < \epsilon_1^2. \tag{107}$$

■

The following lemma proves the uniform convergence result for the function class used in Lemma 17.

Lemma 18 *In the setting of Lemma 17, when $n \geq \Omega(\mathrm{poly}(B, N_{k,d}, \ln(1/\delta), 1/\epsilon_1))$, for any $\delta > 0$, with probability at least $1 - \delta$ we have*

$$\sup_{g \in \mathcal{F}} \left| \|g - f\|_2^2 - \hat{\mathcal{L}}(g) \right| \leq \epsilon_1. \tag{108}$$

Proof The proof is essentially the same as the proof of Lemma 11, with the only difference that here we use the Rademacher complexity upper bound for two-layer neural networks (Bartlett and Mendelson, 2002, Theorem 18). ■

The following lemma proves the realizability result for the function class used in Lemma 17.

Lemma 19 *In the setting of Lemma 17, $\Pi_{\leq k} f$ can be represented by an infinite-width two-layer ReLU neural network with weight $c : \mathbb{S}^{d-1} \to \mathbb{R}$ such that $\|c\|_1 \leq 35 c_1 \sqrt{d} \left(\frac{4c_1c_2}{\epsilon}\right)^{3+4/\alpha}$.*

Proof Recall that we can write $\Pi_{\leq k} f(x) = \sum_{l=0}^{k} \sum_{j=1}^{N_{k,d}} a_{l,j} Y_{l,j}(\cdot)$. Let

$$\lambda_l = N_{l,d}^{-1/2} \left\langle \mathrm{ReLU}, \bar{P}_{l,d} \right\rangle_{\mu_d} \tag{109}$$

and define the weight $c : \mathbb{S}^{d-1} \to \mathbb{R}$ by $c(x) = \sum_{l=0}^{k} \lambda_l^{-1} \sum_{j=1}^{N_{l,d}} a_{l,j} Y_{l,j}(\cdot)$. Then by the Funk-Hecke formula (Theorem 1) we get

$$\Pi_{\leq k} f(x) = \mathbb{E}_{w \sim \mathbb{S}^{d-1}} [\sigma(x^\top w) c(w)], \quad \forall x \in \mathbb{S}^{d-1}. \tag{110}$$

Hence, we only need to upper bound $\|c\|_2$, and then the desired result is proved by the fact that $\|c\|_1 \le \|c\|_2$.

Let $\beta_l^2 = \|\Pi_l f\|_2^2 = \sum_{j=1}^{N_{l,d}} a_{l,j}^2$ for all $l \in [0,k]$. Then we have

$$\begin{aligned}\|c\|_2^2 = \sum_{l=0}^{k} \lambda_l^{-2}\beta_l^2 \le \sum_{l=0}^{k} \lambda_l^{-2} c_1 N_{l,d}^{-\alpha} \le 1200 c_1^2 \sum_{l=0}^{k} N_{l,d}^{1-\alpha} l(l+d) \qquad &\text{(By Lemma 21)} \\ \le 1200 c_1^2 N_{k,d} k^2 (k+d). \qquad &(111)\end{aligned}$$

By the definition of k we have

$$2c_1c_2k^{3/2}(N_{k,d})^{-\alpha/2} > \epsilon/2. \tag{112}$$

Consequently,

$$N_{k,d} < \left(\frac{4c_1c_2}{\epsilon}\right)^{2/\alpha} k^{3/\alpha}. \tag{113}$$

Applying Proposition 29 we get $k \le \left(\frac{4c_1c_2}{\epsilon}\right)^{\frac{2}{d\alpha-3}}$. Using the assumption that $d > 4/\alpha$ we get

$$c_1^2 N_{k,d} k^2 (k+d) \le dc_1^2 \left(\frac{4c_1c_2}{\epsilon}\right)^{2/\alpha} k^{3+3/\alpha} \tag{114}$$

As a result,

$$\|c\|_1^2 \le \|c\|_2^2 \le 1200 dc_1^2 \left(\frac{4c_1c_2}{\epsilon}\right)^{6+8/\alpha}. \tag{115}$$

■

Appendix C. Decomposition of ReLU in the Legendre Polynomial Space

The following lemma analytically computes the spherical harmonics decomposition of ReLU activation (see also Bach (2017); Mhaskar (2006); Bourgain and Lindenstrauss (1988); Schneider (1967)).

