Journal of Machine Learning Research 27 (2026) 1-64 Submitted 8/25; Revised 10/26; Published 10/26

Sharp Generalization for Nonparametric Regression in Interpolation Space by Over-Parameterized Neural Networks Trained with Preconditioned Gradient Descent and Early Stopping

Yingzhen Yang YINGZHEN.YANG@ASU.EDU
School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ 85281, USA

Ping Li PINGLI98@GMAIL.COM
VecML Inc.
Bellevue, WA 98004, USA

Editor: Brian Kulis

Abstract

We study nonparametric regression using an over-parameterized two-layer neural network trained with algorithmic guarantees in this paper. We consider the setting where the training features are drawn uniformly from the unit sphere in $\mathbb{R}^d$, and the target function lies in an interpolation space commonly studied in statistical learning theory. We demonstrate that training the neural network with a novel Preconditioned Gradient Descent (PGD) algorithm, equipped with early stopping, achieves a sharp regression rate of $\mathcal{O}(n^{-\frac{2\alpha s'}{2\alpha s'+1}})$, where α parametrizes the polynomial eigenvalue decay of the regular Neural Tangent Kernel (NTK), $\lambda_j \asymp j^{-2\alpha}$, and $2\alpha = d/(d-1)$ under the spherical uniform design considered here. The target function is assumed to belong to the interpolation space $[\mathcal{H}_K]^{s'}$ with $s' \geq 3$. This rate is even sharper than the currently known nearly-optimal rate of $\mathcal{O}(n^{-\frac{2\alpha s'}{2\alpha s'+1}})\log^2(1/\delta)$ (Li et al., 2024), where n is the size of the training data and $\delta \in (0,1)$ is a small probability. This rate is also sharper than the standard kernel regression rate of $\mathcal{O}(n^{-\frac{2\alpha}{2\alpha+1}})$ obtained under the regular Neural Tangent Kernel (NTK) regime when training the neural network with the vanilla gradient descent (GD), where $2\alpha = d/(d-1)$. Our analysis is based on two key technical contributions. First, we present a principled decomposition of the network output at each PGD step into a function in the reproducing kernel Hilbert space (RKHS) of a newly induced integral kernel, and a residual function with small L^∞-norm. Second, leveraging this decomposition, we apply local Rademacher complexity theory to tightly control the complexity of the function class comprising all the neural network functions obtained in the PGD iterates. Our results show how PGD changes the effective kernel in the reference training dynamics, inducing an integral kernel with lower complexity than the regular NTK arising from vanilla GD.

Keywords: Nonparametric Regression, Over-Parameterized Neural Network, Preconditioned Gradient Descent, Minimax Optimal Rate

©2026 Yingzhen Yang and Ping Li.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-2066.html.

1. Introduction

The remarkable success of deep learning across numerous domains (LeCun et al., 2015) has spurred intense interest in understanding the generalization properties of neural networks within statistical and deep learning theory. A substantial body of work has addressed the optimization dynamics of deep neural networks (DNNs), establishing that gradient descent (GD) and stochastic gradient descent (SGD) can achieve vanishing training loss under various settings (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). Parallel efforts have focused on generalization theory, aiming to provide algorithmic guarantees, that is, bounds for the generalization error of neural networks trained with gradient-based or other optimization methods. A prominent line of research has shown that in the over-parameterized regime, where hidden layers have a sufficiently large number of neurons, the training dynamics of DNNs closely resemble those of kernel methods with the Neural Tangent Kernel (NTK) (Jacot et al., 2018). Specifically, the NTK induced by the architecture of the network characterizes this behavior. The core insight behind NTK-based analysis is that, for highly over-parameterized networks, weights remain close to their random initialization throughout training. This enables first-order Taylor expansion around the initialization point, allowing for tractable approximation and generalization analysis (Cao and Gu, 2019; Arora et al., 2019; Ghorbani et al., 2021). In addition, certain infinite-width networks can learn features (Yang and Hu, 2021).

