Shallow Neural Networks Learn Low-Degree Spherical Polynomials with Feature Learning by Learnable Channel Attention

Yingzhen Yang YINGZHEN.YANG@ASU.EDU
School of Computing and Augmented Intelligence, Arizona State University

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

We study the problem of learning a low-degree spherical polynomial of degree $\ell_0 = \Theta(1) \geq 1$ defined on the unit sphere in $\mathbb{R}^d$ by training an over-parameterized two-layer neural network (NN) with channel attention in this paper. Our main result is the significantly improved sample complexity for learning such low-degree polynomials. We show that, for any regression risk $\varepsilon \in (0, 1)$, a carefully designed two-layer NN with channel attention and finite width trained by the vanilla gradient descent (GD) requires the lowest sample complexity of $n \asymp \Theta(d^{\ell_0}/\varepsilon)$ with high probability, in contrast with the representative sample complexity $\Theta\left(d^{\ell_0} \max\left\{\varepsilon^{-2}, \log d\right\}\right)$, where n is the training data size. Moreover, such sample complexity is not improvable since the trained network renders a sharp rate of the nonparametric regression risk of the order $\Theta(d^{\ell_0}/n)$ with high probability. On the other hand, the minimax optimal rate for the regression risk with a kernel of rank $\Theta(d^{\ell_0})$ is $\Theta(d^{\ell_0}/n)$, so that the rate of the nonparametric regression risk of the network trained by GD is minimax optimal. The training of the two-layer NN with channel attention is a two-stage process. In stage one, a novel and provable learnable channel selection algorithm, as a learnable harmonic-degree selection process, is employed to select the ground truth channel number in the target function, ℓ_0, among the initial $L \geq \ell_0$ channels in its activation function in the first layer with high probability. Such learnable channel selection is performed by efficient one-step GD on both layers of the NN, which achieves the goal of feature learning in learning low-degree polynomials. In stage two, the second layer of the network is trained by standard GD using the activation function with selected channels. To the best of our knowledge, this is the first time a minimax optimal risk bound is obtained by training an over-parameterized but finite-width neural network with feature learning capability to learn low-degree spherical polynomials.

Keywords: Nonparametric Regression, Low-Degree Spherical Polynomial, Neural Network, Gradient Descent, Learnable Channel Attention, Feature Learning, Minimax Optimal Rate

1. Introduction

With deep learning achieving remarkable breakthroughs across a wide range of machine learning tasks (LeCun et al., 2015), understanding the generalization capability of neural networks has become a central topic in both statistical learning theory and theoretical deep learning. A large body of work has established that gradient descent (GD) and stochastic gradient descent (SGD) can provably minimize training loss in deep neural networks (DNNs) (Du et al., 2019b; Allen-Zhu et al., 2019; Du et al., 2019a; Arora et al., 2019; Zou and Gu, 2019; Su and Yang, 2019). Beyond optimization, many studies investigate the generalization behavior of DNNs trained via gradient-based methods, deriving algorithmic generalization bounds. A key insight from this line of work is that with sufficient over-parameterization, meaning a large number of neurons, training dynamics can be effectively described using a kernel method, particularly the Neural Tangent Kernel (NTK) (Jacot et al., 2018) determined by the network's architecture. Other results, such as (Yang and Hu,

2021), demonstrate that infinite-width neural networks can still perform feature learning. The NTK framework reveals that for highly over-parameterized models, the network weights stay close to initialization, enabling a linearized approximation via first-order Taylor expansion that facilitates generalization analysis (Cao and Gu, 2019; Arora et al., 2019; Ghorbani et al., 2021).

The generalization properties of neural networks can also be studied through the lens of learning low-degree polynomials. This direction is motivated by analyses of spectral bias in neural networks (Rahaman et al., 2019; Cao et al., 2021; Choraria et al., 2022), which show that neural networks tend to prioritize learning target functions lying within subspaces spanned by eigenfunctions associated with NTK eigenvalues. For example, on uniformly distributed data over the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$, degree-ℓ polynomials can be expressed via spherical harmonics up to degree ℓ, as formalized in Section B and Theorem 14. While (Yang and Hu, 2021) shows infinite-width networks can perform feature learning, several works attempt to overcome the linear NTK regime to learn low-degree polynomials on spheres in $\mathbb{R}^d$. The QuadNTK method introduced in (Bai and Lee, 2020) applies a second-order Taylor expansion to improve over NTK's linearization, achieving more effective learning of sparse "one-directional" polynomials. Extending this idea, (Nichani et al., 2022) shows that combining NTK and QuadNTK can capture dense polynomials with an additional sparse high-degree term. Further contributions include (Damian et al., 2022), which uses two-stage optimization for learning low-degree polynomials, and (Takakura and Suzuki, 2024), which explores feature learning in the mean-field regime.

Despite these advances, existing work on training over-parameterized neural networks to learn low-degree polynomials, such as (Ghorbani et al., 2021; Bai and Lee, 2020; Nichani et al., 2022; Damian et al., 2022; Takakura and Suzuki, 2024), often lacks sharp characterizations of regression risk. For instance, (Nichani et al., 2022) establishes that the regression risk ε holds when sample size $n \gtrsim d^{\ell_0} \max\{\varepsilon^{-2}, \log d\}$. Separately, (Ghorbani et al., 2021) shows that for $\tilde{\Theta}(d^{\ell_0}) \leq n \leq \Theta(d^{\ell_0+1-\delta})$ with $\tilde{\Theta}(d^{\ell_0})/d^{\ell_0} \to \infty$ as $d \to \infty$, NTK-based regression risk converges to zero under restrictive conditions, but no convergence rate or sharpness is established. Moreover, in practical settings where d is finite, which is commonly considered in sharp rate analyses for nonparametric regression (Hu et al., 2021; Suh et al., 2022; Li et al., 2024; Yang and Li, 2024; Yang, 2025), the results from (Ghorbani et al., 2021) fail to guarantee even the vanishing regression risk.

Understanding the sharpness of regression risk in learning low-degree polynomials remains a significant open problem in statistical learning theory and theoretical deep learning. Furthermore, it is an open problem how to explore the feature learning effect of neural networks in learning such polynomials with sharp rates. In this paper, we consider a target function f^* that belongs to the Reproducing Kernel Hilbert Space (RKHS) associated with a positive definite (PD) kernel induced by an over-parameterized two-layer NN, where f^* is a degree-ℓ_0 polynomial defined on the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$ with $\ell_0 = \Theta(1) \geq 1$. Our main result, Theorem 2, shows that training such a neural network using the vanilla GD achieves the minimax optimal nonparametric regression risk of the order $\Theta(d^{\ell_0}/n)$ with high probability. Comparatively, the minimax optimal rate for kernel regression risk with a positive definite kernel of rank $r_0 = \Theta(d^{\ell_0})$ is known to be $\Theta(r_0/n) = \Theta(d^{\ell_0}/n)$, as established in (Raskutti et al., 2012, Theorem 2(a)), indicating that our result is in fact minimax optimal. Our training algorithm includes two stages. In the first stage, a novel and provable learnable channel selection algorithm is employed to select the channels in the activation function in the first layer of the network by one-step GD, where each channel covers a particular degree of spherical harmonics. It is proved that the number of selected channels is the ground truth channel number, ℓ_0, in the target function. In the second stage, the second-layer weights are trained by GD

with the fixed activation function with selected channels in the first layer. Our analysis demonstrates the potential of a new combination of feature learning and NTK-based analysis, where the feature learning effect of the network is implemented by learnable channel attention, which is followed by training the over-parameterized network by GD in the NTK regime. The discussion of existing empirical and theoretical works about channel attention is deferred to Section E of the appendix. To the best of our knowledge, our work is among the first to reveal the theoretical benefit of channel attention with a novel and provable learnable channel selection algorithm for learning low-degree spherical polynomials with a minimax optimal rate.

Feature Learning Capability of Our Method. The feature learning capability of our training algorithm is in the training stage one, where the novel Algorithm 1 is used to decide the channel number of the activation function of the NN, which is guaranteed to be ℓ_0 with high probability (w.h.p) by Theorem 1. In this way, w.h.p. stage two performs kernel regression with the oracle kernel, achieving the minimax optimal rate. The estimation of ℓ_0 is an important goal of feature learning for learning the target function of degree-ℓ_0 with sharp regression risk. To see this, as a well-known fact widely discussed (Damian et al., 2022; Ghorbani et al., 2021), the target function lies in a subspace with all spherical harmonics of degree $\leq \ell_0$ as its orthogonal basis. Therefore, only the optimal sample complexity of $\Theta(d^{\ell_0}/\varepsilon)$ is required to learn such a target function with any risk $\varepsilon > 0$ by our training stage two w.h.p. Furthermore, an inaccurate and conservative estimate $\ell' > \ell_0$ leads to worse sample complexity $\Theta(d^{\ell'}/\varepsilon)$ compared to our optimal sample complexity. The literature studying the feature learning effect, such as (Lee et al., 2024; Damian et al., 2022), learns the features of the subspace that the target function lies in so as to achieve sharp regression risk. Our estimate of ℓ_0 achieves the feature learning effect under the similar principle. During the two-stage training, the kernel evolves as the activation function changes from stage one to two. Thanks to the feature learning capability of our method, our result is stronger than the literature in terms of learning general low-degree spherical polynomials. For example, existing works (Wei et al., 2019; Glasgow, 2024; Lee et al., 2024; Abbe et al., 2022) do not consider the regression problem where the target function is a degree-ℓ_0 spherical polynomial with our sharp and minimax optimal regression risk rate of $\Theta(d^{\ell_0}/n)$. In particular, (Wei et al., 2019) does not consider the regression problem with the target function being a polynomial. The results of (Glasgow, 2024) are limited to a very specific case where the target function is a quadratic XOR function. In (Lee et al., 2024), the target function is a single-index function $f^*(\mathbf{x}) = \sigma^*(\langle \mathbf{x}, \boldsymbol{\theta} \rangle)$ where the function σ^* has information exponent p, so that f^* is limited to be a polynomial of a particular direction, parameterized by $\boldsymbol{\theta}$, of the variable $\mathbf{x}$, instead of being a more general non-single-index spherical polynomial considered in this paper. (Abbe et al., 2022) studies the case that the target function f^* is a low-dimensional latent function of dimension P in the ambient space of dimension d with $P \leq d$, and shows necessary and nearly sufficient condition that f^* is strongly SGD-learnable in the mean-field regime.

We organize this paper as follows. We first introduce in Section 2 the problem setup. The training algorithm of the network is described in Section 3. Our main result is summarized in Section 4 with the novel training algorithm by GD and the sharp risk bound for learning low-degree spherical polynomials. The roadmap of proofs, the summary of the approaches and the key technical results in the proofs, and the novel proof strategies of this work are presented in Section 5.

Notations. We use bold letters for matrices and vectors, and regular lower letters for scalars throughout this paper. $\mathbf{A}^{(i)}$ is the i-th column of a matrix $\mathbf{A}$. A bold letter with subscripts indicates the corresponding rows or elements of a matrix or a vector. We put an arrow on top of a letter with subscript

if it denotes a vector, e.g., $\vec{\mathbf{x}}_i$ denotes the i-th training feature. $\|\cdot\|_F$ and $\|\cdot\|_p$ denote the Frobenius norm and the vector ℓ^p-norm or the matrix p-norm. $[m:n]$ denotes all the integers between m and n inclusively, and $[1:n]$ is also written as $[n]$. $\mathrm{Var}[\cdot]$ denotes the variance of a random variable. $\mathbf{I}_n$ is a $n \times n$ identity matrix. $\mathbb{1}_{\{E\}}$ is an indicator function which takes the value of 1 if event E happens, or 0 otherwise. The complement of a set A is denoted by A^c, and $|A|$ is the cardinality of the set A. $\mathrm{vec}(\cdot)$ denotes the vectorization of a matrix or a set of vectors, and $\mathrm{tr}(\cdot)$ is the trace of a matrix. We denote the unit sphere in d-dimensional Euclidean space by $\mathbb{S}^{d-1} \coloneqq \{\mathbf{x}\colon \mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 = 1\}$. Let $\mathcal{X}$ denote the input space, and $L^p(\mathcal{X}, \mu)$ with $p \geq 1$ denote the space of p-th power integrable functions on $\mathcal{X}$ with probability measure μ, and the inner product $\langle \cdot, \cdot \rangle_{L^p(\mu)}$ and $\|\cdot\|^2_{L^p(\mu)}$ are defined as $\langle f, g \rangle_{L^p(\mu)} \coloneqq \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\mu(x)$ and $\|f\|^p_{L^p(\mu)} \coloneqq \int_{\mathcal{X}} |f|^p(x)\mathrm{d}\mu(x) < \infty$. $\mathbf{B}(\mathbf{x}; r)$ is the Euclidean closed ball centered at $\mathbf{x}$ with radius r. Given a function $g\colon \mathcal{X} \to \mathbb{R}$, its L^∞-norm is denoted by $\|g\|_\infty \coloneqq \sup_{\mathbf{x}\in\mathcal{X}} |g(\mathbf{x})|$, and L^∞ is the function class whose elements have bounded L^∞-norm. $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ denote the inner product and the norm in the Hilbert space $\mathcal{H}$. $a = \mathcal{O}(b)$ or $a \lesssim b$ indicates that there exists a constant $c > 0$ such that $a \leq cb$. $\tilde{\mathcal{O}}$ indicates there are specific requirements in the constants of the $\mathcal{O}$ notation. $a = o(b)$ and $a = w(b)$ indicate that $\lim |a/b| = 0$ and $\lim |a/b| = \infty$, respectively. $a \asymp b$ or $a = \Theta(b)$ denotes that there exists constants $c_1, c_2 > 0$ such that $c_1 b \leq a \leq c_2 b$. $\mathrm{Unif}\left(\mathbb{S}^{d-1}\right)$ denotes the uniform distribution on $\mathbb{S}^{d-1}$. The constants defined throughout this paper may change from line to line. We use $\mathbb{E}_P[\cdot]$ to denote the expectation with respect to the distribution P. $\mathbb{P}_{\mathcal{S}}$ denotes the orthogonal projection onto the space $\mathcal{S}$, and $\mathrm{Span}(\mathbf{A})$ denotes the linear space spanned by the columns of the matrix $\mathbf{A}$. $\overline{A}$ denotes the closure of a set A. Throughout this paper we let the input space be $\mathcal{X} = \mathbb{S}^{d-1}$.

2. Problem Setup

We introduce the problem setups for nonparametric regression with the target function as a low-degree spherical polynomial in this section.

2.1. Two-Layer Neural Network with Channel Attention

We are given the training data $\left\{(\vec{\mathbf{x}}_i, y_i)\right\}_{i=1}^n$ where each data point is a tuple of feature vector $\vec{\mathbf{x}}_i \in \mathcal{X}$ and its response $y_i \in \mathbb{R}$. Throughout this paper we assume that no two training features coincide, that is, $\vec{\mathbf{x}}_i \neq \vec{\mathbf{x}}_j$ for all $i, j \in [n]$ and $i \neq j$. We denote the training feature vectors by $\mathbf{S} = \left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n$, and denote by P_n the empirical distribution over $\mathbf{S}$. The response y_i is given by $y_i = f^*(\vec{\mathbf{x}}_i) + w_i$ for $i \in [n]$, where $\{w_i\}_{i=1}^n$ are i.i.d. sub-Gaussian random variables as the noise with mean 0 and variance proxy σ_0^2, that is, $\mathbb{E}[\exp(\lambda w_i)] \leq \exp(\lambda^2\sigma_0^2/2)$ for any $\lambda \in \mathbb{R}$. f^* is the target function to be detailed later. We define $\mathbf{y} \coloneqq [y_1, \ldots, y_n]^\top$, $\mathbf{w} \coloneqq [w_1, \ldots, w_n]^\top$, and use $f^*(\mathbf{S}) \coloneqq \left[f^*(\vec{\mathbf{x}}_1), \ldots, f^*(\vec{\mathbf{x}}_n)\right]^\top$ to denote the clean target labels. The feature vectors in $\mathbf{S}$ are drawn i.i.d. according to the data distribution $P = \mathrm{Unif}\left(\mathbb{S}^{d-1}\right)$ with μ being the probability measure for P. We consider a two-layer linear neural network (NN) with channel attention in this paper whose mapping function is

$$f(\boldsymbol{\tau}, \boldsymbol{a}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \sigma_{\boldsymbol{\tau}}(\mathbf{x}, \vec{\mathbf{q}}_r), \tag{1}$$

where $\mathbf{x} \in \mathcal{X}$ is the input, $\mathbf{Q} = \left\{\vec{\mathbf{q}}_r\right\}_{r=1}^m$ are the random weights drawn i.i.d. according to $P = \mathrm{Unif}(\mathcal{X})$. $\sigma_{\boldsymbol{\tau}}$ is the activation function which is a PD kernel defined as

$$\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}') \coloneqq \sum_{\ell=0}^{L} \sum_{j=1}^{N(d,\ell)} \tau_\ell \mu_{\sigma,\ell} Y_{\ell,j}(\mathbf{x}) Y_{\ell,j}(\mathbf{x}'),\ \mu_{\sigma,\ell} = N^{-1}(d,\ell) \text{ for } 0 \le \ell \le L,\ \ \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \quad (2)$$

Here $\{Y_{\ell j}\}_{j \in [N(d,\ell)]}$ are the spherical harmonics of degree ℓ which form an orthogonal basis of $\mathcal{H}_\ell$ of dimension $N(d,\ell)$, and $\mathcal{H}_\ell$ denotes the space of degree-ℓ homogeneous harmonic polynomials on $\mathcal{X}$. The background about harmonic analysis on $\mathbb{S}^{d-1}$ is deferred to Section B of the appendix. Each $\mu_{\sigma,\ell} Y_{\ell,j}(\mathbf{x}) Y_{\ell,j}(\mathbf{x}')$ with $\ell \in [0:L]$ constitutes a channel in the output of the activation function, and $\boldsymbol{\tau} = \{\tau_\ell\}_{\ell=0}^L$ are the channel attention weights with $L+1$ channels. It is noted that in the two-layer NN (1), the first layer comprises the spherical harmonics as the activation functions with channel attention weights, and $\boldsymbol{a} = [a_1, \ldots, a_m] \in \mathbb{R}^m$ denotes the weights of the second layer. It follows from the background in harmonic analysis on spheres in Section B that for every given $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}')$ can be efficiently computed with $\Theta(L)$ time complexity through $\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=0}^{L} \tau_\ell P_\ell^{(d)}(\langle \mathbf{x}, \mathbf{x}' \rangle)$, where each channel, $P_\ell^{(d)}$, is the ℓ-th Gegenbauer polynomial which can be computed efficiently in $\Theta(1)$ time for each $\ell \in [0:L]$ by dynamic programming, as shown in Lemma 16 in Section B of the appendix. We let $L \ge \ell_0$. Intuitively, each $P_\ell^{(d)}$ covers the information about the spherical harmonics of degree ℓ, so that all the information in the target function is captured with $L \ge \ell_0$. With a constant $\ell_0 \in \Theta(1)$, it is always feasible to set $L \ge \ell_0$ with suitably large L, and the computation of $\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}')$ takes $\Theta(L) = \Theta(1)$ time when $L = \Theta(1)$.

We will first run a learnable channel selection algorithm described in Algorithm 1, which is essentially a learnable harmonic-degree selection algorithm to be detailed in Section 3, to keep only the first $\widehat{\ell}$ channels with the updated attention weights $\left\{\tau_\ell = \mu_{\sigma,\ell}^{-\frac{1}{2}}\right\}_{\ell=0}^{\widehat{\ell}}$, and $\widehat{\ell} \le L$. The activation function after applying such learnable channel selection becomes

$$\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=0}^{\widehat{\ell}} \tau_\ell P_\ell^{(d)}(\langle \mathbf{x}, \mathbf{x}' \rangle) = \sum_{\ell=0}^{\widehat{\ell}} \sum_{j=1}^{N(d,\ell)} \mu_{\sigma,\ell}^{\frac{1}{2}} Y_{\ell,j}(\mathbf{x}) Y_{\ell,j}(\mathbf{x}'). \quad (3)$$

The feature learning effect of the two-layer NN with channel attention (1) is that, the number of selected channels, $\widehat{\ell}$, is the ground truth channel number, ℓ_0, in the target function w.h.p., to be detailed in Section 3. With the updated activation function (3) after learnable channel selection, we will train the second-layer weights $\boldsymbol{a}$ by GD with fixed activation function $\sigma_{\boldsymbol{\tau}}$ in the first layer. Herein we define the following empirical kernel incurred during the training of the two-layer NN (1) with selected channels by GD,

$$\widehat{K}(\mathbf{x}, \mathbf{x}') = \frac{1}{m} \sum_{r=1}^{m} \sigma_{\boldsymbol{\tau}}(\mathbf{x}, \vec{\mathbf{q}}_r) \sigma_{\boldsymbol{\tau}}(\vec{\mathbf{q}}_r, \mathbf{x}'), \quad (4)$$

and its population version

$$K(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\mathbf{w} \sim \mathrm{Unif}(\mathcal{X})} \left[\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{w}) \sigma_{\boldsymbol{\tau}}(\mathbf{w}, \mathbf{x}')\right] = \sum_{\ell=0}^{\widehat{\ell}} \sum_{j=1}^{N(d,\ell)} \mu_{\sigma,\ell} Y_{\ell,j}(\mathbf{x}) Y_{\ell,j}(\mathbf{x}'). \quad (5)$$

K is in fact the NTK of the network (1) with respect to its second-layer weights $\boldsymbol{a}$. We denote by $\widehat{\mathbf{K}} \in \mathbb{R}^{n\times n}$ with $\widehat{\mathbf{K}}_{ij} = \widehat{K}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j)$ for $i, j \in [n]$ the gram matrix of $\widehat{K}$ over the training features $\mathbf{S}$, and let $\widehat{\mathbf{K}}_n = \widehat{\mathbf{K}}/n$. Similarly, the gram matrix of K is $\mathbf{K} \in \mathbb{R}^{n\times n}$ with $\mathbf{K}_{ij} = K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j)$ for $i, j \in [n]$, and $\mathbf{K}_n = \mathbf{K}/n$. Let the eigendecomposition of $\mathbf{K}_n$ be $\mathbf{K}_n = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$ where $\mathbf{U}$ is an $n \times n$ orthogonal matrix, and $\boldsymbol{\Sigma}$ is a diagonal matrix with its diagonal elements $\left\{\widehat{\lambda}_i\right\}_{i=1}^n$ being the eigenvalues of $\mathbf{K}_n$ and sorted in a non-increasing order. It follows from Lemma 28 that $\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} K(\mathbf{x}, \mathbf{x}') = \widehat{\ell} + 1$, so that it can be verified that $\widehat{\lambda}_1 \in (0, \widehat{\ell} + 1]$.

2.2. Kernel and Kernel Regression for Nonparametric Regression

Let $\mathcal{H}_K$ be the Reproducing Kernel Hilbert Space (RKHS) associated with K. Because K is of finite rank, the integral operator $T_K \colon L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu), (T_K f)(\mathbf{x}) \coloneqq \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') \mathrm{d}\mu(\mathbf{x}')$ is a positive, self-adjoint, and compact operator on $L^2(\mathcal{X}, \mu)$. By the spectral theorem and Lemma 28, the eigenfunctions of T_K are $\{Y_{\ell j}\}_{\ell\in[0:\widehat{\ell}], j\in[N(d,\ell)]}$, the spherical harmonics of degree up to $\widehat{\ell}$. $\mu_\ell = \mu_{\sigma,\ell} = N(d, \ell)^{-1}$ is the eigenvalue corresponding to the eigenspace $\mathcal{H}_\ell$, that is, $T_K Y_{\ell,j} = \mu_\ell Y_{\ell,j}$ for every $\ell \in [0 : \widehat{\ell}]$ and $j \in [N(d, \ell)]$. Let $\{\mu_\ell\}_{\ell\geq 0}$ be the distinct eigenvalues associated with T_K, and let m_ℓ be the be the sum of multiplicities of the eigenvalues $\{\mu_{\ell'}\}_{\ell'=0}^{\ell}$. That is, $m_\ell - m_{\ell-1}$ is the multiplicity of μ_ℓ with $m_{-1} = 0$. We define $r_0 \coloneqq m_{\ell_0} = \sum_{\ell=0}^{\ell_0} N(d, \ell)$ as the multiplicity of all the top $\ell_0 + 1$ distinct eigenvalues. For a positive constant γ_0, we define $\mathcal{H}_K(\gamma_0) \coloneqq \{f \in \mathcal{H}_K \colon \|f\|_{\mathcal{H}} \leq \gamma_0\}$ as the closed ball in $\mathcal{H}_K$ centered at 0 with radius γ_0. We note that $\mathcal{H}_K(\gamma_0)$ is also specified by $\mathcal{H}_K(\gamma_0) = \left\{f \in L^2(\mathcal{X}, \mu) \colon f = \sum_{\ell=0}^{\widehat{\ell}} \sum_{j=1}^{N(d,\ell)} \alpha_{\ell,j} Y_{\ell,j}, \sum_{\ell=0}^{\ell_0} \sum_{j=1}^{N(d,\ell)} \alpha_{\ell,j}^2/\mu_\ell \leq \gamma_0^2\right\}$. $\mathcal{H}_K(\gamma_0)$ is in fact formed by the union of the space of homogeneous harmonic polynomials up to degree $\widehat{\ell}$ with RKHS-norm γ_0, and $\mathcal{H}_K$ is a subspace of dimension $m_{\widehat{\ell}}$ in $L^2(\mathcal{X}, \mu)$. We define a PD kernel $K^{(r_0)}(\mathbf{x}, \mathbf{x}') \coloneqq \sum_{\ell=0}^{\ell_0} \sum_{j=1}^{N(d,\ell)} \mu_\ell Y_{\ell,j}(\mathbf{x}) Y_{\ell,j}(\mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, then $K^{(r_0)}$ is a low-rank kernel of rank r_0. It is also shown in Lemma 15 in Section B of the appendix that $r_0 = \Theta(d^{\ell_0})$.