Lemma 20 *Let $\tau_k = \left\langle \mathrm{ReLU}, \bar{P}_{k,d} \right\rangle_{\mu_d}$ be the projection of ReLU function to degree-k Legendre polynomial. Then we have*

$$\tau_k = \begin{cases} (-1)^{\frac{k-2}{2}} \sqrt{N_{k,d}} \frac{1}{2^k\sqrt{\pi}} \frac{\Gamma(d/2)\Gamma(k-1)}{\Gamma(k/2)\Gamma((k+d+1)/2)}, & \textit{when } k \textit{ is even,} \\ \frac{1}{2\sqrt{d}}, & \textit{when } k = 1, \\ 0, & \textit{when } k > 1 \textit{ and } k \textit{ is odd.} \end{cases} \tag{116}$$

Proof Recall that by definition,

$$\tau_k = \int_{-1}^{1} \mathrm{ReLU}(t)\bar{P}_{k,d}(t)\mu_d(t)\mathrm{d}t = \sqrt{N_{k,d}}\int_{0}^{1} tP_{k,d}(t)\mu_d(t)\mathrm{d}t. \tag{117}$$

When k is odd we have $P_{k,d}(-t) = -P_{k,d}(t)$. As a result,

$$\int_{0}^{1} tP_{k,d}(t)\mu_d(t)\mathrm{d}t = \frac{1}{2}\int_{-1}^{1} tP_{k,d}(t)\mu_d(t)\mathrm{d}t. \tag{118}$$

Recall that $P_{1,d}(t) = t$, and we have

$$\int_{-1}^{1} tP_{k,d}(t)\mu_d(t)\mathrm{d}t = \int_{-1}^{1} P_{1,d}(t)P_{k,d}(t)\mu_d(t)\mathrm{d}t = \frac{1}{N_{k,d}}\mathbb{I}\left[k=1\right].$$

It follows directly that (1) $\tau_k = 0$ if $k > 1$ and k is odd, and (2) $\tau_1 = \frac{1}{2\sqrt{N_{1,d}}} = \frac{1}{2\sqrt{d}}$.

Now we focus on the case when k is even. By the Rodrigues representation formula (Atkinson and Han, 2012, Theorem 2.23) we get

$$P_{k,d}(t) = (-1)^k \frac{\Gamma(\frac{d-1}{2})}{2^k\Gamma(k+\frac{d-1}{2})}(1-t^2)^{-\frac{d-3}{2}}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^k (1-t^2)^{k+\frac{d-3}{2}}. \tag{119}$$

As a result,

$$\begin{aligned}
&\int_{0}^{1} tP_{k,d}(t)\mu_d(t)\mathrm{d}t && (120)\\
=&(-1)^k \frac{\Gamma(\frac{d-1}{2})}{2^k\Gamma(k+\frac{d-1}{2})}\frac{\Gamma(d/2)}{\Gamma((d-1)/2)}\frac{1}{\sqrt{\pi}}\int_{0}^{1} t\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^k (1-t^2)^{k+\frac{d-3}{2}}\mathrm{d}t && (121)\\
=&(-1)^{k+1} \frac{\Gamma(\frac{d-1}{2})}{2^k\Gamma(k+\frac{d-1}{2})}\frac{\Gamma(d/2)}{\Gamma((d-1)/2)}\frac{1}{\sqrt{\pi}}\int_{0}^{1} \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{k-1} (1-t^2)^{k+\frac{d-3}{2}}\mathrm{d}t \\
& && \text{(integration by parts)}\\
=&(-1)^{k+1} \frac{\Gamma(\frac{d-1}{2})}{2^k\Gamma(k+\frac{d-1}{2})}\frac{\Gamma(d/2)}{\Gamma((d-1)/2)}\frac{1}{\sqrt{\pi}}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{k-2} (1-t^2)^{k+\frac{d-3}{2}}\Bigg|_0^1 && (122)\\
=&(-1)^{k} \frac{\Gamma(\frac{d-1}{2})}{2^k\Gamma(k+\frac{d-1}{2})}\frac{\Gamma(d/2)}{\Gamma((d-1)/2)}\frac{1}{\sqrt{\pi}}\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{k-2} (1-t^2)^{k+\frac{d-3}{2}}\Bigg|_{t=0}. && (123)
\end{aligned}$$

By binomial theorem, we have

$$\begin{aligned}
&\left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{k-2} (1-t^2)^{k+\frac{d-3}{2}}\Bigg|_{t=0} = \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^{k-2} \sum_{j=0}^{k+\frac{d-3}{2}} \binom{k+\frac{d-3}{2}}{j}(-1)^j t^{2j}\Bigg|_{t=0} && (124)\\
=&(-1)^{\frac{k-2}{2}}(k-2)!\binom{k+\frac{d-3}{2}}{\frac{k-2}{2}} = (-1)^{\frac{k-2}{2}}\frac{\Gamma(k-1)\Gamma(k+\frac{d-1}{2})}{\Gamma(k/2)\Gamma(\frac{k+d+1}{2})}. && (125)
\end{aligned}$$

Combining Eq. (123) and Eq. (125) we get

$$\int_0^1 tP_{k,d}(t)\mu_d(t)\mathrm{d}t \tag{126}$$

$$=(-1)^k\frac{\Gamma(\frac{d-1}{2})}{2^k\Gamma(k+\frac{d-1}{2})}\frac{\Gamma(d/2)}{\Gamma((d-1)/2)}\frac{1}{\sqrt{\pi}}(-1)^{\frac{k-2}{2}}\frac{\Gamma(k-1)\Gamma(k+\frac{d-1}{2})}{\Gamma(k/2)\Gamma(\frac{k+d+1}{2})} \tag{127}$$

$$=(-1)^{\frac{k-2}{2}}\frac{\Gamma(\frac{d}{2})\Gamma(k-1)}{2^k\Gamma(\frac{k}{2})\Gamma(\frac{k+d+1}{2}))}\frac{1}{\sqrt{\pi}}. \tag{128}$$