While NTK captures important aspects of learning in wide networks, it operates within a linearized regime of the NTK, limiting its capacity to capture feature learning. Existing works make efforts to learn features beyond NTK's linear approximation. For instance, the QuadNTK (Bai and Lee, 2020) employs higher-order approximations for improved generalization. Other works, such as Nichani et al. (2022), combine NTK and QuadNTK to learn structured polynomial targets, while Damian et al. (2022) and Takakura and Suzuki (2024) explore optimization procedures and mean-field analyses to capture the feature learning effects in two-layer neural networks.

Although much of the generalization literature focuses on clean data, a fundamental question in statistical learning concerns how neural networks perform in the presence of noise, particularly for nonparametric regression. Prior work has demonstrated that various DNN architectures achieve minimax optimal rates for both smooth (Yarotsky, 2017; Bauer and Kohler, 2019; Schmidt-Hieber, 2020; Jiao et al., 2023; Zhang and Wang, 2023) and non-smooth (Imaizumi and Fukumizu, 2019) target functions. However, these results often lack algorithmic guarantees, that is, the network constructions are not generally attainable through standard optimization procedures such as GD. Furthermore, these constructions may assume sparse connectivity structures that are not representative of contemporary over-parameterized architectures. Studying the generalization of neural networks with algorithmic guarantees is particularly important to understand the behavior of neural networks trained in practice, revealed by the efforts in this direction such as Hu et al. (2021); Suh et al. (2022); Li et al. (2024); Yang and Li (2024); Yang (2025).

In this work, we study nonparametric regression using an over-parameterized two-layer neural network with algorithmic guarantees. When the target function resides in an interpolation space characterized by a spectral bias, and the distribution of the covariate admits a polynomial eigenvalue decay rate, we show that early-stopped training by a novel

Preconditioned Gradient Descent (PGD) algorithm achieves the minimax optimal risk rate of $\mathcal{O}(n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}) = \mathcal{O}(n^{-\frac{d(s+2)}{d(s+2)+d-1}})$ for $2\alpha = d/(d-1)$, where $s \geq 1$ is specified by the preconditioner in PGD. This result improves upon existing bounds, as summarized in Section 3. Every function in the interpolation space exhibits a spectral bias, that is, its energy is predominantly concentrated on the leading eigenfunctions of the NTK of the neural network, with low-degree polynomials as examples of such functions. The interpolation space is widely studied in the statistics literature, and its formal definition is deferred to Section 2.2. Although GD-based methods with preconditioning have been used in other over-parameterized settings, such as low-rank matrix sensing (Xu et al., 2023) and non-convex factorization (Zhang et al., 2023), this paper is among the first works to design a PGD method achieving sharp minimax optimal rates for nonparametric regression using over-parameterized neural networks. Furthermore, PGD induces a novel kernel, termed the integral kernel, rather than the well-studied usual NTK induced by the vanilla GD. The integral kernel has lower kernel complexity than the vanilla NTK, and our rate of $\mathcal{O}(n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}})$ is in fact the minimax optimal rate of kernel regression with the integral kernel over the interpolation space.

At a high level, PGD reshapes the training dynamics so that the neural network behaves like kernel regression with a different, smoother kernel than the regular NTK induced by vanilla GD. In particular, PGD induces an integral kernel $K^{(\mathrm{int})}$ whose eigenvalues decay faster than those of the regular NTK. This stronger spectral decay reduces the associated kernel complexity and enables a sharper regression rate when the target function has spectral bias. Our result provides an algorithmic route to a sharp rate under a structured target class, and it clarifies how an optimization design changes the reference kernel in the linearized training dynamics. The target class we study is exactly the interpolation space $[\mathcal{H}_K]^{s'}$, which is widely used in statistical learning theory, and in our setting it coincides with the RKHS associated with the integral kernel $K^{(\mathrm{int})}$. To the best of our knowledge, this is among the first results showing that a carefully designed PGD method yields such a sharp rate for nonparametric regression with an over-parameterized neural network under this setting.