The task of nonparametric regression. We consider the target function

$$f^*(\mathbf{x}) = \sum_{\ell=0}^{\ell_0} \sum_{j=1}^{N(d,\ell)} \beta_{\ell,j} Y_{\ell,j}(\mathbf{x}), \text{ s.t. } \sum_{\ell=0}^{\ell_0} \sum_{j=1}^{N(d,\ell)} \beta_{\ell,j}^2/\mu_\ell \leq \gamma_0^2, \quad \forall \mathbf{x} \in \mathcal{X}, \tag{6}$$

where $\ell_0 = \Theta(1) \geq 1$, and f^* lies in the space of homogeneous harmonic polynomials up to degree ℓ_0. It can be verified that $f^* \in \mathcal{H}_{K^{(r_0)}}(\gamma_0)$, and $\mathcal{H}_{K^{(r_0)}}(\gamma_0) \subseteq \mathcal{H}_K(\gamma_0)$ if $\widehat{\ell} \geq \ell_0$. The task of the analysis for nonparametric regression is to find an estimator $\widehat{f}$ from the training data $\left\{(\vec{\mathbf{x}}_i, y_i)\right\}_{i=1}^n$ so that the risk $\mathbb{E}_P\left[\left(\widehat{f} - f^*\right)^2\right]$ vanishes at a fast rate. In this work, we aim to establish a sharp rate of the risk where the over-parameterized neural network (1) trained by GD serves as the estimator $\widehat{f}$.

Minimax Lower Risk Bound for Learning a Low-Degree Spherical Polynomial. The established result in (Raskutti et al., 2012, Theorem 2(a)) gives the minimax optimal lower bound for kernel regression with the kernel K, that is, $\inf_{\widehat{f}_n} \sup_{f^*\in\mathcal{H}_{K^{(r_0)}}(\gamma_0)} \mathbb{E}_{\mathbf{x}}\left[\left(\widehat{f}_n(\mathbf{x}) - f^*(\mathbf{x})\right)^2\right] \gtrsim d^{\ell_0}/n$,

where the infimum is taken over all measurable functions of the training sample $\left\{\vec{\mathbf{x}}_i, y_i\right\}_{i=1}^n$. This result suggests that the minimax optimal lower bound for the regression risk with K is $\Theta(r_0/n) = \Theta(d^{\ell_0}/n)$, which is provably achieved by the two-layer NN (1) trained by GD, to be shown by our main result in the next section.

3. Training the Two-Layer Neural Network by Gradient Descent

In the training process of our two-layer NN (1), both the channel attention weights $\boldsymbol{\tau}$ and the second-layer weights $\boldsymbol{a}$ are optimized, and the first-layer weights $\mathbf{Q} = \left\{\vec{\mathbf{q}}_r\right\}_{r=1}^m$ are randomly sampled and then fixed during the training. The following quadratic loss function is minimized during the training process:

$$L(\boldsymbol{\tau}, \boldsymbol{a}) \coloneqq \frac{1}{2n} \sum_{i=1}^{n} \left(f(\boldsymbol{a}, \vec{\mathbf{x}}_i) - y_i\right)^2. \tag{7}$$

The training process of the two-layer NN (1) consists of two stages. In the first stage, one step of GD is applied to learn the channel attention weights $\boldsymbol{\tau}$. With the channel attention weights learned, the activation function is set to (3), that is, $\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=0}^{\widehat{\ell}} \mu_{\sigma,\ell}^{-\frac{1}{2}} P_\ell^{(d)}(\langle \mathbf{x}, \mathbf{x}' \rangle)$. We then train the second-layer weights $\boldsymbol{a}$ by minimizing the objective (7) through GD in the second training stage. We introduce the following notations for the training process. Let $\{Y_j\}_{j=0}^{m_L-1} = \{Y_{\ell j}\}_{0 \le \ell \le \widehat{\ell}, j \in [N(d,\ell)]}$ as the enumeration of all the spherical harmonics of up to degree L. We define $\mathbf{Y}(\mathbf{S}, m_L) \in \mathbb{R}^{n \times m_L}$ where $[\mathbf{Y}(\mathbf{S}, m_L)]_{ij} = Y_{j-1}(\vec{\mathbf{x}}_i)$ for every $i \in [n]$ and $j \in [m_L]$, $\mathbf{Y}(\mathbf{S}, r_0) = \mathbf{Y}(\mathbf{S}, m_{\ell_0}) \in \mathbb{R}^{n \times r_0}$ is defined similarly, and $\mathbf{Y}(\mathbf{S}, \ell) \in \mathbb{R}^{n \times N(d,\ell)}$ where $[\mathbf{Y}(\mathbf{S}, \ell)]_{ij} = Y_{\ell,j}(\vec{\mathbf{x}}_i)$ for all $i \in [n]$ and $j \in [N(d, \ell)]$. Similarly, $\mathbf{Y}(\mathbf{Q}, m_L) \in \mathbb{R}^{m \times m_L}$ with $[\mathbf{Y}(\mathbf{Q}, m_L)]_{rj} = Y_{j-1}(\vec{\mathbf{q}}_r)$ every $r \in [m]$ and $j \in [m_L]$, and $[\mathbf{Y}(\mathbf{Q}, \ell)]_{rj} = Y_{\ell,j}(\vec{\mathbf{q}}_r)$ for all $r \in [m]$ and $j \in [N(d, \ell)]$.

Training Stage One: Learning the Channel Attention Weights τ. We have the initialization $\boldsymbol{a}(0) = \mathbf{0}$ and $\tau_\ell(0) = 0$ for all $\ell \in [0 : L]$, where $\mathbf{0}$ denotes a vector whose elements are all 0. In this training stage, we first perform the one-step GD for $\boldsymbol{a}$ to obtain

$$\boldsymbol{a}(1) = \boldsymbol{a}(0) - \eta_1 \nabla_{\boldsymbol{a}} L(\boldsymbol{\tau}, \boldsymbol{a})|_{\boldsymbol{a}=\mathbf{0}, \boldsymbol{\tau}_\ell = \mu_{\sigma,\ell}^{-1}, \forall \ell \in [0:L]} = \frac{1}{n\sqrt{m}} \mathbf{Y}(\mathbf{Q}, m_L) \mathbf{Y}^\top(\mathbf{S}, m_L) \mathbf{y}, \tag{8}$$

where the learning rate $\eta_1 = 1$. $\boldsymbol{\tau}(1)$ is then obtained by one-step of GD with $\boldsymbol{a} = \boldsymbol{a}(1)$ by

$$\tau_\ell(1) = \tau_\ell(0) - \eta_2 \left. \frac{\mathrm{d}\partial L(\boldsymbol{\tau}, \boldsymbol{a})}{\partial \tau_\ell} \right|_{(\boldsymbol{\tau}, \boldsymbol{a}) = (\mathbf{0}, \boldsymbol{a}(1))} = \frac{1}{n\sqrt{m}} \mathbf{y}^\top \mathbf{Y}(\mathbf{S}, \ell) \mathbf{Y}^\top(\mathbf{Q}, \ell) \boldsymbol{a}(1) \tag{9}$$

for all $\ell \in [0 : L]$, where $\eta_2 = N(d, \ell)$. We note that the initialization of $\boldsymbol{\tau}(0) = \mathbf{0}$ is used in the one-step GD update for $\boldsymbol{\tau}(1)$ in (9), and a different initialization $\boldsymbol{\tau}(0)$ is used in (8). Theorem 1 below shows that w.h.p., when $\min\{n, m\} \ge \Theta(m_L) \log(4m_L/\delta)$, after performing the one-step GD update for the channel attention weights by (9), the channel attention weights of all the informative channels, defined as the channels with indices in $[0 : \ell_0]$, are not smaller than $2\varepsilon_0$ for a positive threshold $\varepsilon_0 \in (0, \beta_0^2/3]$. The absolute channel attention weights for the redundant channels, defined as the channels with indices in $[\ell_0 + 1 : L]$, are smaller than ε_0. As a result, Theorem 1 gives

the strong theoretical guarantee for a novel and principled learnable channel selection algorithm, described in Algorithm 1, which assigns updated attention weights $\mu_{\sigma,\ell}^{-\frac{1}{2}}$ to every informative channel with index ℓ, and assigns updated attention weights 0 to all redundant channels. We use $\widehat{\ell}$ to denote the number of channels with nonzero channel attention weights after running Algorithm 1, and Theorem 1 guarantees that $\widehat{\ell} = \ell_0$ in (3), the activation function after running the learnable channel selection by Algorithm 1. We note that Theorem 1 needs the minimum absolute value condition on the target function that $\min_{\ell\in[0:\ell_0],j\in[N(d,\ell)]]} |\boldsymbol{\beta}_{\ell,j}| \geq \beta_0\sqrt{\mu_{\sigma,\ell}}$ for some positive constant β_0. Due to the presence of noise in the response vector $\mathbf{y}$, similar minimum absolute value conditions on the target signal are in fact necessary and broadly used in standard compressive sensing literature such as (Aeron et al., 2010) for signal recovery.

Theorem 1 *Assume that the minimum absolute value condition on the target function holds, that is,* $\min_{\ell\in[0:\ell_0],j\in[N(d,\ell)]]} |\boldsymbol{\beta}_{\ell,j}| \geq \beta_0\sqrt{\mu_{\sigma,\ell}}$ *holds for some positive constant* β_0. ε_0 *is a positive threshold such that* $\varepsilon_0 \in (0, \beta_0^2/3]$. *Let* $\{\tau_\ell(1)\}_{\ell=0}^L = \boldsymbol{\tau}(1)$ *be computed by the one-step GD (9). Suppose that*

$$m > \max\left\{\frac{256\gamma_0^4}{\varepsilon_0^2}, 4\right\} m_L \log\left(\frac{4m_L}{\delta}\right), \tag{10}$$

$$n > \max\left\{\max\left\{\frac{400\gamma_0^4}{\varepsilon_0^2}, 4\right\} m_L \log\left(\frac{4m_L}{\delta}\right), \frac{32m_L(\sigma_0^2+1)}{\varepsilon_0}, \frac{8192\gamma_0^2 m_L(\sigma_0^2+1)}{\varepsilon_0^2}\right\}, \tag{11}$$

then for every $\delta \in (0,1)$, *with probability at least* $1 - \exp(-\Theta(m_L)) - \delta$, *we have*

$$\begin{cases} \tau_\ell(1) \geq 2\varepsilon_0, & \ell \in [0:\ell_0], \\ |\tau_\ell(1)| \leq \varepsilon_0, & \ell_0 < \ell \leq L. \end{cases} \tag{12}$$

Training Stage Two: Learning the Second-Layer Weights $\boldsymbol{a}$. We use GD to train the two-layer NN (1) with the channels attention weights updated in its activation function (3) in the first training stage. In the $(t+1)$-th step of GD with $t \geq 0$, the second-layer weights $\boldsymbol{a}$ are updated by one-step GD through

$$\boldsymbol{a}(t+1) = \boldsymbol{a}(t) - \frac{\eta}{n}\mathbf{Z}(t)(\widehat{\mathbf{y}}(t) - \mathbf{y}), \tag{13}$$

where $\mathbf{y}_i = y_i$, $\widehat{\mathbf{y}}(t) \in \mathbb{R}^n$ with $[\widehat{\mathbf{y}}(t)]_i = f(\boldsymbol{a}(t), \vec{\mathbf{x}}_i)$. We also denote $f(\boldsymbol{a}(t), \cdot)$ as $f_t(\cdot)$ as the neural network function with weighting vectors $\boldsymbol{a}(t)$ obtained right after the t-th step of GD. We define $\mathbf{Z}(t) \in \mathbb{R}^{r\times n}$ which is computed by $[\mathbf{Z}(t)]_{ri} = 1/\sqrt{m} \cdot \sigma_{\boldsymbol{\tau}}(\vec{\mathbf{x}}_i, \vec{\mathbf{q}}_r)$ for every $r \in [m]$ and $i \in [n]$ where $\sigma_{\boldsymbol{\tau}}$ is specified by (3). We employ the initialization $\boldsymbol{a}(0) = \mathbf{0}$ so that $\widehat{\mathbf{y}}(0) = \mathbf{0}$, that is, the initial output of the two-layer NN (1) is zero. The two-layer NN is trained by GD with T steps for $T \geq 1$. In the second training stage the channel attention weights $\boldsymbol{\tau}$ are not updated, so we abbreviate the two-layer NN (1) mapping function $f(\boldsymbol{\tau}, \boldsymbol{a}, \mathbf{x})$ as $f(\boldsymbol{a}, \mathbf{x})$.

4. Main Result

We present our main result about the sharp risk bound in Theorem 2, with its proof deferred to Section C.1 of the appendix.

Algorithm 1 Learnable Channel Selection

1: $\boldsymbol{\tau} \leftarrow$ Channel-Attention($\mathbf{S}, \mathbf{y}, \varepsilon_0$)
2: input: $\mathbf{S}, \mathbf{y}$
3: Compute the channel attention weights $\boldsymbol{\tau}(1) = \{\tau_\ell(1)\}_{\ell=0}^{L}$ by the one-step GD (9).
4: For each $\ell \in [0:L]$, set $\tau_\ell = \mathbb{1}_{\{\tau_\ell(1) \geq 2\varepsilon_0\}} \mu_{\sigma,\ell}^{-\frac{1}{2}}$.
5: **return** the channel attention weights $\boldsymbol{\tau} = \{\tau_\ell\}_{\ell=0}^{L}$

Algorithm 2 Training the Two-Layer NN by GD

1: $\boldsymbol{a}(T) \leftarrow$ Training-by-GD($T, \mathbf{Q}, \boldsymbol{a}$)
2: input: $T, \mathbf{Q}, \eta, \boldsymbol{a}(0) = \mathbf{0}$
3: **for** $t = 1, \ldots, T$ **do**
4: Perform the t-th step of GD by (13)
5: **end for**
6: **return** $\boldsymbol{a}(T)$

Theorem 2 *Suppose the minimum absolute value condition Theorem 1 holds, and $\widehat{\ell} \leq L$ nonzero attention weights are returned by the learnable channel selection algorithm described in Algorithm 1 with the threshold $\varepsilon_0 \in (0, \beta_0^2/3]$, $c_t \in (0,1]$ is an arbitrary positive constant. Suppose the network width m satisfies*

$$m \gtrsim \frac{n^4 \log(2n/\delta)}{d^{2\ell_0}}, \tag{14}$$

and the neural network $f(\boldsymbol{a}(t), \cdot)$ is trained by GD using Algorithm 2 with the constant learning rate $\eta = \Theta(1) \in (0, 1/(\ell_0+1))$ and $T \asymp n/d^{\ell_0}$. Then for every $t \in [c_t T : T]$ and every $\delta \in (0, 1/2)$, with probability at least $1 - 7\exp(-\Theta(r_0)) - \exp(-\Theta(n)) - \exp(-\Theta(m_L)) - 2\delta$ over the random noise $\mathbf{w}$, the random training features $\mathbf{S}$, and the random initialization $\mathbf{Q}$, $f(\boldsymbol{a}(t), \cdot) = f_t$ satisfies

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] \lesssim \Theta\left(\frac{d^{\ell_0}}{n}\right). \tag{15}$$

Here $r_0 = m_{\ell_0} = \Theta(d^{\ell_0})$.

Theorem 2 shows that the neural network (1) trained by GD enjoys a sharp rate of the regression risk for learning a degree-ℓ_0 spherical polynomial, $\Theta(d^{\ell_0}/n)$, which is minimax optimal as explained in Section 2.2. As an immediate result, (15) shows that the two-layer NN (1) trained GD enjoys a sample complexity of $n \asymp \Theta(d^{\ell_0}/\varepsilon)$ for any regression risk $\varepsilon \in (0,1)$, much lower than the sample complexity $\Theta\left(d^{\ell_0} \max\left\{\varepsilon^{-2}, \log d\right\}\right)$ in the representative work (Nichani et al., 2022). We herein compare our result with the competing results in learning low-degree spherical polynomials in Table 1 from the perspective of the sharpness of the regression risk and the algorithmic guarantees, that is, whether a finite-width neural network is trained to obtain the corresponding bound for the regression risk.

It is shown in (Nichani et al., 2022, Theorem 1) that a regression risk $\varepsilon > 0$ can be achieved with sample complexity $n \gtrsim d^{\ell_0} \max\left\{\varepsilon^{-2}, \log d\right\}$, implying a convergence rate of order $\Theta(\sqrt{d^{\ell_0}/n})$ when the regression risk is below $1/\sqrt{\log d}$. This rate is not minimax optimal and is considerably less sharp than our bound. The two-stage feature learning method of (Damian et al., 2022) requires the restrictive assumption that the target function depends only on $r \ll d$ input directions. Under this assumption, vanilla GD ensures that the learned network function lies in a subspace of rank r within the RKHS. Without it (i.e., $r = d$), the L^1-risk bound in (Damian et al., 2022, Theorem 1) is at least $\tilde{\Theta}(\sqrt{d^{\ell_0+1}/n})$. In contrast, since L^p-norm risks are non-decreasing in p, our L^2-risk bound in Theorem 2 immediately yields a sharper L^1-risk bound of $\Theta(\sqrt{d^{\ell_0}/n})$. Furthermore, (Ghorbani

Table 1: Comparison between our result and the existing works on learning low-degree polynomials on the spheres of $\mathbb{R}^d$ by training over-parameterized neural networks with or without algorithmic guarantees. Almost all the results here are under a common and popular setup that $f^* \in \mathcal{H}_{\tilde{K}}$ where $\tilde{K}$ is the NTK of a specific studied neural studied in each work, and the responses $\{y_i\}_{i=1}^n$ are corrupted by i.i.d. Gaussian noise with zero mean, with (Nichani et al., 2022) being the only exception where the responses are noise-free. It is remarked that the sample complexity can be straightforwardly obtained from the regression risk. The regression risk of (Damian et al., 2022, Theorem 1) is for the risk less than $1/\sqrt{\log d}$, with the meaning of r explained in Section 4, and $\tilde{\Theta}$ hides a logarithmic factor of $\log(mnd)$.

Existing Works and Our Result	Finite-Width NN is Trained	Sharpness of the Regression Risk
(Ghorbani et al., 2021, Theorem 4)	No	Only matching the lower bound for pointwise kernel learning, not minimax optimal
(Bai and Lee, 2020, Theorem 7)	Yes	Not minimax optimal
(Nichani et al., 2022, Theorem 1)	Yes	$\Theta(\sqrt{d^{\ell_0}/n})$, not minimax optimal
(Damian et al., 2022, Theorem 1)	Yes	L^1-norm regression risk $\tilde{\Theta}(\sqrt{dr^{\ell_0}/n} + \sqrt{r^p/m})$, not minimax optimal
Our Result (Theorem 2)	Yes	Minimax optimal, $\Theta\left(\frac{d^{\ell_0}}{n}\right)$

et al., 2021) shows that for $\tilde{\Theta}(d^{\ell_0}) \le n \le \Theta(d^{\ell_0+1-\delta})$ with $\tilde{\Theta}(d^{\ell_0})/d^{\ell_0} \to \infty$ as $d \to \infty$, the NTK-based regression risk converges to zero. However, their result requires restrictive conditions on the activation function and assumes infinite network width ($m \to \infty$). In sharp contrast, our result establishes that the minimax-optimal regression risk can be achieved by training finite-width neural networks with the feature learning capability by channel attention.

Beyond such feature learning approaches that aim to escape the linear NTK regime (Table 1), the statistical learning literature has long established sharp convergence rates for nonparametric kernel regression (Stone, 1985; Yang and Barron, 1999; Raskutti et al., 2014; Yuan and Zhou, 2016). In particular, training over-parameterized shallow (Hu et al., 2021, Theorem 5.2) or deep (Suh et al., 2022, Theorem 3.11) neural networks with spherical-uniform training features on the unit sphere achieves the minimax-optimal rate $\mathcal{O}(n^{-d/(2d-1)})$ for the regression risk, when the target function lies in $\mathcal{H}_{\tilde{K}}(\gamma_0)$ where $\tilde{K}$ is the NTK of the respective network.

As discussed in Section 2.2, since the target function f^* is a degree-ℓ_0 spherical polynomial, it lies in the union of eigenspaces up to degree ℓ_0. Therefore, learning requires identifying the subspace $\cup_{\ell=0}^{\ell_0} \mathcal{H}_\ell$ of dimension $r_0 = m_{\ell_0}$, rather than the full $L^2(\mathcal{X}, \mu)$. Crucially, with a carefully designed learnable channel selection algorithm described in Algorithm 1, the goal of feature learning is achieved by setting the number of channels in the activation function of the first layer to $\ell = \ell_0$ w.h.p. In this way, the NTK of the two-layer NN (1) in the second training stage becomes a low-rank kernel $K = K^{(r_0)}$ (5) of rank r_0, whose eigenspaces corresponding to nonzero eigenvalues span all and only spherical harmonics of degree up to ℓ_0. Consequently, vanilla GD on such a two-layer NN with sufficient width m can fit the target f^* using the r_0 eigenfunctions of K, thereby attaining the minimax-optimal regression rate. The roadmap for the proof of this main result is provided in Section 5, following the necessary background on kernel complexity.

5. Roadmap of Proofs

The summary of the approaches and key technical results in the proofs are presented as follows. We first introduce kernel complexity in Section 5.1, a key concept in our results and their proofs. Section 5.2 details the roadmap, key technical results in the proofs, our novel proof strategies and insights from our theoretical results.

5.1. Kernel Complexity

The local kernel complexity has been studied by (Bartlett et al., 2005; Koltchinskii, 2006; Mendelson, 2002). Let $\{\lambda_i\}_{i=1}^{m_{\widehat{\ell}}}$ be the enumeration of the distinct eigenvalues of the integral operator T_K, $\{\mu_\ell\}_{\ell=0}^{\widehat{\ell}}$, where each eigenvalue repeat as many times as its multiplicity in the sequence $\{\lambda_i\}_{i=1}^{m_{\widehat{\ell}}}$. We let $\lambda_i = 0$ for all $i > m_{\widehat{\ell}}$. For the PD kernel K, we define the empirical kernel complexity $\widehat{R}_K$ and the population kernel complexity R_K as

$$\widehat{R}_K(\varepsilon) \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \min\left\{\widehat{\lambda}_i, \varepsilon^2\right\}}, \quad R_K(\varepsilon) \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^{\infty} \min\left\{\lambda_i, \varepsilon^2\right\}}. \tag{16}$$

It can be verified that both $\sigma_0 R_K(\varepsilon)$ and $\sigma_0 \widehat{R}_K(\varepsilon)$ are sub-root functions (Bartlett et al., 2005) in terms of ε^2. The formal definition of sub-root functions is deferred to Definition 10 in the appendix. For a given noise ratio σ_0, the critical empirical radius $\widehat{\varepsilon}_n > 0$ is the smallest positive solution to the inequality $\widehat{R}_K(\varepsilon) \leq \varepsilon^2/\sigma_0$, where $\widehat{\varepsilon}_n^2$ is the also the fixed point of $\sigma_0 \widehat{R}_K(\varepsilon)$ as a function of ε^2: $\sigma_0 \widehat{R}_K(\widehat{\varepsilon}_n) = \widehat{\varepsilon}_n^2$. Similarly, the critical population rate ε_n is defined to be the smallest positive solution to the inequality $R_K(\varepsilon) \leq \varepsilon^2/\sigma_0$, where ε_n^2 is the fixed point of $\sigma_0 \widehat{R}_K(\varepsilon)$ as a function of ε^2: $\sigma_0 R_K(\varepsilon_n) = \varepsilon_n^2$. In this paper we consider the case that $n\varepsilon_n^2 \to \infty$ as $n \to \infty$, which is also used in standard analysis of nonparametric regression with minimax rates by kernel regression (Raskutti et al., 2014). We also define $\eta_t \coloneqq \eta t$ for all $t \geq 0$.