Finally, combining with Eq. (117) we prove the desired result. ■

Lemma 21 *Let $\tau_k = \left\langle \mathrm{ReLU}, \bar{P}_{k,d}\right\rangle_{\mu_d}$ be the projection of ReLU to degree-k Legendre polynomial. Then we have $|\tau_k| = \Theta(d^{1/4}k^{-5/4}(k+d)^{-3/4})$. In particular, for all dimension $d \geq 3$ and even degree $k \geq 4$ the following upper and lower bounds hold:*

$$\frac{2^{5/4}\pi^{3/4}}{\exp(13/2)}d^{1/4}k^{-5/4}(k+d)^{-3/4} \leq |\tau_k| \leq \frac{\exp(13/2)}{2\pi^2}d^{1/4}k^{-5/4}(k+d)^{-3/4}. \tag{129}$$

Proof Recall that Stirling's formula states

$$\sqrt{2\pi}k^{k+1/2}e^{-k} \leq \Gamma(k+1) \leq ek^{k+1/2}e^{-k}. \tag{130}$$

We first prove the upper bound. By Lemma 20, when k is even we have

$$|\tau_k| = \sqrt{N_{k,d}}\frac{1}{2^k\sqrt{\pi}}\frac{\Gamma(d/2)\Gamma(k-1)}{\Gamma(k/2)\Gamma((k+d+1)/2)} \tag{131}$$

$$= \sqrt{\frac{2k+d-2}{k+d-2}\frac{\Gamma(k+d-1)}{\Gamma(k+1)\Gamma(d-1)}}\frac{1}{2^k\sqrt{\pi}}\frac{\Gamma(d/2)\Gamma(k-1)}{\Gamma(k/2)\Gamma((k+d+1)/2)} \tag{132}$$

$$\leq \frac{\sqrt{2}}{\sqrt{\pi}}\left(\sqrt{\frac{\Gamma(k+d-1)}{\Gamma(k+1)\Gamma(d-1)}}\frac{1}{2^k}\frac{\Gamma(d/2)\Gamma(k-1)}{\Gamma(k/2)\Gamma((k+d+1)/2)}\right) \tag{133}$$

$$\leq \frac{\sqrt{2}}{\sqrt{\pi}}\left(\sqrt{\frac{e}{2\pi}\frac{(k+d-2)^{k+d-\frac{3}{2}}}{k^{k+\frac{1}{2}}(d-2)^{d-\frac{3}{2}}}}\frac{1}{2^k}\frac{\exp(7/2)}{2\pi}\frac{(\frac{d}{2}-1)^{\frac{d-1}{2}}(k-2)^{k-\frac{3}{2}}}{(\frac{k}{2}-1)^{\frac{k-1}{2}}(\frac{k+d-1}{2})^{\frac{k+d}{2}}}\right) \tag{134}$$

$$\leq \frac{\exp(4)}{2\pi^2}\left(\exp(5/2)\sqrt{\frac{(k+d)^{k+d-\frac{3}{2}}}{k^{k+\frac{1}{2}}d^{d-\frac{3}{2}}}}\frac{1}{2^k}\frac{(\frac{d}{2})^{\frac{d-1}{2}}k^{k-\frac{3}{2}}}{(\frac{k}{2})^{\frac{k-1}{2}}(\frac{k+d}{2})^{\frac{k+d}{2}}}\right) \quad \text{(Since } (1-1/t)^t = \Theta(1)\text{)}$$

$$\leq \frac{\exp(13/2)}{2\pi^2}\left(\sqrt{\frac{(k+d)^{k+d-\frac{3}{2}}}{k^{k+\frac{1}{2}}d^{d-\frac{3}{2}}}}\frac{d^{\frac{d-1}{2}}k^{k-\frac{3}{2}}}{k^{\frac{k-1}{2}}(k+d)^{\frac{k+d}{2}}}\right) \tag{135}$$

$$\leq \frac{\exp(13/2)}{2\pi^2}\left((k+d)^{-3/4}k^{-5/4}d^{1/4}\right). \tag{136}$$