We organize this paper as follows. The rest of this section introduces the necessary notation. Section 2 presents the problem setup, including the definition of the interpolation space. Section 3 summarizes the main results. Section 4 introduces GD and PGD for the over-parameterized two-layer neural network, and Section 5 states the main results in detail. Section 6 presents the proof roadmap, the key technical results, the differences from existing kernel learning theory, and the novel proof strategy. Section 7 discusses the extension of PGD to general multi-layer neural networks. Section 8 reports the simulations, and Section 9 concludes the paper.

Notations. We use bold letters for matrices and vectors, and regular lowercase letters for scalars throughout this paper. The bold letter with a single superscript indicates the corresponding column of a matrix, e.g., $\mathbf{A}^{(i)}$ is the i-th column of matrix $\mathbf{A}$, and bold letters with subscripts indicate 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. $\|\cdot\|_{\mathrm{F}}$ 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 $L^2(\mathbb{S}^{d-1}, \mu)$ denote the space of square-integrable functions on $\mathbb{S}^{d-1}$ with probability measure μ, and the inner product $\langle \cdot, \cdot \rangle_{L^2}$ and $\|\cdot\|_{L^2}^2$ are defined as $\langle f, g \rangle_{L^2} \coloneqq \int_{\mathbb{S}^{d-1}} f(x)g(x)\mathrm{d}\mu(x)$ and $\|f\|_{L^2}^2 \coloneqq \int_{\mathbb{S}^{d-1}} f^2(x)\mathrm{d}\mu(x) < \infty$. $\mathbf{B}(\mathbf{x}; r)$ is the Euclidean closed ball centered at $\mathbf{x}$ with radius r. Given a function $g\colon \mathbb{S}^{d-1} \to \mathbb{R}$, its L^∞-norm is denoted by $\|g\|_\infty \coloneqq \sup_{\mathbf{x}\in\mathbb{S}^{d-1}} |g(\mathbf{x})|$. L^∞ is the function class whose elements have almost surely 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 exist constants $c_1, c_2 > 0$ such that $c_1 b \leq a \leq c_2 b$. Throughout this paper we let the input space $\mathcal{X} = \mathbb{S}^{d-1}$, and $\mathrm{Unif}(\mathcal{X})$ denotes the uniform distribution on $\mathcal{X}$. The constants defined throughout this paper may change from line to line. For a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$, $\mathcal{H}(\mu_0)$ denotes the ball centered at the origin with radius μ_0 in $\mathcal{H}$. We use $\mathbb{E}_P[\cdot]$ to denote the expectation with respect to the distribution P.

2. Problem Setup

We first introduce the problem setup for nonparametric regression with a two-layer neural network where the target function belongs to an interpolation space with spectral bias.

2.1 Two-Layer Neural Network

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 a feature vector $\vec{\mathbf{x}}_i \in \mathcal{X}$ and its response $y_i \in \mathbb{R}$. 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 $\mathbf{S}$. All the responses are stacked as a vector $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$. 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 representing the noise with mean 0 and variance proxy $\sigma_0^2 > 0$, 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 $\mathbf{w} \coloneqq [w_1, \ldots, w_n]^\top$, 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 responses. The noise is independent of the covariates and the algorithmic randomness. The feature vectors in $\mathbf{S}$ are drawn i.i.d. according to a distribution P with probability measure μ. We assume $\mathrm{supp}(\mu) = \mathcal{X}$, so the continuous Mercer eigenfunctions and RKHS functions are identified pointwise on $\mathcal{X}$. This population support condition permits atoms and coincident training features.