5.2. Detailed Roadmap and Key Results

We present the roadmap of our theoretical results which lead to the main result, Theorem 2, in this section. Before presenting the key technical results, we note the by performing learnable channel selection algorithm described in Algorithm 1, Theorem 1 guarantees that $\widehat{\ell} = \ell_0$ w.h.p. Therefore, the condition on ℓ is satisfied in all the results of this section and Theorem 2. Moreover, all the technical results in this section are for the second training stage, that is, training the second-layer weights $\boldsymbol{a}$ by the standard GD. Our main result, Theorem 2, is built upon the following three significant technical results of independent interest.

First, we can have the following principled decomposition of the neural network function at any step of GD into a function in the RKHS associated with the NTK (5), which is $\mathcal{H}_K(B_h)$, and an error function with a small L^∞-norm.

Theorem 3 *Suppose $\widehat{\ell} = \Theta(1) \geq \ell_0$, the network width m is sufficiently large and finite, and the neural network $f_t = f(\boldsymbol{a}(t), \cdot)$ is trained by GD with constant learning rate $\eta = \Theta(1) \in (0, 1/\widehat{\ell})$. Then for every $t \in [T]$, w.h.p., f_t has the following decomposition on $\mathcal{X}$: $f_t = h_t + e_t$, where $h_t \in \mathcal{H}_K(B_h)$ with B_h defined in (37) of the appendix, $e_t \in L^\infty$ with sufficient small $\|e_t\|_\infty$.*

The proof of Theorem 3 relies on the uniform convergence of the empirical kernel $\widehat{K}$ to the corresponding population kernel K, established by the following theorem, which is proved by the concentration inequality for independent random variables taking values in the RKHS associated with the PD activation function σ, $\mathcal{H}_\sigma$.

Theorem 4 *Suppose $\widehat{\ell} = \Theta(1)$. For any fixed $\mathbf{x}' \in \mathcal{X}$ and every $\delta \in (0,1)$, with probability at least $1-\delta$ over the random initialization $\mathbf{Q} = \left\{\vec{\mathbf{q}}_r\right\}_{r=1}^m$, we have*

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{K}(\mathbf{x},\mathbf{x}') - K(\mathbf{x},\mathbf{x}')\right| \lesssim d^{\widehat{\ell}}\sqrt{\frac{\log 2/\delta}{m}}.$$

Theorem 4 is proved as Theorem 25 in the appendix. Theorem 3 shows that, w.h.p., the neural network function $f(\boldsymbol{a}(t),\cdot)$ right after the t-th step of GD can be decomposed into two functions by $f(\boldsymbol{a}(t),\cdot) = f_t = h + e$, where $h \in \mathcal{H}_K(B_h)$ is a function in the RKHS associated with K with a bounded $\mathcal{H}_K$-norm. The error function e has a small L^∞-norm, that is, $\|e\|_\infty \le w$ with w being a small number controlled by the network width m, and larger m leads to smaller w.

Second, local Rademacher complexity is employed to tightly bound the risk of nonparametric regression in Theorem 5 below, which is based on the Rademacher complexity of a localized subset of the function class $\mathcal{F}(B_h, w)$ in Lemma 21 in the appendix. We use Theorem 3 and Lemma 21 to derive Theorem 5.

Theorem 5 *Suppose $\widehat{\ell} = \Theta(1) \ge \ell_0$, the network width m is sufficiently large and finite, and the neural network $f_t = f(\boldsymbol{a}(t),\cdot)$ is trained by GD with constant learning rate $\eta > 0$. Then for every $t \in [T]$, w.h.p.,*

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim \frac{d^{\widehat{\ell}}}{n} + w. \tag{17}$$

It is remarked that the regression risk $\mathbb{E}_P\left[(f_t - f^*)^2\right]$ is bounded by the sum of the training loss and a small term $d^{\widehat{\ell}}/n + w$ through Theorem 5. w is an arbitrarily small positive number with sufficiently large network width m. The sharp rate d^{ℓ_0}/n on the regression risk bound (17) in Theorem 5 is due to the finite rank $m_{\widehat{\ell}} = \Theta(d^{\widehat{\ell}})$ of the kernel K with $\widehat{\ell} = \Theta(1)$.

Third, we have the following sharp upper bound for the training loss $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$.

Theorem 6 *Suppose $\widehat{\ell} = \Theta(1) \ge \ell_0$, the neural network trained after the t-th step of GD, $f_t = f(\boldsymbol{a}(t),\cdot)$, satisfies $\mathbf{u}(t) = f_t(\mathbf{S}) - \mathbf{y} = \mathbf{v}(t) + \mathbf{e}(t)$ with $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$. If $\eta \in (0, 1/\widehat{\ell})$ and τ is suitably small, then for every $t \in [T]$, w.h.p., we have*

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \le \Theta\left(\frac{\gamma_0^2}{\eta t}\right). \tag{18}$$

We then obtain Theorem 2 using the upper bound (17) for the regression risk in Theorem 5 where w is set to $d^{\widehat{\ell}}/n$, with the empirical loss $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$ bounded by $\Theta(d^{\widehat{\ell}}/n)$ w.h.p. by (18) in Theorem 6, and $\widehat{\ell} = \ell_0$ w.h.p.

Detailed proofs of all the technical results of this paper are deferred to the appendix. In particular, Theorem 17, Theorem 25, Theorem 18, and Theorem 19 in the appendix are the formal versions of Theorem 3, Theorem 4, Theorem 5, and Theorem 6 in this section. The proof of Theorem 2 is presented in Section C.1 of the appendix.

5.3. Novel Proof Strategies

We remark that the proof strategies of our main result, Theorem 2, summarized above are significantly different from the existing works in training over-parameterized neural networks for nonparametric regression with minimax rates (Hu et al., 2021; Suh et al., 2022; Li et al., 2024) and existing works about learning low-degree polynomials (Ghorbani et al., 2021; Bai and Lee, 2020; Nichani et al., 2022; Damian et al., 2022).

First, a novel learnable channel selection algorithm is used to select the informative channels in the activation function of the first-layer of the network (1), and the selected channel number ℓ is the ground truth channel number ℓ_0 in the target function w.h.p. Such channel selection ensures that the kernel K is in fact the low-rank kernel $K^{(r_0)}$, ensuring the sharp regression risk bound for the second training stage.

Second, GD is carefully incorporated into the analysis about the uniform convergence results for NTK (5) in Theorem 4, leading to the crucial decomposition of the neural network function f_t in Theorem 3. It is remarked that while existing works such as (Li et al., 2024) also has uniform convergence results for over-parameterized neural network, our results about the uniform convergence for the NTK, rooted in the martingale based concentration inequality for Banach space-valued process (Pinelis, 1992), do not depend on the Hölder continuity of the NTK.

Third, to the best of our knowledge, Theorem 5 is the first result about the sharp upper bound of the order $\Theta(d^{\widehat{\ell}}/n)$ with $w = d^{\widehat{\ell}}/n$ for the regression risk of the neural network function which has the decomposition in Theorem 3. We note that the regression risk in Theorem 5 is $\Theta(d^{\widehat{\ell}}/n) = \Theta(d^{\ell_0}/n)$ w.h.p., which has the expected and the desired order since the target function is in a r_0-dimensional subspace of the RKHS $\mathcal{H}_K(\gamma_0)$ with $r_0 = \Theta(d^{\ell_0})$. Moreover, the proof of Theorem 3, Theorem 5, and Theorem 6 employ the kernel complexity introduced in Section 5.1. In fact, the term $\Theta(d^{\widehat{\ell}}/n)$ corresponds to the fixed point of the kernel complexity R_K.

6. Conclusion

We study nonparametric regression by training an over-parameterized two-layer neural network with channel attention where the target function is in the RKHS associated with the NTK of the neural network and also a degree-ℓ_0 spherical polynomial on the unit sphere in $\mathbb{R}^d$. We show that, through the feature learning capability of the network by a novel learnable channel selection algorithm, the neural network with channel attention trained by the vanilla Gradient Descent (GD) renders a sharp and minimax optimal regression risk bound of $\Theta(d^{\ell_0}/n)$. Novel proof strategies are employed to achieve this result, and we compare our results to the current state-of-the-art with a detailed roadmap of our technical approaches and results.

7. Acknowledgement

This work is supported by the 2023 Mayo Clinic and Arizona State University Alliance for Health Care Collaborative Research Seed Grant Program under Award No. AWD00038846 and by the NIH grant under Award No. 1OT2OD037955-01.

References

Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 4782–4887. PMLR, 2022.

Shuchin Aeron, Venkatesh Saligrama, and Manqi Zhao. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, 2010. doi: 10.1109/TIT.2010.2059891.

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Xcit: Cross-covariance image transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20014–20027, 2021.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 2019.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 2019.

Francis R. Bach. Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.*, 18:19:1–19:53, 2017.

Yu Bai and Jason D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*. OpenReview.net, 2020.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 12873–12884, 2019.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 10835–10845, 2019.

Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In Zhi-Hua Zhou, editor, *International Joint Conference on Artificial Intelligence*, pages 2205–2211. ijcai.org, 2021.

Ziheng Chen, Yue Song, Xiaojun Wu, Gaowen Liu, and Nicu Sebe. Understanding matrix function normalizations in covariance pooling through the lens of riemannian geometry. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

T.S. Chihara. *An Introduction to Orthogonal Polynomials*. Dover Books on Mathematics. Dover Publications, 2011. ISBN 9780486479293.

Moulik Choraria, Leello Tadesse Dadi, Grigorios Chrysos, Julien Mairal, and Volkan Cevher. The spectral bias of polynomial neural networks. In *International Conference on Learning Representations*. OpenReview.net, 2022.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452. PMLR, 2022.

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 2019a.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b.

Costas Efthimiou and Christopher Frye. *Spherical Harmonics in p Dimensions*. World Scientific Co., 2014. doi: 10.1142/9134.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3146–3154. Computer Vision Foundation / IEEE, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *Ann. Statist.*, 49(2):1029 – 1054, 2021.

Margalit Glasgow. SGD finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the XOR problem. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4376–4386. PMLR, 2020.

Tianyang Hu, Wenjia Wang, Cong Lin, and Guang Cheng. Regularization matters: A nonparametric perspective on overparametrized neural network. In Arindam Banerjee and Kenji Fukumizu, editors, *International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 829–837. PMLR, 2021.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 8580–8589, 2018.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 12 2006.

Nicolai V. Krylov. Basics of harmonic polynomials and spherical functions. Technical report. URL https://www-users.cse.umn.edu/~nkrylov/Moscow_2019_Sphrcal.pdf.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

Michel. Ledoux. *Probability in Banach Spaces [electronic resource] : Isoperimetry and Processes / by Michel Ledoux, Michel Talagrand.* Classics in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 1991. edition, 1991.

Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Yicheng Li, Zixiong Yu, Guhan Chen, and Qian Lin. On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains. *Journal of Machine Learning Research*, 25(82):1–47, 2024.

Shahar Mendelson. Geometric parameters of kernel machines. In Jyrki Kivinen and Robert H. Sloan, editors, *Conference on Computational Learning Theory*, volume 2375 of *Lecture Notes in Computer Science*, pages 29–43. Springer, 2002.

Eshaan Nichani, Yu Bai, and Jason D. Lee. Identifying good directions to escape the NTK regime and efficiently learn low-degree plus sparse polynomials. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information*

Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A. Smith, and Lingpeng Kong. Random feature attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Iosif Pinelis. *An Approach to Inequalities for the Distributions of Infinite-Dimensional Martingales*, pages 128–134. Birkhäuser Boston, Boston, MA, 1992. ISBN 978-1-4612-0367-4. doi: 10.1007/978-1-4612-0367-4_9.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 09–15 Jun 2019.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15(1):335–366, 2014.

Yue Song, Nicu Sebe, and Wei Wang. Why approximate matrix square root outperforms accurate SVD in global covariance pooling? In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1095–1103. IEEE, 2021.

Charles J. Stone. Additive Regression and Other Nonparametric Models. *Ann. Statist.*, 13(2):689 – 705, 1985.

Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2019.

Namjoon Suh, Hyunouk Ko, and Xiaoming Huo. A non-parametric regression viewpoint : Generalization of overparametrized deep RELU network under noisy observations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

G. Szegő. *Orthogonal Polynomials*. American Math. Soc: Colloquium publ. Amer. Math. Soc., 1975. ISBN 9780821810231.

Shokichi Takakura and Taiji Suzuki. Mean-field analysis on two-layer neural networks from a kernel perspective. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11531–11539. Computer Vision Foundation / IEEE, 2020.

Qilong Wang, Zhaolin Zhang, Mingze Gao, Jiangtao Xie, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Towards a deeper understanding of global covariance pooling in deep learning: An optimization perspective. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15802–15819, 2023.

Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 9709–9721, 2019.

F. T. Wright. A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables Whose Distributions are not Necessarily Symmetric. *Ann. Probab.*, 1(6):1068 – 1070, 1973.

Greg Yang and Edward J. Hu. Tensor programs IV: feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 2021.

Yingzhen Yang. Sharp generalization for nonparametric regression by over-parameterized neural networks: A distribution-free analysis in spherical covariate. In *International Conference on Machine Learning (ICML)*, 2025.

Yingzhen Yang and Ping Li. Gradient descent finds over-parameterized neural networks with sharp generalization for nonparametric regression. *arXiv preprint arXiv:2411.02904*, 2024. URL https://arxiv.org/abs/2411.02904.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564 – 1599, 1999.

Ming Yuan and Ding-Xuan Zhou. Minimax optimal rates of estimation in high dimensional additive models. *Ann. Statist.*, 44(6):2564 – 2593, 2016.

Lin Zheng, Jianbo Yuan, Chong Wang, and Lingpeng Kong. Efficient attention via control variates. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 2053–2062, 2019.

The appendix of this paper is organized as follows. We present the basic mathematical results employed in our proofs in Section A, and then introduce the detailed technical background about harmonic analysis on spheres in Section B. Detailed proofs are presented in Section C.

Appendix A. Mathematical Tools

The Rademacher complexity of a function class and its empirical version are defined below.

Definition 7 *Let $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^n$ be n i.i.d. random variables such that* $\Pr[\sigma_i = 1] = \Pr[\sigma_i = -1] = \frac{1}{2}$. *The Rademacher complexity of a function class $\mathcal{F}$ is defined as*

$$\mathfrak{R}(\mathcal{F}) = \mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n, \{\sigma_i\}_{i=1}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i) \right]. \tag{19}$$

The empirical Rademacher complexity is defined as

$$\widehat{\mathfrak{R}}(\mathcal{F}) = \mathbb{E}_{\{\sigma_i\}_{i=1}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i) \right], \tag{20}$$

For simplicity of notations, Rademacher complexity and empirical Rademacher complexity are also denoted by $\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i)\right]$ *and* $\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i)\right]$, *respectively.*

For data $\left\{\vec{\mathbf{x}}\right\}_{i=1}^n$ and a function class $\mathcal{F}$, we define the notation $R_n\mathcal{F}$ by $R_n\mathcal{F} \coloneqq \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i)$.

Theorem 8 ((Bartlett et al., 2005, Theorem 2.1)) *Let $\mathcal{X}, P$ be a probability space, $\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n$ be independent random variables distributed according to P. Let $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ into $[a,b]$. Assume that there is some $r > 0$ such that for every $f \in \mathcal{F}$,*$\mathrm{Var}\left[f(\vec{\mathbf{x}}_i)\right] \le r$. *Then, for every $x > 0$, with probability at least $1 - e^{-x}$,*

$$\sup_{f\in\mathcal{F}} \left(\mathbb{E}_P[f(\mathbf{x})] - \mathbb{E}_{P_n}[f(\mathbf{x})]\right) \le \inf_{\alpha>0} \left(2(1+\alpha)\mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n, \{\sigma_i\}_{i=1}^n}[R_n\mathcal{F}] + \sqrt{\frac{2rx}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{x}{n} \right), \tag{21}$$

and with probability at least $1 - 2e^{-x}$,

$$\sup_{f\in\mathcal{F}} \left(\mathbb{E}_P[f(\mathbf{x})] - \mathbb{E}_{P_n}[f(\mathbf{x})]\right) \le \inf_{\alpha\in(0,1)} \left(\tfrac{2(1+\alpha)}{1-\alpha}\mathbb{E}_{\{\sigma_i\}_{i=1}^n}[R_n\mathcal{F}] + \sqrt{\tfrac{2rx}{n}} + (b-a)\left(\tfrac{1}{3} + \tfrac{1}{\alpha} + \tfrac{1+\alpha}{2\alpha(1-\alpha)}\right)\tfrac{x}{n} \right). \tag{22}$$

P_n is the empirical distribution over $\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n$ with $\mathbb{E}_{P_n}[f(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^n f(\vec{\mathbf{x}}_i)$. Moreover, the same results hold for $\sup_{f\in\mathcal{F}} \left(\mathbb{E}_{P_n}[f(\mathbf{x})] - \mathbb{E}_P[f(\mathbf{x})]\right)$.

In addition, we have the contraction property for Rademacher complexity, which is due to Ledoux and Talagrand (Ledoux, 1991).

Theorem 9 *Let ϕ be a contraction,that is, $|\phi(x) - \phi(y)| \le \mu |x - y|$ for $\mu > 0$. Then, for every function class $\mathcal{F}$,*

$$\mathbb{E}_{\{\sigma_i\}_{i=1}^n}\left[R_n\phi\circ\mathcal{F}\right] \le \mu\mathbb{E}_{\{\sigma_i\}_{i=1}^n}\left[R_n\mathcal{F}\right], \tag{23}$$

where $\phi\circ\mathcal{F}$ is the function class defined by $\phi\circ\mathcal{F} = \{\phi\circ f\colon f\in\mathcal{F}\}$.

Definition 10 (Sub-root function,(Bartlett et al., 2005, Definition 3.1)) *A function $\psi\colon [0,\infty) \to [0,\infty)$ is sub-root if it is nonnegative, nondecreasing and if $\frac{\psi(r)}{\sqrt{r}}$ is nonincreasing for $r > 0$.*

Theorem 11 ((Bartlett et al., 2005, Theorem 3.3)) *Let $\mathcal{F}$ be a class of functions with ranges in $[a,b]$ and assume that there are some functional $T\colon \mathcal{F} \to \mathbb{R}+$ and some constant $\bar{B}$ such that for every $f \in \mathcal{F}$, $\mathrm{Var}[f] \le T(f) \le \bar{B}P(f)$. Let ψ be a sub-root function and let r^* be the fixed point of ψ. Assume that ψ satisfies that, for any $r \ge r^*$, $\psi(r) \ge \bar{B}\mathfrak{R}(\{f \in \mathcal{F}\colon T(f) \le r\})$. Fix $x > 0$, then for any $K_0 > 1$, with probability at least $1 - e^{-x}$,*

$$\forall f \in \mathcal{F}, \quad \mathbb{E}_P[f] \le \frac{K_0}{K_0 - 1}\mathbb{E}_{P_n}[f] + \frac{704K_0}{\bar{B}}r^* + \frac{x\left(11(b-a) + 26\bar{B}K_0\right)}{n}.$$

Also, with probability at least $1 - e^{-x}$,

$$\forall f \in \mathcal{F}, \quad \mathbb{E}_{P_n}[f] \le \frac{K_0 + 1}{K_0}\mathbb{E}_P[f] + \frac{704K_0}{\bar{B}}r^* + \frac{x\left(11(b-a) + 26\bar{B}K_0\right)}{n}.$$

Lemma 12 ((Bartlett et al., 2005, Lemma 3.4)) *If a function class $\mathcal{F}$ is star-shaped around a function $\widehat{f}$, and $T\colon \mathcal{F} \to \mathbb{R}^+$ with $\mathbb{R}^+$ being the set of all nonnegative real numbers is a (possibly random) function that satisfies $T(\alpha f) \le \alpha^2 T(f)$ for every $f \in \mathcal{F}$ and any $\alpha \in [0,1]$, then the (random) function ψ defined for $r \ge 0$ by $\psi(r) \coloneqq \mathbb{E}_{\{\sigma_i\}_{i=1}^n}\left[R_n\left\{f - \widehat{f}\colon f \in \mathcal{F}, T(f - \widehat{f}) \le r\right\}\right]$ is sub-root and $r \to \mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n}[\psi(r)]$ is also sub-root.*

Appendix B. Detailed Technical Background about Harmonic Analysis on Spheres

In this section, we provide background materials on spherical harmonic analysis needed for our study of the RKHS. We refer the reader to (Chihara, 2011; Efthimiou and Frye, 2014; Szegő, 1975) for further information on these topics. As mentioned above, expansions in spherical harmonics were used in the past in the statistics literature, such as (Bach, 2017; Bietti and Mairal, 2019).

With $\ell \ge 0$, let $\mathcal{P}_\ell^{(\mathrm{hom})}$ denote the space of all the degree-ℓ homogeneous polynomials on $\mathcal{X} = \mathbb{S}^{d-1}$, and let $\mathcal{H}_\ell$ denote the space of degree-ℓ homogeneous harmonic polynomials on $\mathcal{X}$, or the degree-ℓ spherical harmonics. That is,

$$\mathcal{H}_\ell = \left\{P\colon \mathcal{X} \to \mathbb{R}\colon P(\mathbf{x}) = \sum_{|\alpha|=\ell} c_\alpha \mathbf{x}^\alpha, \Delta P = 0\right\}, \tag{24}$$

where $\alpha = [\alpha_1, \ldots, \alpha_d]$, $\mathbf{x}^\alpha = \prod_{i=1}^d \mathbf{x}_i^{\alpha_i}$, $|\alpha| = \sum_{i=1}^d \alpha_i$, and Δ is the Laplacian operator. For $\ell \neq \ell'$, the elements of $\mathcal{H}_\ell$ and $\mathcal{H}_{\ell'}$ are orthogonal to each other. All the functions in the following text of this section are assumed to be elements of $L^2(\mathcal{X}, v_{d-1})$, where v_{d-1} stands for the uniform distribution on the sphere $\mathcal{X} = \mathbb{S}^{d-1}$. We have $\langle f, g\rangle_{L^2} \coloneqq \int_{\mathcal{X}} f(x)g(x)\mathrm{d}v_{d-1}(x)$. We denote by $\{Y_{kj}\}_{j\in[N(d,k)]}$ the spherical harmonics of degree k which form an orthogonal basis of $\mathcal{H}_k$, where $N(d,k) = \frac{2k+d-2}{k}\binom{k+d-3}{d-2}$ is the dimension of $\mathcal{H}_k$. They form an orthonormal basis of $L^2(\mathcal{X}, v_{d-1})$. We have $\sum_{j=1}^{N(d,k)} Y_{kj}(\mathbf{x})Y_{kj}(\mathbf{x}') = N(d,k)P_k^{(d)}(\langle \mathbf{x}, \mathbf{x}'\rangle)$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$,

where $P_k^{(d)}$ is the k-th Legendre polynomial in dimension d, which is also known as Gegenbauer polynomials, given by the Rodrigues formula:

$$P_k^{(d)}(t) = (-\frac{1}{2})^k \frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(k+\frac{d-1}{2}\right)} \left(1-t^2\right)^{(3-d)/2} \left(\frac{\mathrm{d}}{\mathrm{d}t}\right)^k \left(1-t^2\right)^{k+(d-3)/2}. \tag{25}$$

The polynomials $\left\{P_k^{(d)}\right\}$ are orthogonal in $L^2(\mathcal{X}, \mathrm{d}v_{d-1})$ where the measure $\mathrm{d}v_{d-1}$ is given by $\mathrm{d}v_{d-1}(t) = (1-t^2)^{(d-3)/2}\mathrm{d}t$, and we have

$$\int_{-1}^{1} {P_k^{(d)}}^2(t)(1-t^2)^{(d-3)/2}\mathrm{d}t = \frac{w_{d-1}}{w_{d-2}} \frac{1}{N(d,k)},$$

where $w_{d-1} \coloneqq \frac{2\pi^{d/2}}{\Gamma(d/2)}$ denotes the surface of the unit sphere $\mathbb{S}^{d-1}$. It follows from the orthogonality of spherical harmonics that

$$\int_{\mathcal{X}} P_j^{(d)}(\langle \mathbf{x}, \mathbf{w}\rangle) P_j^{(d)}(\langle \mathbf{x}', \mathbf{w}\rangle)\mathrm{d}v_{d-1}(\mathbf{w}) = \frac{\delta_{jk}}{N(d,k)} P_k^{(d)}(\langle \mathbf{x}, \mathbf{x}'\rangle),$$

where $\delta_{jk} = \mathbb{1}_{\{j=k\}}$. We have the following recurrence relation (Efthimiou and Frye, 2014, Equation 4.36),

$$tP_k^{(d)}(t) = \frac{k}{2k+d-2} P_{k-1}^{(d)}(t) + \frac{k+d-2}{2k+d-2} P_{k+1}^{(d)}(t) \tag{26}$$

for all $k \geq 1$, and $tP_0^{(d)}(t) = P_1^{(d)}(t)$, and $P_0^{(d)} \equiv 1$. It follows that $P_k^{(d)}(1) = 1$ for all $k \geq 0$, and it can be verified that $\left|P_k^{(d)}(t)\right| \leq 1$ for all $k \geq 0$ and $t \in [-1, 1]$.