Now we prove the lower bound. Similarly,

$$|\tau_k| = \sqrt{N_{k,d}}\frac{1}{2^k\sqrt{\pi}}\frac{\Gamma(d/2)\Gamma(k-1)}{\Gamma(k/2)\Gamma((k+d+1)/2)} \tag{137}$$

$$= \sqrt{\frac{2k+d-2}{k+d-2}\frac{\Gamma(k+d-1)}{\Gamma(k+1)\Gamma(d-1)}}\frac{1}{2^k\sqrt{\pi}}\frac{\Gamma(d/2)\Gamma(k-1)}{\Gamma(k/2)\Gamma((k+d+1)/2)} \tag{138}$$

$$\geq \frac{1}{\sqrt{\pi}}\left(\sqrt{\frac{\Gamma(k+d-1)}{\Gamma(k+1)\Gamma(d-1)}}\frac{1}{2^k}\frac{\Gamma(d/2)\Gamma(k-1)}{\Gamma(k/2)\Gamma((k+d+1)/2)}\right) \tag{139}$$

$$\geq \frac{1}{\sqrt{\pi}}\left(\sqrt{\frac{\sqrt{2\pi}}{e^2}\frac{(k+d-2)^{k+d-\frac{3}{2}}}{k^{k+\frac{1}{2}}(d-2)^{d-\frac{3}{2}}}}\frac{1}{2^k}\frac{2\pi}{\exp(1/2)}\frac{(\frac{d}{2}-1)^{\frac{d-1}{2}}(k-2)^{k-\frac{3}{2}}}{(\frac{k}{2}-1)^{\frac{k-1}{2}}(\frac{k+d-1}{2})^{\frac{k+d}{2}}}\right) \tag{140}$$

$$\geq \frac{2^{5/4}\pi^{3/4}}{\exp(3/2)}\left(\exp(-5)\sqrt{\frac{(k+d)^{k+d-\frac{3}{2}}}{k^{k+\frac{1}{2}}d^{d-\frac{3}{2}}}}\frac{1}{2^k}\frac{(\frac{d}{2})^{\frac{d-1}{2}}k^{k-\frac{3}{2}}}{(\frac{k}{2})^{\frac{k-1}{2}}(\frac{k+d}{2})^{\frac{k+d}{2}}}\right) \quad (\text{Since } (1-1/t)^t = \Theta(1))$$

$$\geq \frac{2^{5/4}\pi^{3/4}}{\exp(13/2)}\left(\sqrt{\frac{(k+d)^{k+d-\frac{3}{2}}}{k^{k+\frac{1}{2}}d^{d-\frac{3}{2}}}}\frac{d^{\frac{d-1}{2}}k^{k-\frac{3}{2}}}{k^{\frac{k-1}{2}}(k+d)^{\frac{k+d}{2}}}\right) \tag{141}$$

$$\geq \frac{2^{5/4}\pi^{3/4}}{\exp(13/2)}\left((k+d)^{-3/4}k^{-5/4}d^{1/4}\right). \tag{142}$$

■

Appendix D. Random Spherical Harmonics

In this section, we prove the L_∞-norm bound for random spherical harmonics. Burq and Lebeau (2014) prove a similar result without explicitly computes the d-dependency in Eq. (14).

Proof of Lemma 6 For any fixed $x \in \mathbb{S}^{d-1}$, by Lemma 22 we get

$$g(x) = \sqrt{N_{k,d}}\mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[g(\xi)\bar{P}_{k,d}(x^\top\xi)]. \tag{143}$$

Since $\{Y_{k,j}\}_{j=1}^{N_{k,d}}$ is a set of orthonormal basis, there exists weights $\{u_j\}_{j=1}^{N_{k,d}}$ (that depends on x) such that

$$\bar{P}_{k,d}(x^\top\xi) = \sum_{j=1}^{N_{k,d}} u_j Y_{k,j}(\xi), \quad \forall \xi \in \mathbb{S}^{d-1}, \tag{144}$$

and $\sum_{j=1}^{N_{k,d}} u_j^2 = \mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[\bar{P}_{k,d}(x^\top\xi)^2] = 1$. Define $a = [a_1, \cdots, a_{N_{k,d}}] \in \mathbb{R}^{N_{k,d}}$ and $u = [u_1, \cdots, u_{N_{k,d}}] \in \mathbb{R}^{N_{k,d}}$. Then we have

$$g(x) = \sqrt{N_{k,d}}\mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[g(\xi)\bar{P}_{k,d}(x^\top\xi)] = \sqrt{N_{k,d}}\mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}\left[\left(\sum_{j=1}^{N_{k,d}} a_j Y_{k,j}(\xi)\right)\left(\sum_{j=1}^{N_{k,d}} u_j Y_{k,j}(\xi)\right)\right]$$

$$= \sqrt{N_{k,d}}\sum_{j=1}^{N_{k,d}} a_j u_j = \sqrt{N_{k,d}}a^\top u. \tag{145}$$

In addition, $\|g\|_2^2=\sum_{j=1}^{N_{k,d}} a_j^2=\|a\|_2^2$. Hence, by Lemma 27 we get

$$\forall t>0, \Pr\left(\frac{|a^\top u|}{\|a\|_2}\ge\frac{2t}{\sqrt{N_{k,d}}}\right)\le 3\exp(-t^2/2). \tag{146}$$

Equivalently,

$$\forall x\in\mathbb{S}^{d-1}, t>0, \quad \Pr\left(|g(x)|\ge 2t\|g\|_2\right)\le 3\exp(-t^2/2). \tag{147}$$

In the following, we upper bound $|g(x)|/\|g\|_2$ uniformly over all $x\in\mathbb{S}^{d-1}$ by the covering number argument and uniform concentration.