We consider a two-layer Neural Network (NN) with the ReLU activation $\sigma\colon \mathbb{R} \to \mathbb{R}$, $\sigma(z) = \max\{0, z\}$. Its mapping function is

$$f(\mathbf{W}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma\left(\vec{\mathbf{w}}_r^{\top} \mathbf{x}\right), \tag{1}$$

where $\mathbf{x} \in \mathcal{X}$ is the input, m is the number of neurons, $\mathbf{W} = \left\{\vec{\mathbf{w}}_r\right\}_{r=1}^{m}$ with $\vec{\mathbf{w}}_r \in \mathbb{R}^d$ for $r \in [m]$ denotes the weight vectors in the first layer. $\boldsymbol{a} = [a_1, \ldots, a_m] \in \mathbb{R}^m$ denotes the weights of the second layer. Our goal is to establish a sharp convergence rate of $\mathbb{E}_P\left[(f(\mathbf{W}, \cdot) - f^*)^2\right]$ with $f(\mathbf{W}, \cdot)$ trained from the training data. We also write $\mathbf{W}$ as $\mathbf{W_S}$ to indicate that the weight vectors, or the weights $\mathbf{W}$, are trained on the training features $\mathbf{S}$.

2.2 Kernel and Target Function

The statistical rate we can achieve depends on two ingredients: the spectrum of the kernel induced by training and the regularity (or spectral bias) of the target function. We first introduce the standard kernel K associated with the two-layer network under vanilla GD, and then define the integral kernel $K^{(\mathrm{int})}$ that will arise from PGD. The key difference is spectral: $K^{(\mathrm{int})}$ shares the same eigenfunctions as K but has faster-decaying eigenvalues, which leads to lower kernel complexity under our target-class assumption.

We define the kernel function

$$K(\mathbf{u}, \mathbf{v}) \coloneqq \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{2\pi} \left(\pi - \arccos \langle \mathbf{u}, \mathbf{v} \rangle\right), \quad \forall\, \mathbf{u}, \mathbf{v} \in \mathcal{X}, \tag{2}$$

which is in fact the NTK associated with the two-layer NN (1) (Du et al., 2019b) when trained with the vanilla GD with fixed $\boldsymbol{a}$. Let the gram matrix of K over the training features $\mathbf{S}$ be $\mathbf{K} \in \mathbb{R}^{n \times n}, \mathbf{K}_{ij} = K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j)$ for $i, j \in [n]$, and $\mathbf{K}_n \coloneqq \mathbf{K}/n$. Let the eigen-decomposition of $\mathbf{K}_n$ be $\mathbf{K}_n = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^{\top}$ where $\mathbf{U}$ is a $n \times n$ 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 can be verified that $\widehat{\lambda}_1 \in (0, 1/2]$. Let $\mathcal{H}_K$ be the RKHS associated with K. Because K is continuous on the compact set $\mathcal{X} \times \mathcal{X}$, 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 $\overline{\operatorname{range}(T_K)} \subseteq L^2(\mathcal{X}, \mu)$. By the spectral theorem, there is a countable orthonormal basis $\{e_j\}_{j \geq 1}$ of $\overline{\operatorname{range}(T_K)} = (\ker T_K)^{\perp} \subseteq L^2(\mathcal{X}, \mu)$ and a sequence of positive eigenvalues $\{\lambda_j\}_{j \geq 1}$ with $\frac{1}{2} \geq \lambda_1 \geq \lambda_2 \geq \ldots > 0$ such that e_j is the eigenfunction of T_K with λ_j being the corresponding eigenvalue. That is, $T_K e_j = \lambda_j e_j, j \geq 1$. Let $\{\mu_\ell\}_{\ell \geq 1}$ be the distinct eigenvalues associated with T_K, and let m_ℓ be the sum of the multiplicities of the eigenvalues $\{\mu_{\ell'}\}_{\ell'=1}^{\ell}$, so that $m_\ell - m_{\ell-1}$ is the multiplicity of μ_ℓ with $m_0 = 0$. It is well known that $\left\{v_j = \sqrt{\lambda_j} e_j\right\}_{j \geq 1}$ is an orthonormal basis of $\mathcal{H}_K$. For a positive constant μ_0, we define $\mathcal{H}_K(\mu_0) \coloneqq \{f \in \mathcal{H}_K \colon \|f\|_{\mathcal{H}} \leq \mu_0\}$ as the closed ball in $\mathcal{H}_K$ centered at the origin with radius μ_0. We note that $\mathcal{H}_K(\mu_0)$ is also specified by $\mathcal{H}_K(\mu_0) = \left\{f \in L^2(\mathcal{X}, \mu) \colon f = \sum_{j=1}^{\infty} \beta_j e_j, \sum_{j=1}^{\infty} \beta_j^2/\lambda_j \leq \mu_0^2\right\}$.