The Funk-Hecke formula is helpful for computing Fourier coefficients in the basis of spherical harmonics in terms of Legendre polynomials. For any $j \in [N(d,k)]$, we have

$$\int_{\mathcal{X}} f(\langle \mathbf{x}, \mathbf{x}'\rangle) Y_{kj}(\mathbf{x}')\mathrm{d}v_{d-1}(\mathbf{x}') = \frac{w_{d-2}}{w_{d-1}} Y_{kj}(\mathbf{x}) \int_{-1}^{1} f(t) P_k^{(d)}(t)(1-t^2)^{(d-3)/2}\mathrm{d}t.$$

For a positive-definite kernel $\tilde{K}(\mathbf{x}, \mathbf{x}') = \kappa(\langle \mathbf{x}, \mathbf{x}'\rangle)$ defined on $\mathcal{X}$, we have its Mercer decomposition as follows.

$$\tilde{K}(\mathbf{x}, \mathbf{x}') = \sum_{\ell \geq 0} \mu_\ell \sum_{j=1}^{N(d,\ell)} Y_{\ell,j}(\mathbf{x}) Y_{\ell,j}(\mathbf{x}') = \sum_{\ell \geq 0} \mu_\ell N(d,\ell) P_\ell^{(d)}(\langle \mathbf{x}, \mathbf{x}'\rangle),$$

where μ_ℓ is the eigenvalue of the integral operator $T_{\tilde{K}}$ associated with $\tilde{K}$ corresponding to $\mathcal{H}_\ell$. It follows that

$$\mu_\ell = \frac{w_{d-2}}{w_{d-1}} \int_{-1}^{1} \kappa(t) P_\ell^{(d)}(t)(1-t^2)^{(d-3)/2}\mathrm{d}t.$$

Proposition 13 ((Krylov, Theorem 4.2)) *Let $p \in \mathcal{P}_\ell^{(\mathrm{hom})}$. Then there exists unique $h_{n-2i} \in \mathcal{H}_{n-2i}$ for $i \in \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ such that*

$$p(\mathbf{x}) = h_n + h_{n-2} + \ldots + h_{n-2k}.$$

Theorem 14 *Every polynomial p defined on $\mathbb{S}^{d-1}$ of degree k for $k \geq 0$ can be represented as a linear combination of homogeneous harmonic polynomials up to degree k, that is,*

$$p = \sum_{i=0}^{k} c_i p_i,$$

where $p_i \in \mathcal{H}_i$ for $i \in \{0, 1, \ldots, k\}$.

Proof Every polynomial p defined on $\mathbb{S}^{d-1}$ of degree k can be represented as the sum of homogeneous polynomials on $\mathbb{S}^{d-1}$ by grouping the terms of p of the same degree together. It follows from Proposition 13 that every homogeneous polynomial is a linear combination of homogeneous harmonic polynomials up to degree k. As a result, the conclusion holds. ■

Lemma 15 *For $\ell_0 = \Theta(1)$ and $d > \Theta(1)$, we have*

$$r_0 = \Theta(d^{\ell_0}). \tag{27}$$

Proof It follows from the direct calculation that $N(d, \ell) \asymp d^\ell$ under the given conditions, so that $r_0 = \sum_{\ell=0}^{\ell_0} N(d, \ell) \asymp d^{\ell_0}$. ■

Lemma 16 (Efficient Computation of the Activation Function σ Defined in (2)) *For every given $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and the channel attention weights $\boldsymbol{\tau}$, $\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}')$ can be computed in $\Theta(1)$ time.*

Proof We note that $\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}')$ is computed by $\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}') = \sum_{\ell=0}^{L} \tau_\ell P_\ell^{(d)}(t)$ with $t = \langle \mathbf{x}, \mathbf{x}' \rangle$. Using the recursive formula (26) and standard dynamic programming, $\left\{ P_\ell^{(d)}(t) \right\}_{\ell=0}^{L}$ can be computed in $\Theta(L)$ time. To see this, we note that $P_0^{(d)}(t) = 1$, and the computation of $P_{\ell'}^{(d)}(t)$ for every $\ell' \in [1 : L]$ takes $\Theta(1)$ time by (26) using the stored values of $\left\{ P_\ell^{(d)}(t) \right\}_{\ell=0}^{\ell'-1}$. Summing all the $\tau_\ell P_\ell^{(d)}(t)$ takes $\Theta(L)$, so the computation of $\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}')$ takes $\Theta(L)$ time in total. ■

Appendix C. Detailed Proofs

We present detailed proofs for the theoretical results that lead to our main result, Theorem 2, in this section. The proof of Theorem 2 is presented in Section C.1, followed by the basic definitions and the detailed proofs of our other technical results. We first introduce the definition of stopping time which serves as the upper bound for the number of steps T in Algorithm 2.

Definition of Stopping Time. Recall that $\eta_t = \eta t$ for all $t > 0$, we then define the stopping time $\widehat{T}$ as

$$\widehat{T} := \min \left\{ t \colon \widehat{R}_K(\sqrt{1/\eta_t}) > (\sigma_0 \eta_t)^{-1} \right\} - 1. \tag{28}$$

The stopping time in fact is the upper bound for the number of steps T for Algorithm 2, that is, $T \leq \widehat{T}$. In the proof of Theorem 2, we will show that $\widehat{T} \asymp n/d^{\widehat{\ell}}$, so that it is always feasible to choose $T \leq \widehat{T}$ such that $T \asymp n/d^{\widehat{\ell}}$. Throughout this appendix we let $T \leq \widehat{T}$.

C.1. Proof of Theorem 2

Proof of Theorem 2 We use Theorem 18 and Theorem 19 in this appendix to prove this theorem. Theorem 18 and Theorem 19 are the formal versions of Theorem 5 and Theorem 6, respectively.

First of all, it follows by Theorem 19 that with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over $\mathbf{w}$,

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \leq \Theta\left(\frac{1}{\eta t}\right).$$

Plugging such bound for $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$ in (40) of Theorem 18 leads to

$$\mathbb{E}_{P}\left[(f_t - f^*)^2\right] \lesssim \Theta\left(\frac{1}{\eta t}\right) + \frac{d^{\widehat{\ell}}}{n} + w. \tag{29}$$

By the definition of $\widehat{T}$ and $\widehat{\varepsilon}_n^2$, we have

$$\widehat{\varepsilon}_n^2 \leq \frac{1}{\eta \widehat{T}} \leq \frac{2}{\eta(\widehat{T}+1)} \leq 2\widehat{\varepsilon}_n^2,$$

so that $\widehat{T} \asymp \widehat{\varepsilon}_n^{-2}$. Furthermore, it follows from (Raskutti et al., 2014, Corollary 4) that $\varepsilon_n^2 \asymp r_0/n$. In addition, Lemma 31 suggests that with probability $1 - 4\exp(-\Theta(n\varepsilon_n^2)) = 1 - 4\exp(-\Theta(r_0))$, $\widehat{\varepsilon}_n^2 \asymp \varepsilon_n^2$. As a result, $\widehat{T} \asymp n/d^{\widehat{\ell}}$, and we choose $T \leq \widehat{T}$ such that $T \asymp n/d^{\widehat{\ell}}$.

Due to the setting that $T \asymp n/d^{\widehat{\ell}}$ and $\eta = \Theta(1)$, we have

$$\frac{1}{\eta t} \asymp \frac{1}{\eta T} \asymp \frac{d^{\widehat{\ell}}}{n}. \tag{30}$$

Let $w = d^{\widehat{\ell}}/n$, then $w \in (0,1)$ with $n > d^{\widehat{\ell}}$. (15) then follows from (29) with $w = d^{\widehat{\ell}}/n$, (30) and the union bound. We note that $c_{\mathbf{u}}$ is bounded by a positive constant, so that the condition on m in (38) in Theorem 17, together with $w = d^{\widehat{\ell}}/n$ and (30) leads to the condition on m in (14).

This theorem is then proved by noting that Theorem 1 guarantees that $\widehat{\ell} = \ell_0$ holds with probability at least $1 - \exp\left(-\Theta(m_L)\right) - \delta$, where $\widehat{\ell}$ is the number of channels selected by the learnable channel selection algorithm described in Algorithm 1.

■

C.2. Basic Definitions

We introduce the following definitions for our analysis. We define

$$\mathbf{u}(t) \coloneqq \widehat{\mathbf{y}}(t) - \mathbf{y} \tag{31}$$

as the difference between the network output $\widehat{\mathbf{y}}(t)$ and the training response vector $\mathbf{y}$ right after the t-th step of GD. Let $\tau \leq 1$ be a positive number. For $t \geq 0$ and $T \geq 1$ we define the following quantities: $c_{\mathbf{u}} \coloneqq \Theta(\gamma_0) + \sigma_0 + \tau + 1$,

$$\mathcal{V}_t \coloneqq \left\{\mathbf{v} \in \mathbb{R}^n \colon \mathbf{v} = -\left(\mathbf{I}_n - \eta \mathbf{K}_n\right)^t f^*(\mathbf{S})\right\}, \tag{32}$$

$$\mathcal{E}_{t,\tau} \coloneqq \left\{ \mathbf{e} \colon \mathbf{e} = \vec{\mathbf{e}}_1 + \vec{\mathbf{e}}_2 \in \mathbb{R}^n, \vec{\mathbf{e}}_1 = -\left(\mathbf{I}_n - \eta \mathbf{K}_n\right)^t \mathbf{w}, \left\| \vec{\mathbf{e}}_2 \right\|_2 \leq \sqrt{n}\tau \right\}. \tag{33}$$

In particular, Theorem 20 in the appendix shows that w.h.p. over the random noise $\mathbf{w}$ and the random initialization $\mathbf{Q}$, $\mathbf{u}(t)$ can be composed into two vectors, $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ such that $\mathbf{v}(t) \in \mathcal{V}_t$ and $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$. We then define the set of the neural network weights during the training by GD as follows:

$$\mathcal{A}(\mathbf{S}, \boldsymbol{a}, T) \coloneqq \Bigg\{ \boldsymbol{a} \colon \exists t \in [T] \text{ s.t. } \boldsymbol{a} = -\sum_{t'=0}^{t-1} \frac{\eta}{n} \mathbf{Z}(t') \mathbf{u}(t'),$$
$$\mathbf{u}(t') \in \mathbb{R}^n, \mathbf{u}(t') = \mathbf{v}(t') + \mathbf{e}(t'), \mathbf{v}(t') \in \mathcal{V}_{t'}, \mathbf{e}(t') \in \mathcal{E}_{t',\tau}, \text{ for all } t' \in [0, t-1] \Bigg\}. \tag{34}$$

The set of the functions represented by the neural network with weights in $\mathcal{A}(\mathbf{S}, \boldsymbol{a}, T)$ is then defined as

$$\mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \boldsymbol{a}, T) \coloneqq \left\{ f_t = f(\boldsymbol{a}(t), \cdot) \colon \exists\, t \in [T], \boldsymbol{a}(t) \in \mathcal{A}(\mathbf{S}, \boldsymbol{a}, T) \right\}. \tag{35}$$

We also define the function class $\mathcal{F}(B, w)$ for any $B, w > 0$ as

$$\mathcal{F}(B, w) \coloneqq \left\{ f \colon f = h + e, h \in \mathcal{H}_K(B), \|e\|_\infty \leq w \right\}. \tag{36}$$

We will show by Theorem 3 in the next subsection that w.h.p. over $\mathbf{w}$, $\mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \boldsymbol{a}, T)$ is a subset of $\mathcal{F}(B, w)$, where a smaller w requires a larger network width m, and $B_h > \gamma_0$ is an absolute positive constant defined by

$$B_h \coloneqq \gamma_0 + \sqrt{2} + 1. \tag{37}$$

C.3. Proofs for Results in Section 5.2

We present our key technical results regarding optimization and generalization of the two-layer NN (1) trained by GD in this section. The following theorem, Theorem 17, is the formal version of Theorem 3 in Section 5.2, and it states that w.h.p. over $\mathbf{w}$, $\mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \boldsymbol{a}, T) \subseteq \mathcal{F}(B_h, w)$.

Theorem 17 *Suppose $\widehat{\ell} = \Theta(1) \geq \ell_0$. Suppose $w \in (0, 1)$, the network width m satisfies*

$$m \gtrsim \max\left\{ T^2 d^{2\widehat{\ell}} \log(2n/\delta)/w^2, T^4 d^{2\widehat{\ell}} \log(2n/\delta) \right\}, \tag{38}$$

and the neural network $f_t = f(\boldsymbol{a}(t), \cdot)$ is trained by GD with the constant learning rate $\eta \in (0, 1/(\widehat{\ell} + 1))$ and $\eta = \Theta(1)$. Then for every $t \in [T]$ and every $\delta \in (0, 1)$, with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right) - \exp\left(-\Theta(n)\right) - \delta$ over the random initialization $\mathbf{Q}$ and the random noise $\mathbf{w}$, $f_t \in \mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \boldsymbol{a}, T)$, and f_t has the following decomposition on $\mathcal{X}$:

$$f_t = h_t + e_t, \tag{39}$$

where $h_t \in \mathcal{H}_K(B_h)$ with B_h defined in (37), $e_t \in L^\infty$ with $\|e_t\|_\infty \leq w$.

Based on Theorem 17 and the local Rademacher complexity based analysis (Bartlett et al., 2005), Theorem 18 presents a sharp upper bound for the nonparametric regression risk, $\mathbb{E}_P\left[(f_t - f^*)^2\right]$, where f_t is the function represented by the two-layer NN (1) right after the t-th step of GD. Theorem 18 is the formal version of Theorem 5 in Section 5.2.

Theorem 18 *Suppose $\widehat{\ell} = \Theta(1) \geq \ell_0$, $w \in (0,1)$, m satisfies (38), and the neural network $f_t = f(\boldsymbol{a}(t), \cdot)$ is trained by GD with the constant learning rate $\eta \in (0, 1/(\widehat{\ell}+1))$ and $\eta = \Theta(1)$. Then for every $t \in [T]$ and every $\delta \in (0,1)$, with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right) - \exp(-m_{\widehat{\ell}}) - \exp\left(-\Theta(n)\right) - \delta$ over the random noise $\mathbf{w}$, the random training features $\mathbf{S}$, and the random initialization $\mathbf{Q}$,*

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim \frac{d^{\widehat{\ell}}}{n} + w. \tag{40}$$

Theorem 19 below shows that the empirical loss $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$ is bounded by $\Theta(1/(\eta t))$ w.h.p. over $\mathbf{w}$. Theorem 19 is the formal version of Theorem 6 in Section 5.2. Such upper bound for the empirical loss by Theorem 19 will be plugged in the risk bound in Theorem 18 to prove Theorem 2. The proofs of Theorem 2 and its corollary are presented in the next subsection.

Theorem 19 *Suppose $\widehat{\ell} = \Theta(1) \geq \ell_0$, the neural network trained after the t-th step of GD, $f_t = f(\boldsymbol{a}(t), \cdot)$, satisfies $\mathbf{u}(t) = f_t(\mathbf{S}) - \mathbf{y} = \mathbf{v}(t) + \mathbf{e}(t)$ with $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$. If*

$$\eta \in (0, 1/(\widehat{\ell}+1)), \quad \tau \leq \frac{1}{\eta T}, \tag{41}$$

then for every $t \in [T]$, with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$, we have

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \leq \Theta\left(\frac{1}{\eta t}\right). \tag{42}$$

C.3.1. PROOF OF THEOREM 17

We prove Theorem 17 in this subsection. The proof requires the following theorem, Theorem 20, about our main result about the optimization of the network (1). Theorem 20 states that w.h.p. over the random noise $\mathbf{w}$ and the random initialization $\mathbf{Q}$, the weights of the network $\boldsymbol{a}(t)$ obtained right after the t-th step of GD belongs to $\mathcal{A}(\mathbf{S}, \boldsymbol{a}, T)$. The proof of Theorem 20 is based on Lemma 22 and Lemma 23 deferred to Section C.4 of this appendix.

Theorem 20 *Suppose $\widehat{\ell} = \Theta(1) \geq \ell_0$,*

$$m \gtrsim T^2 d^{2\widehat{\ell}} \log(2n/\delta)/\tau^2, \tag{43}$$

and the neural network $f(\boldsymbol{a}(t), \cdot)$ trained by GD with the constant learning rate $\eta = \Theta(1) \in (0, 1/(\widehat{\ell}+1))$. Then with probability at least $1 - \exp\left(-\Theta(n)\right) - \delta$ over the random noise $\mathbf{w}$ and the random initialization $\mathbf{Q}$, $\boldsymbol{a}(t) \in \mathcal{A}(\mathbf{S}, \boldsymbol{a}, T)$ for every $t \in [T]$. Moreover, for every $t \in [0, T]$, $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ where $\mathbf{u}(t) = \widehat{\mathbf{y}}(t) - \mathbf{y}$, $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, $\|\mathbf{u}(t)\|_2 \leq c_{\mathbf{u}}\sqrt{n}$.

Proof of Theorem 20 First, when $m \gtrsim T^2 d^{2\widehat{\ell}} \log(2n/\delta)/\tau^2$ with a proper constant, it can be verified that $\mathbf{E}_{m,n,\eta} \leq \tau\sqrt{n}/T$ where $\mathbf{E}_{m,n,\eta}$ is specified by (76) of Lemma 23. We then use mathematical induction to prove this theorem. We will first prove that $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ where $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, and $\|\mathbf{u}(t)\|_2 \leq c_{\mathbf{u}}\sqrt{n}$ for all $t \in [0, T]$.

When $t = 0$, we have

$$\mathbf{u}(0) = -\mathbf{y} = \mathbf{v}(0) + \mathbf{e}(0), \tag{44}$$

where $\mathbf{v}(0) \coloneqq -f^*(\mathbf{S}) = -\left(\mathbf{I} - \eta\mathbf{K}_n\right)^0 f^*(\mathbf{S})$, $\mathbf{e}(0) = -\mathbf{w}$ with $\mathbf{e}(0) = -\left(\mathbf{I} - \eta\mathbf{K}_n\right)^0 \mathbf{w}$. Therefore, $\mathbf{v}(0) \in \mathcal{V}_0$ and $\mathbf{e}(0) \in \mathcal{E}_{0,\tau}$. Also, it follows from the proof of Lemma 22 that $\|\mathbf{u}(0)\|_2 \leq c_{\mathbf{u}}\sqrt{n}$ with probability at least $1 - \exp\left(-\Theta(n)\right)$ over the random noise $\mathbf{w}$.

Suppose that for all $t_1 \in [0, t]$ with $t \in [0, T-1]$, $\mathbf{u}(t_1) = \mathbf{v}(t_1) + \mathbf{e}(t_1)$ where $\mathbf{v}(t_1) \in \mathcal{V}_{t_1}$, $\mathbf{e}(t_1) \in \mathcal{E}_{t_1,\tau}$, and $\|\mathbf{u}(t_1)\|_2 \leq c_{\mathbf{u}}\sqrt{n}$. Then it follows from Lemma 23 that the recursion $\mathbf{u}(t'+1) = \left(\mathbf{I} - \eta\mathbf{K}_n\right)\mathbf{u}(t') + \mathbf{E}(t'+1)$ holds for all $t' \in [0, t]$. As a result, we have

$$\begin{aligned}
\mathbf{u}(t+1) &= \left(\mathbf{I} - \eta\mathbf{K}_n\right)\mathbf{u}(t) + \mathbf{E}(t+1) \\
&= -\left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1} f^*(\mathbf{S}) - \left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1}\mathbf{w} + \sum_{t'=1}^{t+1}\left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1-t'}\mathbf{E}(t') \\
&= \mathbf{v}(t+1) + \mathbf{e}(t+1),
\end{aligned} \tag{45}$$

where $\mathbf{v}(t+1)$ and $\mathbf{e}(t+1)$ are defined as

$$\mathbf{v}(t+1) \coloneqq -\left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1} f^*(\mathbf{S}) \in \mathcal{V}_{t+1}, \tag{46}$$

$$\mathbf{e}(t+1) \coloneqq \underbrace{-\left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1}\mathbf{w}}_{\vec{\mathbf{e}}_1(t+1)} + \underbrace{\sum_{t'=1}^{t+1}\left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1-t'}\mathbf{E}(t')}_{\vec{\mathbf{e}}_2(t+1)}. \tag{47}$$

We now prove the upper bound for $\vec{\mathbf{e}}_2(t+1)$. With $\eta \in (0, 1/(\widehat{\ell}+1))$, we have $\|\mathbf{I} - \eta\mathbf{K}_n\|_2 \in (0, 1)$. It follows that

$$\left\|\vec{\mathbf{e}}_2(t+1)\right\|_2 \leq \sum_{t'=1}^{t+1}\|\mathbf{I} - \eta\mathbf{K}_n\|_2^{t+1-t'}\left\|\mathbf{E}(t')\right\|_2 \leq \tau\sqrt{n}, \tag{48}$$

where the last inequality follows from the fact that $\|\mathbf{E}(t)\|_2 \leq \mathbf{E}_{m,n,\eta} \leq \tau\sqrt{n}/T$ for all $t \in [T]$. It follows that $\mathbf{e}(t+1) \in \mathcal{E}_{t+1,\tau}$. Also, since $\widehat{\ell} \geq \ell_0$, it follows from Lemma 22 that

$$\begin{aligned}
\|\mathbf{u}(t+1)\|_2 &\leq \|\mathbf{v}(t+1)\|_2 + \left\|\vec{\mathbf{e}}_1(t+1)\right\|_2 + \left\|\vec{\mathbf{e}}_2(t+1)\right\|_2 \\
&\leq \left(\frac{\gamma_0}{\sqrt{2e\eta}} + \sigma_0 + \tau + 1\right)\sqrt{n} \leq c_{\mathbf{u}}\sqrt{n}.
\end{aligned}$$

The above inequality completes the induction step, which also completes the proof. ■

Proof of Theorem 17 In this proof we abbreviate f_t as f. It follows from Theorem 20 and its proof that conditioned on an event Ω with probability at least $1 - \exp\left(-\Theta(n)\right) - \delta$, $f \in \mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \boldsymbol{a}, T)$. Moreover, $f = f(\boldsymbol{a}, \cdot)$ with $\boldsymbol{a} = \{a_r\}_{r=1}^m \in \mathcal{A}(\mathbf{S}, \boldsymbol{a}, T)$, where $\mathbf{u}(t') \in \mathbb{R}^n$, $\mathbf{u}(t') = \mathbf{v}(t') + \mathbf{e}(t')$ with $\mathbf{v}(t') \in \mathcal{V}_{t'}$ and $\mathbf{e}(t') \in \mathcal{E}_{t',\tau}$ for all $t' \in [0, t-1]$. $\boldsymbol{a}$ is expressed as

$$\boldsymbol{a} = \boldsymbol{a}(t) = -\sum_{t'=0}^{t-1}\frac{\eta}{n}\mathbf{Z}(t')\mathbf{u}(t') \tag{49}$$

for some $t \in [T]$. Using (49), $g(\mathbf{x})$ is expressed as

$$f(\mathbf{x}) = f(\boldsymbol{a}, \mathbf{x}) = -\sum_{t'=0}^{t-1} \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \sigma_{\boldsymbol{\tau}}(\mathbf{x}, \vec{\mathbf{q}}_r) \frac{\eta}{n} \left[\mathbf{Z}(t')\right]_r \mathbf{u}(t')$$
$$= -\sum_{t'=0}^{t-1} \underbrace{\frac{\eta}{n} \sum_{j=1}^{n} \widehat{K}(\mathbf{x}, \vec{\mathbf{x}}_j) \left[\mathbf{u}(t')\right]_j}_{:=G_{t'}(\mathbf{x})}, \tag{50}$$