Let $h(x)=g(x)/\|g\|_2$. First we prove that $h(x)$ is Lipschitz on $\mathbb{S}^{d-1}$ with respect to the great-circle distance $d(x,y)\triangleq\arccos(x^\top y)$. To this end, we only need to upper bound the manifold gradient $\nabla_x^\star h(x)$ on the sphere. By Eq. (143) we get,

$$\|\nabla_x^\star h(x)\|_2=\frac{1}{\|g\|_2}\sqrt{N_{k,d}}\|\mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[g(\xi)\nabla_x^\star\bar{P}_{k,d}(x^\top\xi)]\|_2 \tag{148}$$

$$\le\frac{1}{\|g\|_2}\sqrt{N_{k,d}}\mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[|g(\xi)|\|\nabla_x^\star\bar{P}_{k,d}(x^\top\xi)\|_2] \tag{149}$$

$$\le\frac{1}{\|g\|_2}\sqrt{N_{k,d}}\left(\mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[g(\xi)^2]\mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[\|\nabla_x^\star\bar{P}_{k,d}(x^\top\xi)\|_2^2]\right)^{1/2} \quad \text{(Cauchy-Schwarz inequality)}$$

$$\le\sqrt{N_{k,d}}\sqrt{k(k+d-2)}. \quad \text{(Atkinson and Han (2012, Proposition 3.6))}$$

which implies that $h(x)$ is $(\sqrt{k(k+d-2)N_{k,d}})$-Lipschitz.

Let $\epsilon=(2\sqrt{k(k+d-2)N_{k,d}})^{-1}$ and $\mathcal{C}$ an ϵ-covering of $\mathbb{S}^{d-1}$ with respect to the great-circle distance. By Proposition 23 we get $|\mathcal{C}|\le(3/\epsilon)^d$. In addition, for every $x\in\mathbb{S}^{d-1}$ there exists $\hat{x}\in\mathcal{C}$ such that

$$|h(x)-h(\hat{x})|\le\sqrt{k(k+d-2)N_{k,d}}\epsilon=\frac{1}{2}. \tag{150}$$

By union bound and Eq. (147), with probability at least $1-\delta$ we get,

$$\forall x\in\mathcal{C}, \quad |h(x)|\le 4\sqrt{\ln\frac{3|\mathcal{C}|}{\delta}}\le 4\sqrt{\ln(3/\delta)+d\ln(3/\epsilon)} \tag{151}$$

$$\le 4\sqrt{\ln(3/\delta)+2d^2\ln(k+1)}. \tag{152}$$

Combining with Eq. (150) we get, with probability at least $1-\delta$,

$$\forall x\in\mathbb{S}^{d-1}, \quad \frac{|g(x)|}{\|g\|_2}\le 5\sqrt{\ln(3/\delta)+2d^2\ln(k+1)}, \tag{153}$$

which proves the desired result. ■

The following lemma is an realization of Riesz representation theorem.

Lemma 22 *For any fixed $k\ge 0$ and $f\in\mathbb{Y}_{k,d}$, we have*

$$f(x)=\sqrt{N_{k,d}}\mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[f(\xi)\bar{P}_{k,d}(x^\top\xi)], \quad \forall x\in\mathbb{S}^{d-1}. \tag{154}$$

Proof let $\{Y_{k,j}\}_{j=1}^{N_{k,d}}$ be any set of orthonormal basis for degree k spherical harmonics $\mathbb{Y}_{k,d}$. By addition theorem Atkinson and Han (2012, Theorem 2.9), for any $\xi \in \mathbb{S}^{d-1}$ we get

$$\sqrt{N_{k,d}}\bar{P}_{k,d}(x^\top \xi) = \sum_{j=1}^{N_{k,d}} Y_{k,j}(x)Y_{k,j}(\xi). \tag{155}$$

Since $\{Y_{k,j}\}_{j=1}^{N_{k,d}}$ is a set of orthonormal basis, there exists unique coefficients $\{a_j\}_{j=1}^{N_{k,d}}$ such that $f(x) = \sum_{j=1}^{N_{k,d}} a_j Y_{k,j}(x), \quad \forall x \in \mathbb{S}^{d-1}$. As a result,

$$\sqrt{N_{k,d}}\mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}[f(\xi)\bar{P}_{k,d}(x^\top \xi)] = \mathbb{E}_{\xi\sim\mathbb{S}^{d-1}}\left[\left(\sum_{j=1}^{N_{k,d}} a_j Y_{k,j}(\xi)\right)\left(\sum_{j=1}^{N_{k,d}} Y_{k,j}(x)Y_{k,j}(\xi)\right)\right]$$

$$= \sum_{j=1}^{N_{k,d}} a_j Y_{k,j}(x) = f(x). \tag{156}$$

■

Appendix E. Helper Lemmas

In this section, we present some low-level helper lemmas.