For each $G_{t'}$ in the RHS of (50), we have

$$G_{t'}(\mathbf{x}) = \frac{\eta}{n} \sum_{j=1}^{n} \widehat{K}(\mathbf{x}, \vec{\mathbf{x}}_j) \left[\mathbf{u}(t')\right]_j \overset{①}{=} \frac{\eta}{n} \sum_{j=1}^{n} K(\mathbf{x}, \vec{\mathbf{x}}_j) \left[\mathbf{u}(t')\right]_j + \underbrace{\frac{\eta}{n} \sum_{j=1}^{n} q_j \left[\mathbf{u}(t')\right]_j}_{:=E(\mathbf{x})}. \tag{51}$$

where $q_j := \widehat{K}(\mathbf{x}, \vec{\mathbf{x}}_j) - K(\mathbf{x}, \vec{\mathbf{x}}_j)$ for all $j \in [n]$ in ①. We now analyze each term on the RHS of (51). Let $h(\cdot, t') \colon \mathcal{X} \to \mathbb{R}$ be defined by $h(\mathbf{x}, t') := \frac{\eta}{n} \sum_{j=1}^{n} K(\mathbf{x}, \vec{\mathbf{x}}_j) \left[\mathbf{u}(t')\right]_j$, then $h(\cdot, t') \in \mathcal{H}_K$ for each $t' \in [0, t-1]$. We define

$$h_t(\cdot) := -\sum_{t'=0}^{t-1} h(\cdot, t') \in \mathcal{H}_K, \tag{52}$$

We note that w.h.p., $\mathbf{u}(t') \le c_{\mathbf{u}} \sqrt{n}$. Since $\widehat{\ell} = \Theta(1)$, it follows from (89) in Theorem 25 that $|q_j| \lesssim d^{\widehat{\ell}} \sqrt{\frac{\log(2n/\delta)}{m}}$ for all $j \in [n]$. As a result, we have

$$\|E\|_\infty = \left\| \frac{\eta}{n} \sum_{j=1}^{n} q_j \mathbf{u}_j(t') \right\|_\infty \lesssim \frac{\eta}{n} c_{\mathbf{u}} \sqrt{n} \cdot \sqrt{n} d^{\widehat{\ell}} \sqrt{\frac{\log(2n/\delta)}{m}} \lesssim \eta c_{\mathbf{u}} d^{\widehat{\ell}} \sqrt{\frac{\log(2n/\delta)}{m}}. \tag{53}$$

Combining (51) and (53), any $t' \in [0, t-1]$,

$$\sup_{\mathbf{x} \in \mathcal{X}} \left| G_{t'}(\mathbf{x}) - h(\mathbf{x}, t') \right| \le \|E\|_\infty \lesssim \eta c_{\mathbf{u}} d^{\widehat{\ell}} \sqrt{\frac{\log(2n/\delta)}{m}}. \tag{54}$$

Define $e_t := f(\boldsymbol{a}, \cdot) - h_t$. It then follows from (50) and (54) that

$$\|e_t\|_\infty \le \sup_{\mathbf{x} \in \mathcal{X}} |f(\boldsymbol{a}, \mathbf{x}) - h_t(\mathbf{x})| \le \sum_{t'=0}^{t-1} \sup_{\mathbf{x} \in \mathcal{X}} \left| G_{t'}(\mathbf{x}) - h(\mathbf{x}, t') \right|$$
$$\lesssim \eta c_{\mathbf{u}} T d^{\widehat{\ell}} \sqrt{\frac{\log(2n/\delta)}{m}} := \Delta_{m,n,\eta,T}. \tag{55}$$

It follows that, for any $w \in (0, 1)$, when $m \gtrsim T^2 d^{2\widehat{\ell}} \log(2n/\delta)/w^2$, we have $\Delta_{m,n,\eta,T} \le w$.

It follows from Lemma 24 that with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$, $\|h_t\|_{\mathcal{H}_K} \le B_h$, where B_h is defined in (37), and τ is required to satisfy $\tau \lesssim 1/(\eta T)$.

Theorem 20 requires that $m \gtrsim T^2 d^{2\widehat{\ell}} \log(2n/\delta)/\tau^2$. As a result, we also need to have

$$m \gtrsim \eta^2 T^4 d^{2\widehat{\ell}} \log(2n/\delta),$$

which leads to the condition (38) on m with $\eta = \Theta(1)$. ■

C.3.2. PROOF OF THEOREM 18

We need the following lemma, Lemma 21, which gives a sharp upper bound for the Rademacher complexity of a localized function class as a subset of the function class $\mathcal{F}(B, w)$, and then prove Theorem 18 using Lemma 21.

Lemma 21 *For every $B, w > 0$ every $r > 0$,*

$$\mathfrak{R}\left(\left\{f \in \mathcal{F}(B,w) \colon \mathbb{E}_P\left[f^2\right] \le r\right\}\right) \le \varphi_{B,w}(r), \tag{56}$$

where

$$\varphi_{B,w}(r) \coloneqq \min_{Q\colon Q\ge 0} \left((\sqrt{r} + w)\sqrt{\frac{Q}{n}} + B \left(\frac{\sum_{q=Q+1}^{\infty} \lambda_q}{n} \right)^{1/2} \right) + w. \tag{57}$$

Proof of Theorem 18 It follows from Theorem 17 that for every $t \in [T]$, conditioned on an event Ω with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right) - \exp\left(-\Theta(n)\right) - \delta$ over $\mathbf{Q}$ and $\mathbf{w}$, we have $\boldsymbol{a}(t) \in \mathcal{A}(\mathbf{S}, \boldsymbol{a}, T)$, and $f(\boldsymbol{a}(t), \cdot) = f_t \in \mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \boldsymbol{a}, T)$. Moreover, conditioned on the event Ω, $f_t = h_t + e_t$ where $h_t \in \mathcal{H}_K(B_h)$ and $e_t \in L^\infty$ with $\|e_t\|_\infty \le w$.

We then derive the sharp upper bound for $\mathbb{E}_P\left[(f_t - f^*)^2\right]$ by applying Theorem 11 to the function class $\mathcal{F} = \left\{F = (f - f^*)^2 : f \in \mathcal{F}(B_h, w)\right\}$. Since $B_0 \coloneqq (B_h + \gamma_0)\sqrt{\widehat{\ell} + 1} + 1 \ge (B_h + \gamma_0)\sqrt{\widehat{\ell} + 1} + w$, then $\|F\|_\infty \le B_0^2$ with $F \in \mathcal{F}$, so that $\mathbb{E}_P\left[F^2\right] \le B_0^2 \mathbb{E}_P[F]$. Let $T(F) = B_0^2 \mathbb{E}_P[F]$ for $F \in \mathcal{F}$. Then $\mathrm{Var}[F] \le \mathbb{E}_P\left[F^2\right] \le T(F) = B_0^2 \mathbb{E}_P[F]$.

We have

$$\begin{aligned}
&\mathfrak{R}\left(\{F \in \mathcal{F} \colon T(F) \le r\}\right) = \mathfrak{R}\left(\left\{(f - f^*)^2 \colon f \in \mathcal{F}(B_h, w), \mathbb{E}_P\left[(f - f^*)^2\right] \le \frac{r}{B_0^2}\right\}\right) \\
&\overset{①}{\le} 2B_0 \mathfrak{R}\left(\left\{f - f^* \colon f \in \mathcal{F}(B_h, w), \mathbb{E}_P\left[(f - f^*)^2\right] \le \frac{r}{B_0^2}\right\}\right) \\
&\overset{②}{\le} 4B_0 \mathfrak{R}\left(\left\{f \colon f \in \mathcal{F}(B_h, w), \mathbb{E}_P\left[f^2\right] \le \frac{r}{4B_0^2}\right\}\right).
\end{aligned} \tag{58}$$

where ① is due to the contraction property of Rademacher complexity in Theorem 9. Since $f^* \in \mathcal{F}(B_h, w)$, $f \in \mathcal{F}(B_h, w)$, we have $\frac{f - f^*}{2} \in \mathcal{F}(B_h, w)$ due to the fact that $\mathcal{F}(B_h, w)$ is symmetric and convex, and it follows that ② holds.

It follows from (58) and Lemma 21 that

$$B_0^2 \mathfrak{R}\left(\{F \in \mathcal{F} \colon T(F) \le r\}\right) \le 4B_0^3 \mathfrak{R}\left(\left\{f \colon f \in \mathcal{F}(B_h, w), \mathbb{E}_P\left[f^2\right] \le \frac{r}{4B_0^2}\right\}\right)$$

$$\leq 4B_0^3 \varphi_{B_h,w}\left(\frac{r}{4B_0^2}\right) := \psi(r). \tag{59}$$

ψ defined as the RHS of (59) is a sub-root function since it is nonnegative, nondecreasing and $\frac{\psi(r)}{\sqrt{r}}$ is nonincreasing. Let r^* be the fixed point of ψ, and $0 \leq r \leq r^*$. It follows from (Bartlett et al., 2005, Lemma 3.2) that $0 \leq r \leq \psi(r) = 4B_0^3 \varphi\left(\frac{r}{4B_0^2}\right)$. Therefore, by the definition of φ in (57), for every $0 \leq Q \leq n$, we have

$$\frac{r}{4B_0^3} \leq \left(\frac{\sqrt{r}}{2B_0} + w\right)\sqrt{\frac{Q}{n}} + B_h\left(\frac{\sum_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2} + w. \tag{60}$$

Solving the quadratic inequality (60) for r, we have

$$r \leq \frac{8B_0^4 Q}{n} + 8B_0^3\left(w\left(\sqrt{\frac{Q}{n}}+1\right) + B_h\left(\frac{\sum_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right). \tag{61}$$

(61) holds for every $0 \leq Q \leq n$, so we have

$$r \leq 8B_0^3 \min_{0\leq Q\leq n}\left(\frac{B_0 Q}{n} + w\left(\sqrt{\frac{Q}{n}}+1\right) + B_h\left(\frac{\sum_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right). \tag{62}$$

It then follows from (59) and Theorem 11 that with probability at least $1-\exp(-x)$ over the random training features $\mathbf{S}$,

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - \frac{K_0}{K_0-1}\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] - \frac{x\left(11B_0^2 + 26B_0^2 K_0\right)}{n} \leq \frac{704K_0}{B_0^2} r^*, \tag{63}$$

or

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim r^* + \frac{x}{n}, \tag{64}$$

with $K_0 = 2$ in (63). It follows from (62) and (64) with $Q = m_{\widehat{\ell}}$ that

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim \frac{m_{\widehat{\ell}}}{n} + w\left(\sqrt{\frac{Q}{n}}+1\right) + B_h\left(\frac{\sum_{q=m_{\widehat{\ell}}+1}^{\infty}\lambda_q}{n}\right)^{1/2} + \frac{x}{n}. \tag{65}$$

We note that $\lambda_q = 0$ for all $q > m_{\widehat{\ell}}$ in (65), and the above argument requires Theorem 17 which holds with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right) - \exp\left(-\Theta(n)\right) - \delta$ over the random noise $\mathbf{w}$. Setting $x = m_{\widehat{\ell}}$ in (65) and noting that $m_{\widehat{\ell}} = \Theta(d^{\widehat{\ell}})$ due to $\widehat{\ell} = \Theta(1)$ prove (40).

■

Proof of Theorem 19 We have

$$f_t(\mathbf{S}) = f^*(\mathbf{S}) + \mathbf{w} + \mathbf{v}(t) + \mathbf{e}(t), \tag{66}$$

where $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, $\vec{\mathbf{e}}(t) = \vec{\mathbf{e}}_1(t) + \vec{\mathbf{e}}_2(t)$ with $\vec{\mathbf{e}}_1(t) = -\left(\mathbf{I}_n - \eta\mathbf{K}_n\right)^t \mathbf{w}$ and $\left\|\vec{\mathbf{e}}_2(t)\right\|_2 \leq \sqrt{n}\tau$. We have $\eta\lambda_1 \in (0,1)$ if $\eta \in (0, 1/(\widehat{\ell}+1))$. It follows from (66) that

$$\begin{aligned}
&\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] = \frac{1}{n}\|f_t(\mathbf{S}) - f^*(\mathbf{S})\|_2^2 = \frac{1}{n}\|\mathbf{v}(t) + \mathbf{w} + \mathbf{e}(t)\|_2^2 \\
&= \frac{1}{n}\left\|-\left(\mathbf{I} - \eta\mathbf{K}_n\right)^t f^*(\mathbf{S}) + \left(\mathbf{I}_n - \left(\mathbf{I}_n - \eta\mathbf{K}_n\right)^t\right)\mathbf{w} + \vec{\mathbf{e}}_2(t)\right\|_2^2 \\
&\overset{①}{\leq} \frac{3}{n}\sum_{i=1}^n \left(1 - \eta\widehat{\lambda}_i\right)^{2t}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 + \frac{3}{n}\sum_{i=1}^n \left(1 - \left(1 - \eta\widehat{\lambda}_i\right)^t\right)^2 \left[\mathbf{U}^\top \mathbf{w}\right]_i^2 + \frac{3}{n}\left\|\vec{\mathbf{e}}_2(t)\right\|_2^2 \\
&\overset{②}{\leq} \frac{3\mu_0^2}{2e\eta t} + \frac{3}{n}\sum_{i=1}^n \left(1 - (1 - \eta\lambda_i)^t\right)^2 \left[\mathbf{U}^\top \mathbf{w}\right]_i^2 + 3\tau^2 \\
&\leq \Theta\left(\frac{1}{\eta t}\right) + 3 \cdot \underbrace{\frac{1}{n}\sum_{i=1}^n \left(1 - (1 - \eta\lambda_i)^t\right)^2 \left[\mathbf{U}^\top \mathbf{w}\right]_i^2}_{:= E_\varepsilon} = \Theta\left(\frac{1}{\eta t}\right) + 3E_\varepsilon.
\end{aligned} \tag{67}$$

Here ① follows from the Cauchy-Schwarz inequality, ② follows from (74) in the proof of Lemma 22. We then derive the upper bound for E_ε on the RHS of (67). We define the diagonal matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ with $\mathbf{R}_{ii} = \left(1 - (1 - \eta\lambda_i)^t\right)^2$. Then we have

$$E_\varepsilon = 1/n \cdot \mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top \mathbf{w}\mathbf{w}^\top\right)$$

It follows from (Wright, 1973) that

$$\begin{aligned}
&\Pr\left[1/n \cdot \mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top \mathbf{w}\mathbf{w}^\top\right) - \mathbb{E}\left[1/n \cdot \mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top \mathbf{w}\mathbf{w}^\top\right)\right] \geq u\right] \\
&\leq \exp\left(-c\min\left\{nu/\|\mathbf{R}\|_2, n^2u^2/\|\mathbf{R}\|_{\mathrm{F}}^2\right\}\right)
\end{aligned} \tag{68}$$

holds for all $u > 0$, and c is a positive constant. With $\eta_t = \eta t$ for all $t \geq 0$, we have

$$\begin{aligned}
\mathbb{E}\left[1/n \cdot \mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top \mathbf{w}\mathbf{w}^\top\right)\right] &\leq \frac{\sigma_0^2}{n}\sum_{i=1}^n \left(1 - \left(1 - \eta\widehat{\lambda}_i\right)^t\right)^2 \overset{①}{\leq} \frac{\sigma_0^2}{n}\sum_{i=1}^n \min\left\{1, \eta_t^2\widehat{\lambda}_i^2\right\} \\
&\leq \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^n \min\left\{\frac{1}{\eta_t}, \eta_t\widehat{\lambda}_i^2\right\} \overset{②}{\leq} \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^n \min\left\{\frac{1}{\eta_t}, \widehat{\lambda}_i\right\} \\
&= \sigma_0^2\eta_t\widehat{R}_K^2(\sqrt{1/\eta_t}) \leq \frac{1}{\eta_t}.
\end{aligned} \tag{69}$$

Here ① follows from the fact that $(1-\eta\widehat{\lambda}_i)^t \geq \max\left\{0, 1-t\eta\widehat{\lambda}_i\right\}$, and ② follows from $\min\{a,b\} \leq \sqrt{ab}$ for any nonnegative numbers a, b. Because $t \leq T \leq \widehat{T}$, we have $R_K(\sqrt{1/\eta_t}) \leq 1/(\sigma_0\eta_t)$, so the last inequality holds.

Moreover, we have the upper bounds for $\|\mathbf{R}\|_2$ and $\|\mathbf{R}\|_{\mathrm{F}}$ as follows. First, we have

$$\|\mathbf{R}\|_2 \leq \max_{i\in[n]}\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2 \leq \min\left\{1, \eta_t^2\widehat{\lambda}_i^2\right\} \leq 1. \tag{70}$$

We also have

$$\begin{aligned}\frac{1}{n}\|\mathbf{R}\|_{\mathrm{F}}^2 &= \frac{1}{n}\sum_{i=1}^n\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^4 \leq \frac{\eta_t}{n}\sum_{i=1}^n \min\left\{\frac{1}{\eta_t}, \eta_t^3\widehat{\lambda}_i^4\right\} \\ &\overset{③}{\leq} \frac{\eta_t}{n}\sum_{i=1}^n \min\left\{\widehat{\lambda}_i, \frac{1}{\eta_t}\right\} = \eta_t\widehat{R}_K^2(\sqrt{1/\eta_t}) \leq \frac{1}{\sigma_0^2\eta_t}.\end{aligned} \tag{71}$$

If $1/\eta_t \leq \eta_t^3(\widehat{\lambda}_i)^4$, then $\min\left\{1/\eta_t, \eta_t^3(\widehat{\lambda}_i)^4\right\} = 1/\eta_t$. Otherwise, we have $\eta_t^4\widehat{\lambda}_i^4 < 1$, so that $\eta_t\widehat{\lambda}_i < 1$ and it follows that $\min\left\{1/\eta_t, \eta_t^3(\widehat{\lambda}_i)^4\right\} \leq \eta_t^3\widehat{\lambda}_i^4 \leq \widehat{\lambda}_i$. As a result, ③ holds.

Combining (68)- (71), we have

$$\Pr\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right) - \mathbb{E}\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)\right] \geq u\right] \leq \exp\left(-cn\min\left\{u, u^2\sigma_0^2\eta_t\right\}\right).$$

Let $u = 1/\eta_t$ in the above inequality, we have

$$\exp\left(-cn\min\left\{u, u^2\sigma_0^2\eta_t\right\}\right) = \exp\left(-c'n/\eta_t\right) \leq \exp\left(-c'n\widehat{\varepsilon}_n^2\right),$$

where $c' = c\min\left\{1, \sigma_0^2\right\}$, and the last inequality is due to the fact that $1/\eta_t \geq \widehat{\varepsilon}_n^2$ since $t \leq T \leq \widehat{T}$. It follows that with probability at least $1-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$,

$$E_\varepsilon \leq u + \frac{1}{\eta_t} = \frac{2}{\eta_t}. \tag{72}$$

It then follows from (67), (68)-(72) that

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \leq \Theta\left(\frac{1}{\eta t}\right)$$

holds with probability at least $1-\exp\left(-c'n\widehat{\varepsilon}_n^2\right)$.

■

C.4. Proof of the Lemmas Required for the Proofs in Section C.3

Lemma 22 *Suppose $\widehat{\ell} \geq \ell_0$. Let $t \in [0:T]$, $\mathbf{v} = -\left(\mathbf{I}-\eta\mathbf{K}_n\right)^t f^*(\mathbf{S})$, $\mathbf{e} = -\left(\mathbf{I}-\eta\mathbf{K}_n\right)^t\mathbf{w}$, and $\eta \in (0, 1/(\widehat{\ell}+1))$. Then with probability at least $1-\exp\left(-\Theta(n)\right)$ over the random noise $\mathbf{w}$,*

$$\|\mathbf{v}\|_2 + \|\mathbf{e}\|_2 \leq (\Theta(\gamma_0) + \sigma_0 + 1)\cdot\sqrt{n}. \tag{73}$$

Proof When $t \in [T]$, we have

$$\begin{aligned}\|\mathbf{v}\|_2^2 &= \sum_{i=1}^n \left(1-\eta\widehat{\lambda}_i\right)^{2t} \left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 = \sum_{i=1}^n \left(1-\eta\widehat{\lambda}_i\right)^{2t} \left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \\ &\leq \sum_{i=1}^n \left(1-\eta\widehat{\lambda}_i\right)^{2t} \left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \overset{①}{\leq} \sum_{i=1}^n \frac{1}{2e\eta\widehat{\lambda}_i t} \left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \\ &\overset{②}{\leq} \frac{n\gamma_0^2}{2e\eta t} \leq \Theta(\gamma_0^2)\cdot n. \end{aligned} \tag{74}$$

Here ① follows from Lemma 30. ② follows from Lemma 29. This is because with $\widehat{\ell} \geq \ell_0$, $f^* \in \mathcal{H}_{K^{(r_0)}}(\gamma_0) \subseteq \mathcal{H}_K(\gamma_0)$. Moreover, it follows from the concentration inequality about quadratic forms of sub-Gaussian random variables in (Wright, 1973) that

$$\Pr\left[\|\mathbf{w}\|_2^2 - \mathbb{E}\left[\|\mathbf{w}\|_2^2\right] > n\right] \leq \exp\left(-\Theta(n)\right),$$

so that $\|\mathbf{e}\|_2 \leq \|\mathbf{w}\|_2 \leq \sqrt{\mathbb{E}\left[\|\mathbf{w}\|_2^2\right]} + \sqrt{n} = \sqrt{n}(\sigma_0+1)$ with probability at least $1-\exp\left(-\Theta(n)\right)$. As a result, (73) follows from this inequality and (74) for $t \geq 1$. When $t = 0$, $\|\mathbf{v}\|_2 \leq \Theta(\gamma_0)\sqrt{n}$, so that (73) still holds.

■

Lemma 23 *Suppose $\widehat{\ell} = \Theta(1)$. Let $0 < \eta < 1$, $0 \leq t \leq T-1$ for $T \geq 1$, and suppose that $\|\widehat{\mathbf{y}}(t') - \mathbf{y}\|_2 \leq c_{\mathbf{u}}\sqrt{n}$ holds for all $0 \leq t' \leq t$. Then for every $\delta \in (0,1)$, with probability at least $1-\delta$ over the random initialization $\mathbf{Q}$,*

$$\widehat{\mathbf{y}}(t+1) - \mathbf{y} = \left(\mathbf{I} - \eta\mathbf{K}_n\right)\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) + \mathbf{E}(t+1), \tag{75}$$

where $\|\mathbf{E}(t+1)\|_2 \leq \mathbf{E}_{m,n,\eta}$, and $\mathbf{E}_{m,n,\eta}$ satisfies

$$\mathbf{E}_{m,n,\eta} \lesssim \eta c_{\mathbf{u}} d^{\widehat{\ell}} \sqrt{\frac{\log(2n/\delta)}{m}}\sqrt{n}. \tag{76}$$

Proof Because $\|\widehat{\mathbf{y}}(t') - \mathbf{y}\|_2 \leq \sqrt{n}c_{\mathbf{u}}$ holds for all $t' \in [0,t]$. We have

$$\begin{aligned}\widehat{\mathbf{y}}(t+1) - \widehat{\mathbf{y}}(t) &= \frac{1}{\sqrt{m}} \sum_{r=1}^m \left(a_r(t+1) - a_r(t)\right) \sigma_{\boldsymbol{\tau}}(\vec{\mathbf{x}}_i, \vec{\mathbf{q}}_r) \\ &= -\frac{\eta}{n}\widehat{\mathbf{K}}(\widehat{\mathbf{y}}(t) - \mathbf{y}) \\ &= -\frac{\eta}{n}\mathbf{K}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) + \underbrace{\frac{\eta}{n}\left(\mathbf{K} - \widehat{\mathbf{K}}\right)\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)}_{:=\mathbf{E}(t+1)}. \end{aligned} \tag{77}$$

Since $\widehat{\ell} = \Theta(1)$, it follows from (91) of Theorem 25 that with probability at least $1-\delta$ over $\mathbf{Q}$, $\left\|\widehat{\mathbf{K}}_n - \mathbf{K}_n\right\|_2 \lesssim d^{\widehat{\ell}}\sqrt{\frac{\log(2n/\delta)}{m}}$. As a result, $\|\mathbf{E}(t+1)\|_2$ can be bounded by

$$\|\mathbf{E}(t+1)\|_2 \lesssim \eta c_{\mathbf{u}} \cdot d^{\widehat{\ell}} \sqrt{\frac{\log(2n/\delta)}{m}} \cdot \sqrt{n}. \tag{78}$$

(77) can be rewritten as

$$\widehat{\mathbf{y}}(t+1)-\mathbf{y}=\left(\mathbf{I}-\frac{\eta}{n}\mathbf{K}\right)(\widehat{\mathbf{y}}(t)-\mathbf{y})+\mathbf{E}(t+1),$$

which proves (75) with the upper bound for $\|\mathbf{E}(t+1)\|_2$ in (78). ■

Lemma 24 *Suppose $\widehat{\ell}=\Theta(1)\geq\ell_0$. Let $h_t(\cdot)=\sum_{t'=0}^{t-1}h(\cdot,t')$ for $t\in[T]$, $T\leq\widehat{T}$ where*

$$\begin{aligned}
h(\cdot,t')&=v(\cdot,t')+\widehat{e}(\cdot,t'),\\
v(\cdot,t')&=\frac{\eta}{n}\sum_{j=1}^{n}K(\cdot,\vec{\mathbf{x}}_j)\left[\mathbf{v}(t')\right]_j,\\
\widehat{e}(\cdot,t')&=\frac{\eta}{n}\sum_{j=1}^{n}K(\cdot,\vec{\mathbf{x}}_j)\left[\mathbf{e}(t')\right]_j,
\end{aligned}$$

where $\mathbf{v}(t')\in\mathcal{V}_{t'}$, $\mathbf{e}(t')\in\mathcal{E}_{t',\tau}$ for all $0\leq t'\leq t-1$. Suppose that $\tau\lesssim 1/(\eta T)$, then with probability at least $1-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$,

$$\|h_t\|_{\mathcal{H}_K}\leq B_h=\gamma_0+\sqrt{2}+1, \tag{79}$$

and B_h is also defined in (37).

Proof We have $\mathbf{v}(t)=-\left(\mathbf{I}-\eta\mathbf{K}_n\right)^t f^*(\mathbf{S})$, $\mathbf{e}(t)=\vec{\mathbf{e}}_1(t)+\vec{\mathbf{e}}_2(t)$ with $\vec{\mathbf{e}}_1(t)=-\left(\mathbf{I}-\eta\mathbf{K}_n\right)^t\mathbf{w}$, $\left\|\vec{\mathbf{e}}_2(t)\right\|_2\leq\sqrt{n}\tau$. We define

$$\widehat{e}_1(\cdot,t')\coloneqq-\frac{\eta}{n}\sum_{j=1}^{n}K(\vec{\mathbf{x}}_j,\mathbf{x})\left[\vec{\mathbf{e}}_1(t')\right]_j,\quad \widehat{e}_2(\cdot,t')\coloneqq-\frac{\eta}{n}\sum_{j=1}^{n}K(\vec{\mathbf{x}}_j,\mathbf{x})\left[\vec{\mathbf{e}}_2(t')\right]_j. \tag{80}$$

Let $\mathbf{\Sigma}$ be the diagonal matrix containing eigenvalues of $\mathbf{K}_n$, which are $\widehat{\lambda}_1\geq\widehat{\lambda}_2\ldots\geq\widehat{\lambda}_r\geq\widehat{\lambda}_{r+1}=\ldots\widehat{\lambda}_n=0$ where $r\leq n$ is the rank of the gram matrix $\mathbf{K}_n$. Then we have

$$\begin{aligned}
\sum_{t'=0}^{t-1}v(\mathbf{x},t')&=\frac{\eta}{n}\sum_{j=1}^{n}\sum_{t'=0}^{t-1}\left[\left(\mathbf{I}-\eta\mathbf{K}_n\right)^{t'}f^*(\mathbf{S})\right]_j K(\vec{\mathbf{x}}_j,\mathbf{x})\\
&=\frac{\eta}{n}\sum_{j=1}^{n}\sum_{t'=0}^{t-1}\left[\mathbf{U}\left(\mathbf{I}-\eta\mathbf{\Sigma}\right)^{t'}\mathbf{U}^\top f^*(\mathbf{S})\right]_j K(\vec{\mathbf{x}}_j,\mathbf{x}).
\end{aligned} \tag{81}$$

It follows from (81) that

$$\begin{aligned}
\left\|\sum_{t'=0}^{t-1}v(\cdot,t')\right\|_{\mathcal{H}_K}^2&=\frac{\eta^2}{n^2}f^*(\mathbf{S})^\top\mathbf{U}\sum_{t'=0}^{t-1}\left(\mathbf{I}-\eta\mathbf{\Sigma}\right)^{t'}\mathbf{U}^\top\mathbf{K}\mathbf{U}\sum_{t'=0}^{t-1}\left(\mathbf{I}-\eta\mathbf{\Sigma}\right)^{t'}\mathbf{U}^\top f^*(\mathbf{S})\\
&=\frac{1}{n}\left\|\eta\left(\mathbf{K}_n\right)^{1/2}\mathbf{U}\sum_{t'=0}^{t-1}\left(\mathbf{I}-\eta\mathbf{\Sigma}\right)^{t'}\mathbf{U}^\top f^*(\mathbf{S})\right\|_2^2
\end{aligned}$$

$$\leq \frac{1}{n}\sum_{i=1}^{r}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \leq \frac{1}{n}\sum_{i=1}^{r}\frac{\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2}{\widehat{\lambda}_i} \leq \gamma_0^2, \quad (82)$$

where the last inequality follows from Lemma 29.

Similarly, we have

$$\left\|\sum_{t'=0}^{t-1}\widehat{e}_1(\cdot,t')\right\|_{\mathcal{H}_K}^2 \leq \frac{1}{n}\sum_{i=1}^{r}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i}\left[\mathbf{U}^\top\mathbf{w}\right]_i^2, \quad (83)$$

It then follows from the argument in the proof of (Raskutti et al., 2014, Lemma 9) that the RHS of (83) is bounded w.h.p. We define a diagonal matrix $\mathbf{R}\in\mathbb{R}^{n\times n}$ with $\mathbf{R}_{ii} = \left(1-(1-\eta\widehat{\lambda}_i)^t\right)^2/\widehat{\lambda}_i$ for $i\in[n]$. Then the RHS of (83) is $1/n\cdot\operatorname{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)$. It follows from (Wright, 1973) that

$$\begin{aligned}&\Pr\left[1/n\cdot\operatorname{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)-\mathbb{E}\left[1/n\cdot\operatorname{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)\right]\geq u\right]\\ &\leq\exp\left(-c\min\left\{nu/\|\mathbf{R}\|_2, n^2u^2/\|\mathbf{R}\|_{\mathrm{F}}^2\right\}\right)\end{aligned} \quad (84)$$

for all $u>0$, and c is a positive constant. Let $\eta_t=\eta t$ for all $t\geq 0$, we have

$$\begin{aligned}\mathbb{E}\left[1/n\cdot\operatorname{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)\right] &\leq \frac{\sigma_0^2}{n}\sum_{i=1}^{r}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i} \overset{①}{\leq} \frac{\sigma_0^2}{n}\sum_{i=1}^{r}\min\left\{\frac{1}{\widehat{\lambda}_i}, \eta_t^2\widehat{\lambda}_i\right\}\\ &\leq \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^{r}\min\left\{\frac{1}{\eta_t\widehat{\lambda}_i}, \eta_t\widehat{\lambda}_i\right\} \overset{②}{\leq} \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^{r}\min\left\{1, \eta_t\widehat{\lambda}_i\right\}\\ &= \frac{\sigma_0^2\eta_t^2}{n}\sum_{i=1}^{r}\min\left\{\eta_t^{-1}, \widehat{\lambda}_i\right\} = \sigma_0^2\eta_t^2\widehat{R}_K^2(\sqrt{1/\eta_t})\leq 1.\end{aligned} \quad (85)$$

Here ① follows from the fact that $(1-\eta\widehat{\lambda}_i)^t\geq\max\left\{0, 1-t\eta\widehat{\lambda}_i\right\}$, and ② follows from $\min\{a,b\}\leq\sqrt{ab}$ for any nonnegative numbers a,b. Because $t\leq T\leq\widehat{T}$, we have $\widehat{R}_K(\sqrt{1/\eta_t})\leq 1/(\sigma_0\eta_t)$, so the last inequality holds.

Moreover, we have the upper bounds for $\|\mathbf{R}\|_2$ and $\|\mathbf{R}\|_{\mathrm{F}}$ as follows. First, we have

$$\|\mathbf{R}\|_2 \leq \max_{i\in[r]}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i} \leq \max_{i\in[r]}\min\left\{\frac{1}{\widehat{\lambda}_i}, \eta_t^2\widehat{\lambda}_i\right\}\leq\eta_t. \quad (86)$$

We also have

$$\frac{1}{n}\|\mathbf{R}\|_{\mathrm{F}}^2 = \frac{1}{n}\sum_{i=1}^{r}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^4}{(\widehat{\lambda}_i)^2} \leq \frac{\eta_t^3}{n}\sum_{i=1}^{r}\min\left\{\frac{1}{\eta_t^3\widehat{\lambda}_i^2}, \eta_t\widehat{\lambda}_i^2\right\}$$

$$\overset{③}{\leq} \frac{\eta_t^3}{n}\sum_{i=1}^{r}\min\left\{\widehat{\lambda}_i, \frac{1}{\eta_t}\right\} = \eta_t^3 \widehat{R}_K^2(\sqrt{1/\eta_t}) \leq \frac{\eta_t}{\sigma_0^2}, \tag{87}$$

where ③ follows from

$$\min\left\{\frac{1}{\eta_t^3\widehat{\lambda}_i^2}, \eta_t\widehat{\lambda}_i^2\right\} = \widehat{\lambda}_i \min\left\{\frac{1}{\eta_t^3\widehat{\lambda}_i^3}, \eta_t\widehat{\lambda}_i\right\} \leq \widehat{\lambda}_i.$$

Combining (83)-(87) with $u = 1$ in (84), we have

$$\begin{aligned}&\Pr\left[1/n \cdot \mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top \mathbf{w}\mathbf{w}^\top\right) - \mathbb{E}\left[1/n \cdot \mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top \mathbf{w}\mathbf{w}^\top\right)\right] \geq 1\right] \\ &\leq \exp\left(-c\min\left\{n/\eta_t, n\sigma_0^2/\eta_t\right\}\right) \leq \exp\left(-nc'/\eta_t\right) \leq \exp\left(-c'n\widehat{\varepsilon}_n^2\right),\end{aligned}$$

where $c' = c\min\left\{1, \sigma_0^2\right\}$, and the last inequality is due to the fact that $1/\eta_t \geq \widehat{\varepsilon}_n^2$ since $t \leq T \leq \widehat{T}$. It follows that with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$, $\left\|\sum_{t'=0}^{t-1}\widehat{e}_1(\cdot, t')\right\|_{\mathcal{H}_K}^2 \leq 2$.

We now find the upper bound for $\left\|\sum_{t'=0}^{t-1}\widehat{e}_2(\cdot, t')\right\|_{\mathcal{H}_K}$. We have

$$\left\|\widehat{e}_2(\cdot, t')\right\|_{\mathcal{H}_K}^2 \leq \frac{\eta^2}{n^2}\overrightarrow{\mathbf{e}}_2^\top(t')\mathbf{K}\overrightarrow{\mathbf{e}}_2(t') \leq \eta^2\widehat{\lambda}_1\tau^2,$$

so that

$$\left\|\sum_{t'=0}^{t-1}\widehat{e}_2(\cdot, t')\right\|_{\mathcal{H}_K} \leq \sum_{t'=0}^{t-1}\left\|\widehat{e}_2(\cdot, t')\right\|_{\mathcal{H}_K} \leq T\eta\sqrt{\widehat{\lambda}_1}\tau \leq 1, \tag{88}$$

if $\tau \lesssim 1/(\eta T)$ since $\widehat{\lambda}_1 \in (0, \Theta(1))$ due to the fact that $\widehat{\lambda}_1 \leq \sup_{\mathbf{x}\in\mathcal{X}} K(\mathbf{x}, \mathbf{x}) = \Theta(1)$.

Finally, it follows from (82), (84), and (88) that

$$\|h_t\|_{\mathcal{H}_K} \leq \left\|\sum_{t'=0}^{t-1}\widehat{v}(\cdot, t')\right\|_{\mathcal{H}_K} + \left\|\sum_{t'=0}^{t-1}\widehat{e}_1(\cdot, t')\right\|_{\mathcal{H}_K} + \left\|\sum_{t'=0}^{t-1}\widehat{e}_2(\cdot, t')\right\|_{\mathcal{H}_K} \leq \gamma_0 + \sqrt{2} + 1 = B_h.$$

■

Theorem 25 *Suppose $\widehat{\ell} = \Theta(1)$. For any fixed $\mathbf{x}' \in \mathcal{X}$ and every $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random initialization $\mathbf{Q} = \left\{\overrightarrow{\mathbf{q}}_r\right\}_{r=1}^m$, we have*

$$\sup_{\mathbf{x}\in\mathcal{X}}\left|\widehat{K}(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{x}')\right| \lesssim d^{\widehat{\ell}}\sqrt{\frac{\log 2/\delta}{m}}. \tag{89}$$

As a result, with probability at least $1 - \delta$ over $\mathbf{Q}$,

$$\sup_{\mathbf{x}\in\mathcal{X}, i\in[n]}\left|\widehat{K}(\mathbf{x}, \overrightarrow{\mathbf{x}}_i) - K(\mathbf{x}, \overrightarrow{\mathbf{x}}_i)\right| \lesssim d^{\widehat{\ell}}\sqrt{\frac{\log(2n/\delta)}{m}}, \tag{90}$$

$$\left\|\widehat{\mathbf{K}}_n - \mathbf{K}_n\right\|_2 \lesssim d^{\widehat{\ell}}\sqrt{\frac{\log(2n/\delta)}{m}}. \tag{91}$$

Proof First, it follows from (95) in the proof of Lemma 26 that for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$,

$$\left|\sigma_{\boldsymbol{\tau}}(\mathbf{x}, \mathbf{x}')\right| \leq \sup_{\mathbf{q} \in \mathcal{X}} \|\sigma_{\boldsymbol{\tau}}(\cdot, \mathbf{q})\|_{\mathcal{H}_\sigma}^2 = \sum_{\ell=0}^{\widehat{\ell}} \mu_{\sigma,\ell}^{\frac{1}{2}} N(d, \ell) = \sum_{\ell=0}^{\widehat{\ell}} N^{\frac{1}{2}}(d, \ell) = \Theta(d^{\widehat{\ell}/2}) \coloneqq p_0,$$

which follows from the fact that $N^{\frac{1}{2}}(d, \ell) \asymp d^{\frac{\ell}{2}}$ for every $\ell \in [0 : \widehat{\ell}]$ with $\widehat{\ell} = \Theta(1)$. The following arguments hold for every given $\mathbf{x}' \in \mathcal{X}$. We have

$$\mathbb{E}_{\vec{\mathbf{w}}}\left[\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{w}})\sigma_{\boldsymbol{\tau}}(\vec{\mathbf{w}}, \mathbf{x}')\right] = K(\cdot, \mathbf{x}').$$

It then follows from (94) of Lemma 26 that for every $t > 0$,

$$\Pr\left[\left\|\frac{1}{m}\sum_{r=1}^{m} \sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_r)\sigma_{\boldsymbol{\tau}}(\vec{\mathbf{q}}_r, \mathbf{x}') - K(\cdot, \mathbf{x}')\right\|_{\mathcal{H}_\sigma} < t\right] \geq 1 - 2\exp\left(-\frac{mt^2}{\Theta(d^{3\widehat{\ell}/2})}\right). \tag{92}$$

Noting that $1/m \cdot \sum_{r=1}^{m} \sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_r)\sigma_{\boldsymbol{\tau}}(\vec{\mathbf{q}}_r, \mathbf{x}') = \widehat{K}(\cdot, \mathbf{x}')$, it then follows from (92) that

$$\Pr\left[\left\|\widehat{K}(\cdot, \mathbf{x}') - K(\cdot, \mathbf{x}')\right\|_{\mathcal{H}_\sigma} < t\right] \geq 1 - 2\exp\left(-\frac{mt^2}{\Theta(d^{3\widehat{\ell}/2})}\right). \tag{93}$$

(89) then follows from (93) and the fact that

$$\sup_{\mathbf{x} \in \mathcal{X}} \left|\widehat{K}(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{x}')\right| \leq \left\|\widehat{K}(\cdot, \mathbf{x}') - K(\cdot, \mathbf{x}')\right\|_{\mathcal{H}_\sigma} \cdot \sup_{\mathbf{x} \in \mathcal{X}} \|\sigma_{\boldsymbol{\tau}}(\cdot, \mathbf{x})\|_{\mathcal{H}_\sigma},$$

and (90) and (91) follow from (89) by the union bound. ■

Lemma 26 *Suppose $\widehat{\ell} = \Theta(1)$, and p is a function defined on $\mathcal{X}$ and $\sup_{\mathbf{x} \in \mathcal{X}} |p(\mathbf{x})| \leq p_0$ for a positive number p_0. Then for every $r > 0$,*

$$\Pr\left[\left\|\frac{1}{m}\sum_{r=1}^{m} \sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_r)p(\vec{\mathbf{q}}_r) - \mathbb{E}_{\vec{\mathbf{w}}}\left[\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{w}})p(\vec{\mathbf{w}})\right]\right\|_{\mathcal{H}_\sigma} > r\right] \leq 2\exp\left(-\frac{mr^2}{\Theta(d^{\widehat{\ell}/2})p_0^2}\right). \tag{94}$$

Proof Let $\mathcal{B} = \mathcal{H}_K \subseteq L^2(\mathbb{S}^{d-1}, \mu)$, then $\mathcal{B} \in D(1, 1)$ (Pinelis, 1992). We then construct the martingale $\{f_k\}_{k \in [m]}$. First, for every $\mathbf{q} \in \mathcal{X}$, we have

$$\|\sigma_{\boldsymbol{\tau}}(\cdot, \mathbf{q})\|_{\mathcal{H}_\sigma}^2 = \sigma_{\boldsymbol{\tau}}(\mathbf{q}, \mathbf{q}) = \sum_{\ell=0}^{\widehat{\ell}} \mu_{\sigma,\ell}^{-\frac{1}{2}} P_\ell^{(d)}(1) = \sum_{\ell=0}^{\widehat{\ell}} N^{\frac{1}{2}}(d, \ell) = \Theta(d^{\widehat{\ell}/2}). \tag{95}$$

We define $p_1 \coloneqq 2p_0\|\sigma_{\boldsymbol{\tau}}(\cdot, \mathbf{q})\|_{\mathcal{H}_\sigma} = \Theta(d^{\widehat{\ell}/4})p_0$ for every $\mathbf{q} \in \mathcal{X}$. For each $k \in [m]$, we also define

$$f_k \coloneqq \mathbb{E}\left[\frac{1}{p_1\sqrt{m}}\sum_{r=1}^{m}\left(\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_r)p(\vec{\mathbf{q}}_r) - \mathbb{E}_{\vec{\mathbf{w}}}\left[\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{w}})p(\vec{\mathbf{w}})\right]\right) \middle| \mathcal{F}_k\right], \forall k \in [m],$$

where $\{\mathcal{F}_k\}_{k=0}^m$ is an increasing sequence of σ-algebras, $\mathcal{F}_k$ is the σ-algebra generated by $\left\{\vec{\mathbf{q}}_r\right\}_{r=1}^k$, and $\mathcal{F}_0$ is the trivial σ-algebra so that $f_0 = 0$. We note that

$$f_m = \frac{1}{p_1\sqrt{m}} \sum_{r=1}^m \left(\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_r)p(\vec{\mathbf{q}}_r) - \mathbb{E}_{\vec{\mathbf{w}}}\left[K(\cdot, \vec{\mathbf{w}})p(\vec{\mathbf{w}})\right]\right),$$
$$d_k = f_k - f_{k-1} = \frac{1}{p_1\sqrt{m}} \left(\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_k)p(\vec{\mathbf{q}}_k) - \mathbb{E}_{\vec{\mathbf{w}}}\left[\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{w}})p(\vec{\mathbf{w}})\right]\right), \forall k \in [m],$$

and $f^* = \max_{k\in[m]} \|f_k\|$. For every $k \in [m]$, we have

$$\begin{aligned}\|d_k\|_{\mathcal{H}_K} &= \left\| \frac{1}{p_1\sqrt{m}} \left(\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_k)p(\vec{\mathbf{q}}_k) - \mathbb{E}_{\vec{\mathbf{w}}}\left[\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{w}})p(\vec{\mathbf{w}})\right]\right)\right\|_{\mathcal{H}_\sigma} \\ &\overset{①}{\leq} \frac{1}{p_1\sqrt{m}} \left(p_0 \left\|\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_k)\right\|_{\mathcal{H}_\sigma} + p_0\mathbb{E}_{\vec{\mathbf{w}}}\left[\left\|\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{w}})\right\|_{\mathcal{H}_\sigma}\right]\right) \overset{②}{\leq} \frac{1}{\sqrt{m}}, \end{aligned} \tag{96}$$

where ① follows from the triangle inequality and the Jensen's inequality, and ② follows from (95).

It follows from (96) that $\sum_{k=1}^\infty \|d_k\|^2 \leq 1$. Applying Lemma 27 with the martingale $\{f_k\}_{k=0}^m$ and $\mathcal{B} = \mathcal{H}_\sigma \subseteq L^2(\mathbb{S}^{d-1}, \mu)$ with $B = 1$, we have $\Pr\left[f^* = \max_{k\in[m]} \|f_k\| > r\right] \leq 2\exp\left(-\frac{r^2}{2}\right)$. As a result, for every $r > 0$,

$$\Pr\left[\left\| \frac{1}{p_1\sqrt{m}} \sum_{r=1}^m \left(\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_r)p(\vec{\mathbf{q}}_r) - \mathbb{E}_{\vec{\mathbf{w}}}\left[\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{w}})p(\vec{\mathbf{w}})\right]\right)\right\|_{\mathcal{H}_\sigma} > r\right] \leq 2\exp\left(-\frac{r^2}{2}\right),$$

and it follows that

$$\Pr\left[\left\| \frac{1}{m} \sum_{r=1}^m \sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{q}}_r)p(\vec{\mathbf{q}}_r) - \mathbb{E}_{\vec{\mathbf{w}}}\left[\sigma_{\boldsymbol{\tau}}(\cdot, \vec{\mathbf{w}})p(\vec{\mathbf{w}})\right]\right\|_{\mathcal{H}_\sigma} > r\right] \leq 2\exp\left(-\frac{mr^2}{\Theta(d^{\widehat{\ell}/2})p_0^2}\right),$$

which completes the proof of (94).

■

In order to prove Lemma 26, we need to the following concentration inequality for independent random variables taking values in a Hilbert space $\mathcal{B}$ of functions defined on a measurable space (S, Σ_S, μ_S). Let $\{f_k\}_{n=0}^\infty$ be a martingale over a separable Banach space $(\mathcal{B}, \|\cdot\|)$ with respect to an increasing sequence of σ-algebras $\{\mathcal{F}_k\}_{n=0}^\infty$ and $f_0 = 0$. Define $d_k := f_k - f_{k-1}$ for $k \geq 1$, $d_0 = 0$, and $f^* := \sup_{n\geq 0} \|f_k\|$. The following lemma is about the martingale based concentration inequality for Banach space-valued random process (Pinelis, 1992).