Proposition 23 *Let $N(\epsilon)$ be the ϵ-covering number of $\mathbb{S}^{d-1}$ with respect to the great-circle distance $d(x, y) \triangleq \arccos(x^\top y)$. Then for any $\epsilon \in (0, 1)$ we have*

$$N(\epsilon) \leq (3/\epsilon)^d. \tag{157}$$

Proof Note that any $(\epsilon/2)$-cover of the unit ball B^d (w.r.t. the Euclidean distance) induces an ϵ-covering of the unit sphere $\mathbb{S}^{d-1}$ with the same size. As a result,

$$N(\epsilon) \leq \frac{(1+\epsilon/2)^d}{(\epsilon/2)^d} \leq \left(\frac{3}{\epsilon}\right)^d. \tag{158}$$

■

Proposition 24 *Let $I_\nu(z) \triangleq \sum_{j=0}^\infty \frac{1}{j!\Gamma(\nu+j+1)}\left(\frac{z}{2}\right)^{\nu+2j}$ be the modified Bessel function of the first kind. Then for every $\nu > 1$ we get*

$$\sqrt{2}e^{-1}\frac{e^\nu}{(2\nu)^{\nu+1/2}} \leq I_\nu(1) \leq \frac{e^{1/4}}{\sqrt{\pi}}\frac{e^\nu}{(2\nu)^{\nu+1/2}}. \tag{159}$$

Proof By algebraic manipulation we get,

$$I_\nu(1) = \sum_{j=0}^\infty \frac{1}{j!\Gamma(\nu+j+1)}\left(\frac{1}{2}\right)^{\nu+2j} \leq \frac{(1/2)^\nu}{\Gamma(\nu+1)}\sum_{j=0}^\infty \frac{(1/2)^{2j}}{j!} = \frac{(1/2)^\nu}{\Gamma(\nu+1)}e^{1/4}. \tag{160}$$

By Stirling's formula, we get

$$\frac{(1/2)^\nu}{\Gamma(\nu+1)}e^{1/4} \le \frac{(1/2)^\nu e^\nu}{\sqrt{2\pi}\nu^{\nu+1/2}}e^{1/4} = \frac{e^{1/4}}{\sqrt{\pi}}\frac{e^\nu}{(2\nu)^{\nu+1/2}}. \tag{161}$$

Similarly, we have

$$I_\nu(1) = \sum_{j=0}^{\infty} \frac{1}{j!\Gamma(\nu+j+1)}\left(\frac{1}{2}\right)^{\nu+2j} \ge \frac{(1/2)^\nu}{\Gamma(\nu+1)}. \tag{162}$$

Using Stirling's formula again we have,

$$\frac{(1/2)^\nu}{\Gamma(\nu+1)} \ge \frac{(1/2)^\nu e^\nu}{e\nu^{\nu+1/2}} = \sqrt{2}e^{-1}\frac{e^\nu}{(2\nu)^{\nu+1/2}}. \tag{163}$$

■

Proposition 25 *For any fixed $k \ge 0$, $u \in \mathbb{S}^{d-1}$, and $t > 0$ we have*

$$\Pr_{x\sim\mathbb{S}^{d-1}}(|P_{k,d}(x^\top u)| > t) \le \frac{1}{t^2 N_{k,d}}. \tag{164}$$

Proof By Markov ineqaulity we have

$$\Pr_{x\sim\mathbb{S}^{d-1}}(|P_{k,d}(x^\top u)| > t) = \Pr_{x\sim\mathbb{S}^{d-1}}(P_{k,d}(x^\top u)^2 > t^2) \tag{165}$$

$$\le t^{-2}\mathbb{E}_{x\sim\mathbb{S}^{d-1}}[P_{k,d}(x^\top u)^2] = t^{-2}N_{k,d}^{-1}, \tag{166}$$

which proves the desired result. ■

Lemma 26 (Lemma 1 of Laurent and Massart (2000)) *Let $a_1, \cdots, a_d$ be i.i.d. $\mathcal{N}(0,1)$ Gaussian variables. Then for any $t > 0$,*

$$\Pr\left(\sum_{i=1}^{d} a_i^2 \le d - 2\sqrt{dt}\right) \le \exp(-t). \tag{167}$$

Lemma 27 *Let $a = (a_1, \cdots, a_d) \sim \mathcal{N}(0, I)$ be a Gaussian vector and $u \in \mathbb{R}^d$ a fixed vector with unit norm. Then for any $t > 0$,*

$$\Pr\left(\frac{|\langle a, u\rangle|}{\|a\|_2} \ge \frac{2t}{\sqrt{d}}\right) \le 3\exp(-t^2/2). \tag{168}$$