Lemma 27 ((Pinelis, 1992, Theorem 2)) *Suppose that* $\sum_{k=1}^\infty \operatorname{esssup}\|d_k\|^2 \leq 1$ *where* $\operatorname{esssup}(f) = \inf_{a\in\mathbb{R}}\left\{\mu(f^{-1}(a, +\infty)) = 0\right\}$ *for a function denotes the essential supremum of a function, and* $\mathcal{B} \in D(A_1, A_2)$ *or* $\mathcal{B} \subseteq L^p(S, \Sigma, \mu)$ *with* $p \geq 2$. *Then for every* $r > 0$,

$$\Pr\left[f^* > r\right] \leq 2\exp\left(-\frac{r^2}{2B}\right) \tag{97}$$

with $B = p - 1$ *for* $\mathcal{B} \subseteq L^p(S, \Sigma_S, \mu_S)$.

Lemma 28 *The integral operator $T_K\colon L^2(\mathcal{X},\mu)\to L^2(\mathcal{X},\mu), (T_K f)(\mathbf{x}) \coloneqq \int_{\mathcal{X}} K(\mathbf{x},\mathbf{x}')f(\mathbf{x}')\mathrm{d}\mu(\mathbf{x}')$ is a positive, self-adjoint, and compact operator on $L^2(\mathcal{X},\mu)$. $\{Y_{\ell j}\}_{j\in[N(d,\ell)]}$ are the eigenfunction of T_K with $\mu_\ell = \mu_{\sigma,\ell}$ being the corresponding eigenvalue for every $\ell\in[0:\widehat{\ell}]$. Furthermore,*

$$K(\mathbf{x},\mathbf{x}') = \sum_{\ell=0}^{\widehat{\ell}}\sum_{j=1}^{N(d,\ell)} \mu_\ell Y_{\ell,j}(\mathbf{x})Y_{\ell,j}(\mathbf{x}'), \quad \mathbf{x},\mathbf{x}'\in\mathcal{X}, \tag{98}$$

and $\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}|K(\mathbf{x},\mathbf{x}')| = \widehat{\ell}+1 = \Theta(1)$.

Proof It follows from the definition of the activation function σ in (2) and the definition of K in (5) that

$$\begin{aligned} K(\mathbf{x},\mathbf{x}') &= \int_{\mathcal{X}} \sigma_{\boldsymbol{\tau}}(\mathbf{x},\mathbf{w})\sigma_{\boldsymbol{\tau}}(\mathbf{w},\mathbf{x}')\mathrm{d}\mu(\mathbf{w}) \\ &= \int_{\mathcal{X}} \left(\sum_{\ell=0}^{\widehat{\ell}}\sum_{j=1}^{N(d,\ell)} \mu_{\sigma,\ell}^{\frac{1}{2}} Y_{\ell,j}(\mathbf{x})Y_{\ell,j}(\mathbf{w})\right)\cdot\left(\sum_{\ell=0}^{\widehat{\ell}}\sum_{j=1}^{N(d,\ell)} \mu_{\sigma,\ell}^{\frac{1}{2}} Y_{\ell,j}(\mathbf{w})Y_{\ell,j}(\mathbf{x}')\right)\mathrm{d}\mu(\mathbf{w}) \\ &= \sum_{\ell=0}^{\widehat{\ell}}\sum_{j=1}^{N(d,\ell)} \mu_\ell Y_{\ell,j}(\mathbf{x})Y_{\ell,j}(\mathbf{x}'), \end{aligned} \tag{99}$$

where the last inequality follows from the orthogonality of the orthogonal set $\{Y_{\ell j}\}_{\ell\in[0:\widehat{\ell}],j\in[N(d,\ell)]}$.

It follows from (99) that K is PD kernel of finite rank over the compact set $\mathcal{X}$, so that T_K is a positive, self-adjoint, and compact operator on $L^2(\mathcal{X},\mu)$. Furthermore, for every $\ell\in[0:\widehat{\ell}]$ and every $j\in[N(d,\ell)]$, $T_K Y_{\ell,j} = \mu_\ell Y_{\ell,j}$, showing that μ_ℓ is the eigenvalue for every function in $\{Y_{\ell j}\}_{\ell\in[0:\widehat{\ell}],j\in[N(d,\ell)]}$.

Finally, considering the RKHS associated with the PD kernel K, we have

$$\begin{aligned} \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \left|K(\mathbf{x},\mathbf{x}')\right| &= \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \left|\left\langle K(\cdot,\mathbf{x}), K(\cdot,\mathbf{x}')\right\rangle_{\mathcal{H}_K}\right| \le \sup_{\mathbf{x}\in\mathcal{X}} K(\mathbf{x},\mathbf{x}) \\ &= \sum_{\ell=0}^{\widehat{\ell}} \mu_\ell N(d,\ell)P_\ell^{(d)}(1) = \widehat{\ell}+1 = \Theta(1), \end{aligned}$$

which is due to the fact that $P_k^{(d)}(1) = 1$ for all $k\ge 0$ discussed in Section B of this appendix. ■

Lemma 29 (In the proof of (Raskutti et al., 2014, Lemma 8)) *Let r be the rank of the gram matrix $\mathbf{K}$ for the kernel K over the training features $\mathbf{S}$. Then for any $f\in\mathcal{H}_K(\gamma_0)$, we have*

$$\frac{1}{n}\sum_{i=1}^{r}\frac{\left[\mathbf{U}^\top f(\mathbf{S})\right]_i^2}{\widehat{\lambda}_i} \le \gamma_0^2. \tag{100}$$

Lemma 30 *For any positive real number $a\in(0,1)$ and natural number t, we have*

$$(1-a)^t \le e^{-ta} \le \frac{1}{eta}. \tag{101}$$

Proof The result follows from the facts that $\log(1-a) \le a$ for $a \in (0,1)$ and $\sup_{u\in\mathbb{R}} ue^{-u} \le 1/e$. ■

Lemma 31 ((Yang and Li, 2024, Lemma B.7)) *With probability at least* $1 - 4\exp(-\Theta(n\varepsilon_n^2))$,

$$\varepsilon_n^2 \lesssim \widehat{\varepsilon}_n^2, \quad \widehat{\varepsilon}_n^2 \lesssim \varepsilon_n^2. \tag{102}$$

Proof of Lemma 21 We first decompose the Rademacher complexity of the function class $\{f \in \mathcal{F}(B,w)\colon \mathbb{E}_P\left[f^2\right] \le r\}$ into two terms as follows:

$$\begin{aligned}
&\mathfrak{R}\left(\{f\colon f\in\mathcal{F}(B,w), \mathbb{E}_P\left[f^2\right]\le r\}\right)\\
&\le \underbrace{\frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}(B,w)\colon \mathbb{E}_P[f^2]\le r}\sum_{i=1}^n \sigma_i h(\vec{\mathbf{x}}_i)\right]}_{:=\mathcal{R}_1} + \underbrace{\frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}(B,w)\colon \mathbb{E}_P[f^2]\le r}\sum_{i=1}^n \sigma_i e(\vec{\mathbf{x}}_i)\right]}_{:=\mathcal{R}_2}.
\end{aligned} \tag{103}$$

We now analyze the upper bounds for $\mathcal{R}_1, \mathcal{R}_2$ on the RHS of (103).

Derivation for the upper bound for $\mathcal{R}_1$.

According to Definition 36 and Theorem 17, for any $f \in \mathcal{F}(B,w)$, we have $f = h + e$ with $h \in \mathcal{H}_K(B)$, $e \in L^\infty$, $\|e\|_\infty \le w$.

When $\mathbb{E}_P\left[f^2\right] \le r$, it follows from the triangle inequality that $\|h\|_{L^2} \le \|f\|_{L^2} + \|e\|_{L^2} \le \sqrt{r} + w := r_h$. We now consider $h \in \mathcal{H}_K(B)$ with $\|h\|_{L^2} \le r_h$ in the remaining of this proof. We have

$$\begin{aligned}
\sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i) &= \sum_{i=1}^n \sigma_i \left(h(\vec{\mathbf{x}}_i) + e(\vec{\mathbf{x}}_i)\right)\\
&= \left\langle h, \sum_{i=1}^n \sigma_i K(\cdot, \vec{\mathbf{x}}_i)\right\rangle_{\mathcal{H}_K} + \sum_{i=1}^n \sigma_i e(\vec{\mathbf{x}}_i).
\end{aligned} \tag{104}$$

Because $\{v_q\}_{q\ge 1}$ is an orthonormal basis of $\mathcal{H}_K$, for any $0 \le Q \le n$, we further express the first term on the RHS of (104) as

$$\begin{aligned}
\left\langle h, \sum_{i=1}^n \sigma_i K(\cdot, \vec{\mathbf{x}}_i)\right\rangle_{\mathcal{H}_K} &= \left\langle \sum_{q=1}^Q \sqrt{\lambda_q}\langle h, v_q\rangle_{\mathcal{H}_K} v_q, \sum_{q=1}^Q \frac{1}{\sqrt{\lambda_q}}\left\langle \sum_{i=1}^n \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q\right\rangle_{\mathcal{H}_K} v_q\right\rangle_{\mathcal{H}_K}\\
&\quad + \left\langle h, \sum_{q>Q}\left\langle \sum_{i=1}^n \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q\right\rangle_{\mathcal{H}_K} v_q\right\rangle_{\mathcal{H}_K}.
\end{aligned} \tag{105}$$

Due to the fact that $h \in \mathcal{H}_K$, $h = \sum_{q=1}^\infty \boldsymbol{\beta}_q^{(h)} v_q = \sum_{q=1}^\infty \sqrt{\lambda_q}\boldsymbol{\beta}_q^{(h)} e_q$ with $v_q = \sqrt{\lambda_q} e_q$. Therefore, $\|h\|_{L^2}^2 = \sum_{q=1}^\infty \lambda_q {\boldsymbol{\beta}_q^{(h)}}^2$, and

$$\left\|\sum_{q=1}^Q \sqrt{\lambda_q}\langle h, v_q\rangle_{\mathcal{H}_K} v_q\right\|_{\mathcal{H}_K} = \left\|\sum_{q=1}^Q \sqrt{\lambda_q}\boldsymbol{\beta}_q^{(h)} v_q\right\|_{\mathcal{H}_K} = \sqrt{\sum_{q=1}^Q \lambda_q {\boldsymbol{\beta}_q^{(h)}}^2} \le \|h\|_{L^2} \le r_h. \tag{106}$$

According to Mercer's Theorem, because the kernel K is continuous, symmetric and positive definite, it has the decomposition

$$K(\cdot, \vec{\mathbf{x}}_i) = \sum_{j=1}^{\infty} \lambda_j e_j(\cdot) e_j(\vec{\mathbf{x}}_i),$$

so that we have

$$\begin{aligned}\left\langle \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} &= \left\langle \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{\infty} \lambda_j e_j e_j(\vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} = \left\langle \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{\infty} \sqrt{\lambda_j} e_j(\vec{\mathbf{x}}_i) \cdot v_j, v_q \right\rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^{n} \sigma_i \sqrt{\lambda_q} e_q(\vec{\mathbf{x}}_i). \end{aligned} \tag{107}$$

Combining (105), (106), and (107), we have

$$\begin{aligned}\left\langle h, \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i) \right\rangle &\overset{①}{\leq} \left\| \sum_{q=1}^{Q} \sqrt{\lambda_q} \langle h, v_q \rangle_{\mathcal{H}_K} v_q \right\|_{\mathcal{H}_K} \cdot \left\| \sum_{q=1}^{Q} \frac{1}{\sqrt{\lambda_q}} \left\langle \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} v_q \right\|_{\mathcal{H}_K} \\ &\quad + \|h\|_{\mathcal{H}_K} \cdot \left\| \sum_{q=Q+1}^{\infty} \left\langle \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} v_q \right\|_{\mathcal{H}_K} \\ &\leq \|h\|_{L^2} \left\| \sum_{q=1}^{Q} \sum_{i=1}^{n} \sigma_i e_q(\vec{\mathbf{x}}_i) v_q \right\|_{\mathcal{H}_K} + B \left\| \sum_{q=Q+1}^{\infty} \sum_{i=1}^{n} \sigma_i \sqrt{\lambda_q} e_q(\vec{\mathbf{x}}_i) v_q \right\|_{\mathcal{H}_K} \\ &\leq r_h \sqrt{\sum_{q=1}^{Q} \left(\sum_{i=1}^{n} \sigma_i e_q(\vec{\mathbf{x}}_i) \right)^2} + B \sqrt{\sum_{q=Q+1}^{\infty} \left(\sum_{i=1}^{n} \sigma_i \sqrt{\lambda_q} e_q(\vec{\mathbf{x}}_i) \right)^2}, \end{aligned} \tag{108}$$

where ① is due to Cauchy-Schwarz inequality. Moreover, by Jensen's inequality we have

$$\mathbb{E}\left[\sqrt{\sum_{q=1}^{Q} \left(\sum_{i=1}^{n} \sigma_i e_q(\vec{\mathbf{x}}_i) \right)^2} \right] \leq \sqrt{\mathbb{E}\left[\sum_{q=1}^{Q} \left(\sum_{i=1}^{n} \sigma_i e_q(\vec{\mathbf{x}}_i) \right)^2 \right]} \leq \sqrt{\mathbb{E}\left[\sum_{q=1}^{Q} \sum_{i=1}^{n} e_q^2(\vec{\mathbf{x}}_i) \right]} = \sqrt{nQ}. \tag{109}$$

and similarly,

$$\mathbb{E}\left[\sqrt{\sum_{q=Q+1}^{\infty} \left(\sum_{i=1}^{n} \sigma_i \sqrt{\lambda_q} e_q(\vec{\mathbf{x}}_i) \right)^2} \right] \leq \sqrt{\mathbb{E}\left[\sum_{q=Q+1}^{\infty} \lambda_q \sum_{i=1}^{n} e_q^2(\vec{\mathbf{x}}_i) \right]} = \sqrt{n \sum_{q=Q+1}^{\infty} \lambda_q}. \tag{110}$$

Since (108)-(110) hold for all $Q \geq 0$, it follows that

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}_K(B), \|h\|_{L^2} \leq r_h} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(\vec{\mathbf{x}}_i) \right] \leq \min_{Q:\, Q \geq 0} \left(r_h \sqrt{nQ} + B \sqrt{n \sum_{q=Q+1}^{\infty} \lambda_q} \right). \tag{111}$$

It follows from (103), (104), and (111) that

$$\mathcal{R}_1 \le \frac{1}{n}\mathbb{E}\left[\sup_{h\in\mathcal{H}_K(B),\|h\|_{L^2}\le r_h}\sum_{i=1}^{n}\sigma_i h(\vec{\mathbf{x}}_i)\right] \le \min_{Q\colon Q\ge 0}\left(r_h\sqrt{\frac{Q}{n}} + B\left(\frac{\sum_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right). \quad (112)$$

Derivation for the upper bound for $\mathcal{R}_2$.
Because $\left|1/n\sum_{i=1}^{n}\sigma_i e(\vec{\mathbf{x}}_i)\right| \le w$ when $\|e\|_\infty \le w$, we have

$$\mathcal{R}_2 \le \frac{1}{n}\mathbb{E}\left[\sup_{e\in L^\infty\colon \|e\|_\infty\le w}\sum_{i=1}^{n}\sigma_i e(\vec{\mathbf{x}}_i)\right] \le w. \quad (113)$$

It follows from (112) and (113) that

$$\mathfrak{R}\left(\left\{f\colon f\in\mathcal{F}(B,w),\mathbb{E}_P\left[f^2\right]\le r\right\}\right) \le \min_{Q\colon Q\ge 0}\left(r_h\sqrt{\frac{Q}{n}} + B\left(\frac{\sum_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right) + w.$$

Plugging r_h in the RHS of the above inequality completes the proof. ■

Appendix D. Proofs for Channel Selection

Proof of Theorem 1 We denote $\tau_\ell(1)$ as τ_ℓ for all $\ell\in[0:L]$ in this proof. Let $\boldsymbol{\beta}\in\mathbb{R}^{r_0}$ be the vector with the elements $\{\beta_{\ell,j}\colon \ell\in[0:\ell_0], j\in[N(d,\ell)]\}$.

We note that $f^*(\mathbf{S}) = \mathbf{Y}(\mathbf{S},r_0)\boldsymbol{\beta}$ and $\mathbf{y} = f^*(\mathbf{S}) + \mathbf{w}$, so that $\tau_\ell = \tau_{*,\ell} + \tau_{\mathbf{w},\ell} + \widehat{\tau}_{\mathbf{w},\ell}$, and

$$\begin{aligned}
\tau_{*,\ell} &\coloneqq \frac{1}{n^2 m}\boldsymbol{\beta}^\top\mathbf{Y}^\top(\mathbf{S},r_0)\mathbf{Y}(\mathbf{S},\ell)\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},r_0)\boldsymbol{\beta},\\
\tau_{\mathbf{w},\ell} &\coloneqq \frac{1}{n^2 m}\mathbf{w}^\top\mathbf{Y}(\mathbf{S},\ell)\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{w},\\
\widehat{\tau}_{\mathbf{w},\ell} &\coloneqq \frac{2}{n^2 m}\boldsymbol{\beta}^\top\mathbf{Y}^\top(\mathbf{S},r_0)\mathbf{Y}(\mathbf{S},\ell)\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{w},
\end{aligned}$$

where $\boldsymbol{\beta}\in\mathbb{R}^{r_0}$, and the elements of $\boldsymbol{\beta}$ form the enumeration of $\{\beta_{\ell j}\}_{0\le\ell\le\ell_0, j\in[N(d,\ell)]}$. We let

$$\begin{aligned}
\mathbf{Y}^\top(\mathbf{S},r_0)\mathbf{Y}(\mathbf{S},\ell)/n &= \mathbf{E}_{r_0,\ell} + \boldsymbol{\Delta}_{r_0,\ell}, \mathbf{E}_{r_0,\ell} \coloneqq \mathbb{E}\left[\mathbf{Y}^\top(\mathbf{S},r_0)\mathbf{Y}(\mathbf{S},\ell)\right],\\
\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},r_0)/n &= \mathbf{E}_{m_L,r_0} + \boldsymbol{\Delta}_{m_L,r_0}, \mathbf{E}_{m_L,r_0} \coloneqq \mathbb{E}\left[\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},r_0)\right],\\
\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)/m &= \mathbf{E}_{\ell,m_L} + \boldsymbol{\Delta}_{\ell,m_L}, \mathbf{E}_{\ell,m_L} \coloneqq \mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L).
\end{aligned}$$

Here $\mathbf{E}_{r_0,\ell}, \boldsymbol{\Delta}_{r_0,\ell}, \in \mathbb{R}^{r_0\times N(d.\ell)}$, $\mathbf{E}_{m_L,r_0}, \boldsymbol{\Delta}_{m_L,r_0} \in \mathbb{R}^{m_L\times r_0}$, and $\mathbf{E}_{\ell,m_L}, \boldsymbol{\Delta}_{\ell,m_L} \in \mathbb{R}^{N(d.\ell)\times m_L}$. We let $\mathbf{A}_{[s:t]}$ to denote the submatrix of $\mathbf{A}$ formed by rows of $\mathbf{A}$ with row indices in $[s:t]$, and

$\mathbf{A}^{[s:t]}$ to denote the submatrix of $\mathbf{A}$ formed by columns of $\mathbf{A}$ with columns indices in $[s:t]$. Then if $0 \le \ell \le \ell_0$,

$$[\mathbf{E}_{r_0,\ell}]_{[m_{\ell-1}+1:m_\ell]} = \mathbf{I}_{N(d,\ell)}, \quad [\mathbf{E}_{r_0,\ell}]_j = \mathbf{0} \text{ for all } j \notin [m_{\ell-1}+1 : m_\ell],$$

and $\mathbf{E}_{r_0,\ell} = \mathbf{0}$ if $\ell > \ell_0$. Similarly,

$$[\mathbf{E}_{m_L,r_0}]_{[1:r_0]} = \mathbf{I}_{r_0}, \quad [\mathbf{E}_{m_L,r_0}]_{[r_0+1:m_L]} = \mathbf{0},$$

and

$$[E_{\ell,m_L}]^{[m_{\ell-1}+1:m_\ell]} = \mathbf{I}_{N(d,\ell)}, \quad [E_{\ell,m_L}]^{(j)} = \mathbf{0} \text{ for all } j \notin [m_{\ell-1}+1 : m_\ell].$$

With $\min\{m,n\} > 4m_L \log(4m_L/\delta)$, it follows from Lemma 34 that, with probability at least $1-\delta$ for every $\delta \in (0,1)$,

$$\max\left\{\|\mathbf{\Delta}_{r_0,\ell}\|_2, \|\mathbf{\Delta}_{m_L,r_0}\|_2\right\} \le \sqrt{\log\left(\frac{4m_L}{\delta}\right)\frac{4m_L}{n}} \le 1, \quad \|\mathbf{\Delta}_{\ell,m_L}\|_2 \le \sqrt{\log\left(\frac{4m_L}{\delta}\right)\frac{4m_L}{m}} \le 1, \tag{114}$$

which are due to the fact that $\max\left\{\|\mathbf{\Delta}_{r_0,\ell}\|_2, \|\mathbf{\Delta}_{m_L,r_0}\|_2\right\} \le \|\Delta_{\mathbf{S},m_L}\|_2$ and $\|\mathbf{\Delta}_{\ell,m_L}\|_2 \le \|\Delta_{\mathbf{Q},m_L}\|_2$, where $\Delta_{\mathbf{S},m_L}, \Delta_{\mathbf{Q},m_L}$ are defined in (137)-(138). We have

$$\begin{aligned}\tau_{*,\ell} &= \frac{1}{n^2 m}\boldsymbol{\beta}^\top \mathbf{Y}^\top(\mathbf{S},r_0)\mathbf{Y}(\mathbf{S},\ell)\underbrace{\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},r_0)}_{:=\mathbf{D}_1}\boldsymbol{\beta} \\ &= \frac{1}{nm}\boldsymbol{\beta}^\top E_{r_0,\ell}\mathbf{D}_1\boldsymbol{\beta} + \underbrace{\frac{1}{nm}\boldsymbol{\beta}^\top \mathbf{\Delta}_{r_0,\ell}\mathbf{D}_1\boldsymbol{\beta}}_{:=E_1}.\end{aligned} \tag{115}$$

It follows from (114) that

$$\left\|\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)\right\|_2 \le 2m, \quad \left\|\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},r_0)\right\|_2 \le 2n. \tag{116}$$

It follows from (116) that

$$\|\mathbf{D}_1\|_2 \le 4mn. \tag{117}$$

It then follows from (114) and (117) that

$$|E_1| \le \frac{1}{nm}\cdot\|\boldsymbol{\beta}\|_2^2\|\mathbf{\Delta}_{r_0,\ell}\|_2\|\mathbf{D}_1\|_2 \le 4\gamma_0^2\sqrt{\log\left(\frac{4m_L}{\delta}\right)\frac{4m_L}{n}}. \tag{118}$$

We have

$$\begin{aligned}\frac{1}{nm}\boldsymbol{\beta}^\top E_{r_0,\ell}\mathbf{D}_1\boldsymbol{\beta} &= \frac{1}{n}\boldsymbol{\beta}^\top E_{r_0,\ell}E_{\ell,m_L}\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},r_0)\boldsymbol{\beta} \\ &\quad + \underbrace{\frac{1}{n}\boldsymbol{\beta}^\top E_{r_0,\ell}\mathbf{\Delta}_{\ell,m_L}\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},r_0)\boldsymbol{\beta}}_{:=E_2},\end{aligned} \tag{119}$$

and

$$|E_2| \leq \frac{1}{n}\|\boldsymbol{\beta}\|_2^2 \|\boldsymbol{\Delta}_{\ell,m_L}\|_2 \left\|\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},r_0)\right\|_2 \leq 2\gamma_0^2\sqrt{\log\left(\frac{4m_L}{\delta}\right)\frac{4m_L}{m}}. \tag{120}$$

We further have

$$\begin{aligned}\frac{1}{n}\boldsymbol{\beta}^\top E_{r_0,\ell}E_{\ell,m_L}\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},r_0)\boldsymbol{\beta} &= \boldsymbol{\beta}^\top E_{r_0,\ell}E_{\ell,m_L}E_{m_L,r_0}\boldsymbol{\beta} \\ &\quad + \underbrace{\boldsymbol{\beta}^\top E_{r_0,\ell}E_{\ell,m_L}\boldsymbol{\Delta}_{m_L,r_0}\boldsymbol{\beta}}_{:=E_3},\end{aligned} \tag{121}$$

and

$$|E_3| \leq \gamma_0^2\sqrt{\log\left(\frac{4m_L}{\delta}\right)\frac{4m_L}{n}}. \tag{122}$$

We note that

$$\boldsymbol{\beta}^\top E_{r_0,\ell}E_{\ell,m_L}E_{m_L,r_0}\boldsymbol{\beta} = \begin{cases} 0 & \ell_0 < \ell \leq L \\ \sum_{j\in N(d,\ell)}\beta_{\ell,j}^2 & \ell \in [0:\ell_0]. \end{cases} \tag{123}$$

It follows that when $\ell \in [0:\ell_0]$,

$$\boldsymbol{\beta}^\top E_{r_0,\ell}E_{\ell,m_L}E_{m_L,r_0}\boldsymbol{\beta} \geq N(d,\ell)\min_{\ell\in[0:\ell_0],j\in[N(d,\ell)]]}\beta_{\ell,j}^2 \geq \beta_0^2. \tag{124}$$

For $\varepsilon_0 \leq \beta_0^2/3$, with

$$m \geq \frac{256m_L\gamma_0^4}{\varepsilon_0^2}\log\left(\frac{4m_L}{\delta}\right), n \geq \max\left\{\frac{400m_L\gamma_0^4}{\varepsilon_0^2}\log\left(\frac{4m_L}{\delta}\right), \frac{32m_L(\sigma_0^2+1)}{\varepsilon_0}, \frac{8192\gamma_0^2m_L(\sigma_0^2+1)}{\varepsilon_0^2}\right\},$$

by Lemma 32 and Lemma 33 we have

$$\begin{aligned} E_1 + E_3 \leq 5\gamma_0^2\sqrt{\log\left(\frac{4m_L}{\delta}\right)\frac{4m_L}{n}} \leq \varepsilon_0/2, \quad E_2 \leq 2\gamma_0^2\sqrt{\log\left(\frac{4m_L}{\delta}\right)\frac{4m_L}{m}} \leq \varepsilon_0/4, \\ |\tau_{\mathbf{w},\ell}| \leq \frac{4m_L(\sigma_0^2+1)}{n} \leq \varepsilon_0/8, \quad |\widehat{\tau}_{\mathbf{w},\ell}| \leq 8\sqrt{2}\gamma_0\sqrt{\frac{m_L(\sigma_0^2+1)}{n}} \leq \varepsilon_0/8, \end{aligned}$$

which hold for all $\ell \in [0:L]$. Combining the above results, we have

$$\begin{cases} \tau_{*,\ell} \geq \beta_0^2 - E_1 - E_2 - E_3 \geq \frac{9\varepsilon_0}{4}, & \ell \in [0:\ell_0], \\ |\tau_{*,\ell}| \leq E_1 + E_2 + E_3 \leq \frac{3\varepsilon_0}{4}, & \ell_0 < \ell \leq L. \end{cases} \tag{125}$$

As a result, when $\ell \in [0:\ell_0]$, we have

$$\tau_\ell = \tau_{*,\ell} + \tau_{\mathbf{w},\ell} + \widehat{\tau}_{\mathbf{w},\ell} \geq \frac{9\varepsilon_0}{4} - \frac{\varepsilon_0}{8} - \frac{\varepsilon_0}{8} \geq 2\varepsilon_0. \tag{126}$$

When $\ell_0 < \ell \leq L$, we have

$$|\tau_\ell| \leq |\tau_{*,\ell}| + |\tau_{\mathbf{w},\ell}| + |\widehat{\tau}_{\mathbf{w},\ell}| \leq \frac{3\varepsilon_0}{4} + \frac{\varepsilon_0}{8} + \frac{\varepsilon_0}{8} \leq \varepsilon_0, \tag{127}$$

which completes the proof with the union bound.

■

Lemma 32 *For every $\delta \in (0,1)$, suppose $\min\{n,m\} > 4m_L \log(4m_L/\delta)$. Then with probability at least $1 - \exp(-\Theta(m_L)) - \delta$, for every $\ell \in [0:L]$,*

$$|\tau_{\mathbf{w},\ell}| \leq \frac{4m_L(\sigma_0^2+1)}{n}. \tag{128}$$

Proof We first define $\mathbf{M} = \mathbf{Y}(\mathbf{S},\ell)\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)\mathbf{Y}^\top(\mathbf{S},m_L)/\left(n^2m\right) \in \mathbb{R}^{n\times n}$, then $\tau_{\mathbf{w},\ell} = \mathbf{w}^\top\mathbf{M}\mathbf{w}$. With $n > 4m_L\log(4m_L/\delta)$, it follows from (139) in Lemma 34 that both $\mathbf{Y}(\mathbf{S},\ell)$ and $\mathbf{Y}(\mathbf{S},m_L)$ are of full column rank. We let the singular value decomposition of $\mathbf{Y}(\mathbf{S},m_L)$ and $\mathbf{Y}(\mathbf{S},\ell)$ be

$$\mathbf{Y}(\mathbf{S},m_L) = \mathbf{U}^{(L)}\mathbf{\Sigma}^{(L)}{\mathbf{V}^{(L)}}^\top, \mathbf{Y}(\mathbf{S},\ell) = \mathbf{U}^{(\ell)}\mathbf{\Sigma}^{(\ell)}{\mathbf{V}^{(\ell)}}^\top, \tag{129}$$

where $\mathbf{U}^{(L)} \in \mathbb{R}^{n\times m_L}, \mathbf{V}^{(L)} \in \mathbb{R}^{m_L\times m_L}, \mathbf{U}^{(\ell)} \in \mathbb{R}^{n\times N(d,\ell)}, \mathbf{V}^{(\ell)} \in \mathbb{R}^{N(d,\ell)\times N(d,\ell)}$ are orthogonal matrices, $\mathbf{\Sigma}^{(L)} \in \mathbb{R}^{m_L\times m_L}, \mathbf{\Sigma}^{(\ell)} \in \mathbb{R}^{N(d,\ell)\times N(d,\ell)}$ are diagonal matrices. We can then express $\mathbf{M}$ as

$$\mathbf{M} = \frac{1}{n}\mathbf{U}^{(\ell)}\underbrace{(\mathbf{\Sigma}^{(\ell)}/\sqrt{n}){\mathbf{V}^{(\ell)}}^\top(\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)/m)\mathbf{V}^{(L)}(\mathbf{\Sigma}^{(L)}/\sqrt{n})}_{:=\mathbf{D}}{\mathbf{U}^{(L)}}^\top. \tag{130}$$

The operator norm of $\mathbf{D}$ in (130) can be bounded by

$$\|\mathbf{D}\|_2 \leq 4, \tag{131}$$

which holds with probability at least $1-\delta$. It follows from (139) in Lemma 34 again that $\left\|\mathbf{\Sigma}^{(\ell)}/\sqrt{n}\right\|_2 \leq \sqrt{2}$, $\left\|\mathbf{\Sigma}^{(L)}/\sqrt{n}\right\|_2 \leq \sqrt{2}$, and $\left\|(\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)/m)\right\|_2 \leq 2$, so that (131) holds. Moreover, because the column space of $\mathbf{Y}(\mathbf{S},\ell)$ is a subspace of the column space of $\mathbf{Y}(\mathbf{S},m_L)$, we have $\left\|{\mathbf{U}^{(\ell)}}^\top\mathbf{w}\right\|_2 \leq \left\|{\mathbf{U}^{(L)}}^\top\mathbf{w}\right\|_2$. It then follows from this fact and (130)-(131) that

$$\tau_{\mathbf{w},\ell} = \mathbf{w}^\top\mathbf{M}\mathbf{w} \leq \frac{4}{n}\left\|{\mathbf{U}^{(L)}}^\top\mathbf{w}\right\|_2^2. \tag{132}$$

It follows from the concentration inequality about quadratic forms of sub-Gaussian random variables in (Wright, 1973) that $\Pr\left[\left\|{\mathbf{U}^{(L)}}^\top\mathbf{w}\right\|_2^2 - \mathbb{E}\left[\left\|{\mathbf{U}^{(L)}}^\top\mathbf{w}\right\|_2^2\right] > m_L\right] \leq \exp(-\Theta(m_L))$. Then with probability at least $1-\exp(-\Theta(m_L))$, we have

$$\left\|{\mathbf{U}^{(L)}}^\top\mathbf{w}\right\|_2^2 \leq \mathbb{E}\left[\left\|{\mathbf{U}^{(L)}}^\top\mathbf{w}\right\|_2^2\right] + m_L \leq \sigma_0^2\mathrm{tr}\left(\mathbf{U}^{(L)}{\mathbf{U}^{(L)}}^\top\right) + m_L = m_L(\sigma_0^2+1). \tag{133}$$

(128) then follows from (132) and (133). ■

Lemma 33 *For every $\delta \in (0,1)$, suppose $\min\{n,m\} > 4m_L \log(4m_L/\delta)$. Then with probability at least $1-\exp(-\Theta(m_L))-\delta$, for every $\ell \in [0:L]$,*

$$|\widehat{\tau}_{\mathbf{w},\ell}| \le 8\sqrt{2}\gamma_0\sqrt{\frac{m_L(\sigma_0^2+1)}{n}}. \tag{134}$$

Proof Recall that the singular value decomposition of $\mathbf{Y}(\mathbf{S},m_L)$ is $\mathbf{Y}(\mathbf{S},m_L) = \mathbf{U}^{(L)}\mathbf{\Sigma}^{(L)}\mathbf{V}^{(L)\top}$ as that in (129). Then we have

$$\begin{aligned}\widehat{\tau}_{\mathbf{w},\ell} &= \frac{2}{\sqrt{n}}\boldsymbol{\beta}^\top \underbrace{(\mathbf{Y}^\top(\mathbf{S},r_0)\mathbf{Y}(\mathbf{S},\ell)/n)(\mathbf{Y}^\top(\mathbf{Q},\ell)\mathbf{Y}(\mathbf{Q},m_L)/m)\mathbf{V}^{(L)}(\mathbf{\Sigma}^{(L)}/\sqrt{n})}_{:=\mathbf{D}_\mathbf{u}}\mathbf{U}^{(L)\top}\mathbf{w}\\ &= \frac{2}{\sqrt{n}}\boldsymbol{\beta}^\top\mathbf{D}_\mathbf{u}\mathbf{w}.\end{aligned}$$

It follows from Lemma 34 that with probability at least $1-\delta$,

$$\|\mathbf{D}_\mathbf{u}\|_2 \le 4\sqrt{2}. \tag{135}$$

Moreover, it follows from (133) in the proof of Lemma 32 that with probability at least $1-\exp(-\Theta(m_L))$,

$$\left\|\mathbf{U}^{(L)\top}\mathbf{w}\right\|_2 \le \sqrt{m_L(\sigma_0^2+1)}. \tag{136}$$

(134) then follows from (135), (136), and the fact that $\|\boldsymbol{\beta}\|_2 \le \gamma_0$. ■

Lemma 34 *With $\tau_\ell = 1$ for all $\ell \in [0:L]$ in the activation function*

$$\sigma_{\boldsymbol{\tau}}(\mathbf{x},\mathbf{x}') = \sum_{\ell=0}^{L}\sum_{j=1}^{N(d,\ell)}\tau_\ell\mu_{\sigma,\ell}Y_{\ell,j}(\mathbf{x})Y_{\ell,j}(\mathbf{x}'),$$

we have $\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}}|\sigma_{\boldsymbol{\tau}}(\mathbf{x},\mathbf{x}')| \le L+1$. Moreover, define

$$\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},m_L)/n - \mathbf{I}_{m_L} := \Delta_{\mathbf{S},m_L}, \tag{137}$$

$$\mathbf{Y}^\top(\mathbf{Q},m_L)\mathbf{Y}(\mathbf{Q},m_L)/m) - \mathbf{I}_{m_L} := \Delta_{\mathbf{Q},m_L}. \tag{138}$$

When $n,m \ge 4m_L\log(4m_L/\delta)$, for every $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\max\{\Delta_{\mathbf{S},m_L},\Delta_{\mathbf{Q},m_L}\} \le 1. \tag{139}$$

Proof First, we note that with $\tau_\ell = 1$ for every $\ell \in [0:L]$,

$$\sigma_{\boldsymbol{\tau}}(\mathbf{x},\mathbf{x}') = \sum_{\ell=0}^{L}P_\ell^{(d)}(\langle\mathbf{x},\mathbf{x}'\rangle) \le L+1,$$

which follows from the fact that $\sup_{t\in[-1,1],k\geq 0}\left|P_k^{(d)}(t)\right| \leq 1$ in Section B of the appendix. Furthermore, it follows from Lemma 35 that with probability at least $1-\delta$ for every $\delta \in (0,1)$,

$$\max\left\{\left\|\mathbf{Y}^\top(\mathbf{S},m_L)\mathbf{Y}(\mathbf{S},m_L)/n - \mathbf{I}_{m_L}\right\|_2, \left\|(\mathbf{Y}^\top(\mathbf{Q},m_L)\mathbf{Y}(\mathbf{Q},m_L)/m) - \mathbf{I}_{m_L}\right\|_2\right\}$$
$$\leq \max\left\{\sqrt{\log\left(\frac{4m_L}{\delta}\right)\frac{4m_L}{n}}, \sqrt{\log\left(\frac{4m_L}{\delta}\right)\frac{4m_L}{m}}\right\} \leq 1,$$

which proves (139). It is noted that we use $\mathbf{S}$ and $\mathbf{Q}$ to replace the sample $\left\{\vec{\mathbf{w}}_r\right\}$ in Lemma 35 to obtain (139). ■

Lemma 35 *Recall that $\{Y_j\}_{j=0}^{m_L-1} = \{Y_{\ell j}\}_{0\leq\ell\leq L, j\in[N(d,\ell)]}$ as the enumeration of all the spherical harmonics of up to degree L. Suppose A, B are two nonempty subsets of $[0:m_L-1]$ containing consecutive integers starting with 0 with $|A|=r_1$, $|B|=r_2$, and $Y_A=\{Y_j : j\in A\}$ and $Y_B=\{Y_j : j\in B\}$. For any vector $\mathbf{w}\in\mathcal{X}$, we define $Y_A(\mathbf{w})\in\mathbb{R}^{r_1}$ as a vector whose elements are $\{Y_j(\mathbf{w}) : j\in A\}$, and $Y_B(\mathbf{w})$ is defined similarly. Let $\left\{\vec{\mathbf{w}}_r\right\}_{r\in[m]} \overset{iid}{\sim} \mathrm{Unif}(\mathcal{X})$. We define $\mathbf{A}^{(r)}\in\mathbb{R}^{r_1}$ with $\mathbf{A}^{(r)} = Y_A(\vec{\mathbf{w}}_r)$ for all $r\in[m]$, and $\mathbf{A}=\left[\mathbf{A}^{(1)},\dots\mathbf{A}^{(m)}\right]\in\mathbb{R}^{r_1\times m}$. Similarly, we define $\mathbf{B}^{(r)}\in\mathbb{R}^{r_2}$ with $\mathbf{B}^{(r)} = Y_B(\vec{\mathbf{w}}_r)$ for all $r\in[m]$, and $\mathbf{B}=\left[\mathbf{B}^{(1)},\dots\mathbf{B}^{(m)}\right]\in\mathbb{R}^{r_2\times m}$. Suppose that $\|Y_A(\mathbf{w})\|_2^2$ and $\|Y_B(\mathbf{w})\|_2^2$ are not varying with $\mathbf{w}$, and $\|Y_A(\mathbf{w})\|_2^2\in[1,m_L]$, $\|Y_B(\mathbf{w})\|_2^2\in[1,m_L]$. Then for every $\delta\in(0,1)$, when $m\geq 4m_L\log(2m_L/\delta)$,*

$$\Pr\left[\left\|\frac{\mathbf{A}\mathbf{B}^\top}{m} - \mathbb{E}\left[\frac{\mathbf{A}\mathbf{B}^\top}{m}\right]\right\|_2 \geq \sqrt{\log\left(\frac{2m_L}{\delta}\right)\frac{4m_L}{m}}\right] \leq \delta. \tag{140}$$

Remark 36 *When Y_A contains spherical harmonics of several degrees, for example, there exists $\ell_1,\ell_2\in[0:L]$ and $\ell_1\leq\ell_2$ such that $Y_A=\{Y_{\ell j}\}_{\ell_1\leq\ell\leq\ell_2, j\in[N(d,\ell)]}$, then it can be verified that $\|Y_A(\mathbf{w})\|_2^2 = \sum_{\ell=\ell_1}^{\ell_2} N(d,\ell)$ which does not vary with $\mathbf{w}\in\mathcal{X}$. The same argument applies to Y_B. Throughout this paper we would apply Lemma 35 for such cases.*

Proof First, we have

$$\frac{\mathbf{A}\mathbf{B}^\top}{m} = \frac{1}{m}\sum_{r=1}^m \mathbf{A}^{(r)}{\mathbf{B}^{(r)}}^\top, \quad \mathbb{E}\left[\mathbf{A}\mathbf{B}^\top/m\right] := \mathbf{E}\in\mathbb{R}^{r_1\times r_2}.$$

Let $A=\{0,1,\dots,r_1-1\}$ and $B=\{0,1,\dots,r_2-1\}$, then it follows from the orthogonality of $\{Y_j\}_{j=0}^{m_L-1}$ that $\mathbf{E}_{st} = \mathbb{1}_{\{s=t\}}\cdot\mathbb{1}_{\{s\leq\min\{r_1,r_2\}\}}$ for all $s\in[r_1]$ and $j\in[r_2]$. It follows that the off-diagonal elements of $\mathbf{E}\mathbf{E}^\top$ and $\mathbf{E}^\top\mathbf{E}$ are 0, and the diagonal elements of $\mathbf{E}\mathbf{E}^\top$ and $\mathbf{E}^\top\mathbf{E}$ are either 0 or 1. We now apply the matrix Bernstein inequality in Theorem 37. We define $\mathbf{X}^{(r)} := \mathbf{A}^{(r)}{\mathbf{B}^{(r)}}^\top - \mathbf{E}\in\mathbb{R}^{r_1\times r_2}$. Then we have $\mathbb{E}\left[\mathbf{X}^{(r)}\right]=0$, and

$$\left\|\mathbf{X}^{(r)}\right\|_2 \leq \left\|\mathbf{A}^{(r)}\right\|_2\left\|\mathbf{B}^{(r)}\right\|_2 + 1 \leq m_L+1, \tag{141}$$

where we use the fact that $\max\left\{\|Y_A(\mathbf{w})\|_2^2, \|Y_B(\mathbf{w})\|_2^2\right\} \le m_L$. Let $V = \left\|\sum_{r=1}^{m} \mathbb{E}\left[\mathbf{X}^{(r)}\mathbf{X}^{(r)\top}\right]\right\|_2$, then we have

$$\begin{aligned} V &\le \sum_{r=1}^{m} \left\|\mathbb{E}\left[\left(\mathbf{A}^{(r)}\mathbf{B}^{(r)\top} - \mathbf{E}\right)\left(\mathbf{A}^{(r)}\mathbf{B}^{(r)\top} - \mathbf{E}\right)^{\top}\right]\right\|_2 \\ &= \sum_{r=1}^{m} \left\|\mathbb{E}\left[\mathbf{A}^{(r)}\mathbf{B}^{(r)\top}\mathbf{B}^{(r)}\mathbf{A}^{(r)\top} - \mathbf{A}^{(r)}\mathbf{B}^{(r)\top}\mathbf{E}^{\top} - \mathbf{E}\mathbf{B}^{(r)}\mathbf{A}^{(r)\top} + \mathbf{E}\mathbf{E}^{\top}\right]\right\|_2 \\ &\overset{①}{=} \sum_{r=1}^{m} \left\|\mathbf{I}_{r_1}\left\|\mathbf{B}^{(r)}\right\|_2^2 - \mathbf{E}\mathbf{E}^{\top}\right\|_2 \overset{②}{\le} m(m_L - 1), \end{aligned} \tag{142}$$

where ① follows from $\mathbb{E}\left[\mathbf{A}^{(r)}\mathbf{A}^{(r)\top}\right] = \mathbf{I}_{r_1}$ due to the orthogonality of the set Y_A. ② follows from the fact that $\left\|\mathbf{B}^{(r)}\right\|_2^2$ is a constant and $1 \le \left\|\mathbf{B}^{(r)}\right\|_2^2 \le m_L$. It can be verified in a way similar to (142) that $\left\|\sum_{r=1}^{m} \mathbb{E}\left[\mathbf{X}^{(r)\top}\mathbf{X}^{(r)}\right]\right\|_2 \le m(m_L - 1)$.

As a result, it follows from the matrix Bernstein inequality in Theorem 37, (141), and (142) that, for every $t \in (0, 1]$,

$$\begin{aligned} \Pr\left[\left\|\frac{\mathbf{A}\mathbf{B}^{\top}}{m} - \mathbb{E}\left[\frac{\mathbf{A}\mathbf{B}^{\top}}{m}\right]\right\|_2 \ge t\right] &\le 2m_L \exp\left(-\frac{m^2t^2}{2m(m_L-1) + 2(m_L+1)mt/3}\right) \\ &\le 2m_L \exp\left(-\frac{mt^2}{4m_L}\right), \end{aligned}$$

which proves (140). ■

Theorem 37 (Matrix Bernstein Inequality, (Tropp, 2015, Theorem 6.1.1)) *Let $\left\{\mathbf{X}^{(r)}\right\}_{i=1}^{n}$ be independent, centered, self-adjoint random matrices in $\mathbb{R}^{d_1 \times d_2}$ such that $\mathbb{E}\left[\mathbf{X}^{(r)}\right] = 0$, $\left\|\mathbf{X}^{(r)}\right\|_2 \le L$ for all $i \in [n]$. Let the total variance be*

$$\sigma^2 := \max\left\{\left\|\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{X}^{(r)}\mathbf{X}^{(r)\top}\right]\right\|_2, \left\|\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{X}^{(r)\top}\mathbf{X}^{(r)}\right]\right\|_2\right\}.$$

Then, for all $t \ge 0$,

$$\Pr\left[\left\|\sum_{i=1}^{n} \mathbf{X}^{(r)}\right\|_2 \ge t\right] \le (d_1 + d_2)\exp\left(\frac{-t^2/2}{\sigma^2 + Lt/3}\right). \tag{143}$$

Appendix E. Existing Empirical and Theoretical Works about Channel Attention and General Attention Mechanism

Channel attention mechanisms (Fu et al., 2019; Wang et al., 2020; Ali et al., 2021) have emerged as an effective method to enhance feature representations learned by DNNs by adaptively reweighting channel responses. DANet (Fu et al., 2019) incorporates a channel attention branch alongside spatial

attention to capture inter-channel relationships, enabling feature refinement for the image segmentation task. Following that, ECA-Net (Wang et al., 2020) introduces a parameter-efficient channel attention module based on the 1D convolution. XCiT (Ali et al., 2021) interprets channel attention as a cross-covariance operation across feature dimensions, and demonstrates its effectiveness for image classification by replacing the self-attention module in the vision transformer. More recently, (Chen et al., 2025) establishes a theoretical framework for covariance-based channel interactions, which is also referred to as covariance pooling, demonstrating that matrix function normalizations, such as logarithm, power, or square-root, applied to Symmetric Positive Definite (SPD) covariance matrices implicitly induce Riemannian classifiers, thereby offering a principled explanation of how second-order channel statistics improve discriminability and enhance the stability of DNNs for image classification.

Building on the same theoretical perspective, (Song et al., 2021) analyzes why approximate matrix square root computations via Newton–Schulz iteration consistently outperform exact singular value decomposition (SVD) in covariance pooling, attributing the superiority of the approximate method to improved numerical stability and gradient smoothness. Furthermore, (Wang et al., 2023) investigates covariance pooling from an optimization perspective, showing that it smooths the loss landscape, yields flatter local minima, and acts as a feature-based preconditioner on gradients, thereby explaining its ability to accelerate convergence, improve robustness, and enhance generalization of deep architectures.

Kernelizable attention has been investigated in (Choromanski et al., 2021; Peng et al., 2021; Zheng et al., 2023) for efficient approximation of attention matrices, and (Hron et al., 2020) analyzes multi-head attention architectures in the Gaussian process limit with infinitely many heads. Although a few works, such as (Kim et al., 2024), study the optimality of attention-based neural networks for in-context learning (ICL) tasks, the theoretical benefits of attention mechanisms, particularly channel attention, for standard nonparametric regression tasks remain largely unexplored.

However, to the bet of our knowledge, most existing works in attention mechanisms, including channel attention, do not give sharp rates for nonparametric regression with target function being low-degree spherical polynomials. Our work is among the first to reveal the theoretical benefit of channel attention with a novel and provable learnable channel selection algorithm for learning low-degree spherical polynomials with a minimax optimal rate.