Proof Since $a \sim \mathcal{N}(0, I)$ is a Gaussian vector, we have $\langle a, u\rangle \sim \mathcal{N}(0,1)$. Hence,

$$\Pr(|\langle a, u\rangle| > t) \le 2\exp(-t^2/2). \tag{169}$$

By Lemma 26 with $t = \frac{9d}{64}$, we also have

$$\Pr\left(\|a\|_2 \le \sqrt{d}/2\right) = \Pr\left(\|a\|_2^2 \le d/4\right) \le \exp\left(-\frac{9}{64}d\right) \le \exp\left(-\frac{1}{8}d\right). \tag{170}$$

By union bound we have

$$\Pr\left(\frac{|\langle a,u\rangle|}{\|a\|_2}\geq\frac{2t}{\sqrt{d}}\right)\leq\Pr\left(|\langle a,u\rangle|>t\right)+\Pr\left(\|a\|_2\leq\sqrt{d}/2\right) \tag{171}$$

$$\leq 2\exp(-t^2/2)+\exp\left(-\frac{1}{8}d\right). \tag{172}$$

Note that when $t>\sqrt{d}/2$, the desired result is trivial because $|\langle a,u\rangle|\leq\|a\|_2$ with probability 1. Therefore, when $t\leq\sqrt{d}/2$ we have

$$\Pr\left(\frac{|\langle a,u\rangle|}{\|a\|_2}\geq\frac{2t}{\sqrt{d}}\right)\leq 2\exp(-t^2/2)+\exp\left(-\frac{1}{8}d\right)\leq 3\exp(-t^2/2). \tag{173}$$

■

Lemma 28 *Let f be a infinite-width two-layer neural network with activation $\sigma:[-1,1]\to[-1,1]$, defined by*

$$f(x)=\mathbb{E}_{w\sim\mathbb{S}^{d-1}}[\sigma(x^\top w)c(w)] \tag{174}$$

for some weight $c:\mathbb{S}^{d-1}\to\mathbb{R}$ with $\|c\|_1<\infty$. For any $\epsilon>0$, there exists a neural network $\hat{f}$ with $m=4\|c\|_1^2/\epsilon^2$ neurons, defined by $\hat{f}(x)=\sum_{j=1}^m a_i\sigma(w_i^\top x)$, such that $\|\hat{f}-f\|_2\leq\epsilon$, and $\sum_{j=1}^m|a_i|\leq\|c\|_1$.

Proof We prove this theorem by probabilistic method. Let $p:\mathbb{S}^{d-1}\to\mathbb{R}_+$ be a function given by $p(w)=|c(w)|/\|c\|_1$. Then p is a probability density function. Let $m=4\|c\|_1^2/\epsilon^2$. We sample $w_1,\cdots,w_m$ independently from p and let $a_i=\operatorname{sign}(c(w_i))\frac{\|c\|_1}{m}$. Define the two-layer neural network $\hat{f}$ by $\hat{f}(x)\triangleq\sum_{j=1}^m a_i\sigma(w_i^\top x)$. In the following we prove that $\mathbb{E}[\|\hat{f}-f\|_2^2]\leq\epsilon^2$ where the expectation is taken over the random variables $w_1\cdots,w_m$.

For any fixed $x\in\mathbb{S}^{d-1}$, we have

$$\hat{f}(x)-f(x)=\sum_{j=1}^m a_i\sigma(w_i^\top x)-f(x)=\frac{\|c\|_1}{m}\sum_{j=1}^m\left(\operatorname{sign}(c(w_i))\sigma(w_i^\top x)-\frac{f(x)}{\|c\|_1}\right). \tag{175}$$

Let $X_i\triangleq\operatorname{sign}(c(w_i))\sigma(w_i^\top x)-\frac{f(x)}{\|c\|_1}$. By basic algebra we have

$$\mathbb{E}_{w_i}[X_i]=\int_{\mathbb{S}^{d-1}}p(w_i)\operatorname{sign}(c(w_i))\sigma(w_i^\top x)\mathrm{d}w_i-\frac{f(x)}{\|c\|_1} \tag{176}$$

$$=\frac{1}{\|c\|_1}\left(\int_{\mathbb{S}^{d-1}}|c(w_i)|\operatorname{sign}(c(w_i))\sigma(w_i^\top x)\mathrm{d}w_i-f(x)\right) \tag{177}$$

$$=\frac{1}{\|c\|_1}\left(\int_{\mathbb{S}^{d-1}}c(w_i)\sigma(w_i^\top x)\mathrm{d}w_i-f(x)\right)=0. \tag{178}$$

In addition, $|X_i|\leq|\operatorname{sign}(c(w_i))\sigma(w_i^\top x)|+|\frac{f(x)}{\|c\|_1}|\leq 2$. It follows that

$$\mathbb{E}_{\hat{f}}[(\hat{f}(x)-f(x))^2]=\frac{\|c\|_1^2}{m^2}\left(\sum_{j=1}^m X_i\right)^2=\frac{\|c\|_1^2}{m^2}\sum_{j=1}^m X_i^2\leq\frac{4\|c\|_1^2}{m}. \tag{179}$$

Consequently,

$$\mathbb{E}_{\hat{f}}[\|\hat{f} - f\|_2^2] = \mathbb{E}_{\hat{f}}[\mathbb{E}_{x \in \mathbb{S}^{d-1}}[\|\hat{f} - f\|_2^2]] \leq \frac{4\|c\|_1^2}{m} \leq \epsilon^2. \tag{180}$$

By the probabilistic method, there exists $\hat{f}$ such that $\|\hat{f} - f\|_2^2 \leq \epsilon^2$, which proves the first part of the lemma.

By construction, we also have

$$\sum_{j=1}^{m} |a_j| = \sum_{j=1}^{m} \frac{\|c\|_1}{m} \leq \|c\|_1 \tag{181}$$

almost surely, which proves the second part of the lemma. ■

Proposition 29 *For any fixed $\alpha \in (0,1], c_1, c_2 > 0, \epsilon > 0$, let*

$$k = \inf_{l \geq 0} \{2c_1c_2(l+1)^{3/2}(N_{l+1,d})^{-\alpha/2} \leq \epsilon/2\}.$$

When $d > \max\{2e, 4/\alpha\}$ we have $k \leq \max\{2e, (4c_1c_2/\epsilon)^{\frac{2}{d\alpha-3}}\}$ and $N_{k,d} \leq \left(\frac{4c_1c_2}{\epsilon}\right)^{8/\alpha}$.

Proof Let $c = c_1c_2$. By the definition k we get

$$2c(k+1)^{3/2} N_{k+1,d}^{-\alpha/2} \leq \epsilon/2. \tag{182}$$

Consequently, by the fact that $N_{k+1,d} = \binom{d+k}{d-1} - \binom{d+k-2}{d-1} \leq \binom{d+k+1}{d} \leq \left(\frac{e(d+k+1)}{d}\right)^d$ we get

$$4c(k+1)^{3/2} \leq \epsilon N_{k+1,d}^{\alpha/2} \leq \epsilon \left(\frac{e(k+d+1)}{d}\right)^{d\alpha/2}. \tag{183}$$

When $d \geq 2e$ and $k \geq 2e$, we get $\frac{e(k+d+1)}{d} \leq k+1$. As a result,

$$4c(k+1)^{3/2} \leq \epsilon(k+1)^{d\alpha/2}, \tag{184}$$

which implies that

$$k \leq \left(\frac{\epsilon}{4c_1c_2}\right)^{\frac{2}{3-d\alpha}} = \left(\frac{4c_1c_2}{\epsilon}\right)^{\frac{2}{d\alpha-3}}. \tag{185}$$

By the definition k we also have

$$2ck^{3/2} N_{k,d}^{-\alpha/2} > \epsilon/2. \tag{186}$$

Hence,

$$N_{k,d} \leq \left(\frac{4c}{\epsilon}\right)^{2/\alpha} k^{3/\alpha} \leq \left(\frac{4c}{\epsilon}\right)^{8/\alpha} \tag{187}$$

■

Proposition 30 *For any $k \geq 0$, let $\sigma_k(t) = \sum_{l=0}^{k} \left\langle \mathrm{ReLU}, \bar{P}_{l,d} \right\rangle_{\mu_d} \bar{P}_{l,d}(t)$. Then for all $k \geq 0$ we have*

$$\sup_{t \in [-1,1]} |\sigma_k(t)| \leq 1200\sqrt{N_{k,d}}, \tag{188}$$

$$\sup_{t \in [-1,1]} |\sigma_k'(t)| \leq 1200k\sqrt{N_{k,d}}. \tag{189}$$

Proof Let $\tau_l = \left\langle \text{ReLU}, \bar{P}_{l,d} \right\rangle_{\mu_d}$. Then we have

$$\sup_{t\in[-1,1]} |\sigma_k(t)| \leq \sum_{l=0}^{k} \tau_l \sup_{t\in[-1,1]} |\bar{P}_{l,d}(t)| = \sum_{l=0}^{k} \tau_l \sqrt{N_{l,d}} \leq k\tau_k \sqrt{N_{k,d}}. \tag{190}$$

By Lemma 21 we get $l\tau_l \leq 1200$. As a result,

$$\sup_{t\in[-1,1]} |\sigma_l(t)| \leq 1200\sqrt{N_{l,d}}. \tag{191}$$

By Atkinson and Han (2012, Eq. (2.89)) we have $\sup_{t\in[-1,1]} |\bar{P}'_{l,d}(t)| \leq \frac{l(l+d-2)}{d-1}$. As a result,

$$\sup_{t\in[-1,1]} |\sigma'_l(t)| \leq \sum_{l=0}^{k} \tau_l \sup_{t\in[-1,1]} |\bar{P}'_{l,d}(t)| = \sum_{l=0}^{k} \tau_l \frac{l(l+d-2)}{d-1} \sqrt{N_{l,d}} \tag{192}$$

$$\leq \tau_k \frac{k^2(k+d-2)}{d-1} \sqrt{N_{k,d}} \leq 1200k\sqrt{N_{k,d}}. \tag{193}$$

■