Target Function in an Interpolation Space with Spectral Bias. Extensive theoretical and empirical studies find that it is easy for neural networks to learn spectrally biased target functions or low-frequency information in the training data (Rahaman et al., 2019;

Arora et al., 2019; Cao et al., 2021; Choraria et al., 2022). For example, the studies in Arora et al. (2019); Cao et al. (2021) reveal that it is easier for over-parameterized neural networks to learn target functions with spectral bias, for example, polynomials of low-degree with spherical uniform data distribution on $\mathcal{X}$, or the low-rank part of the ground truth training class labels, or simple patterns of low-frequency. This observation motivates us to restrict the target function f^* to a smaller class than $\mathcal{H}_K(\mu_0)$, which is $\mathcal{H}_{K^{(\mathrm{int})}}(\mu_0)$ to be studied in this paper.

We then define the integral kernel $K^{(\mathrm{int})}$, and explain why functions in $\mathcal{H}_{K^{(\mathrm{int})}}(\mu_0)$ have stronger spectral bias than functions in $\mathcal{H}_K(\mu_0)$. Given the kernel K, we first define another kernel $K^{(s)}$ with $s \geq 1$ by

$$K^{(s)}(\mathbf{x}, \mathbf{x}') \coloneqq \sum_{j \geq 1} \lambda_j^s e_j(\mathbf{x}) e_j(\mathbf{x}'), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{3}$$

It can be verified that $K^{(s)}$ is positive-definite (PD), and it is shown in Theorem C.1 deferred to Section C of the appendix that $K^{(s)}$ is well-defined since the RHS of (3) converges uniformly on $\mathcal{X} \times \mathcal{X}$ with $s \geq 1$, and $K^{(s)}$ is bounded with $\sup_{\mathbf{x},\mathbf{x}' \in \mathcal{X}} \left|K^{(s)}(\mathbf{x}, \mathbf{x}')\right| \leq 1/2$. We define $\mathbf{K}^{(s)} \in \mathbb{R}^{N \times N}$ as the kernel gram matrix of $K^{(s)}$ specified by $\mathbf{K}^{(s)}_{ij} = K^{(s)}(\vec{\mathbf{q}}_i, \vec{\mathbf{q}}_j)$ for all $i, j \in [N]$, and $\mathbf{K}^{(s)}_N = \mathbf{K}^{(s)}/N$. We then define the integral kernel $K^{(\mathrm{int})}$ as well as its empirical version of $\widehat{K}^{(\mathrm{int})}$ as

$$K^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') \coloneqq \int_{\mathcal{X} \times \mathcal{X}} K(\mathbf{x}, \mathbf{v}) K^{(s)}(\mathbf{v}, \mathbf{v}') K(\mathbf{v}', \mathbf{x}') \mathrm{d}\mu(\mathbf{v}) \otimes \mu(\mathbf{v}'), \tag{4}$$

$$\widehat{K}^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') \coloneqq \frac{1}{N^2} \sum_{i,j=1}^{N} K(\mathbf{x}, \vec{\mathbf{q}}_i) K^{(s)}(\vec{\mathbf{q}}_i, \vec{\mathbf{q}}_j) K(\vec{\mathbf{q}}_j, \mathbf{x}'). \tag{5}$$

An i.i.d. sample $\mathbf{Q} = \left\{\vec{\mathbf{q}}_i\right\}_{i=1}^N$ such that $\vec{\mathbf{q}}_i \sim P$ for all $i \in [N]$ is used in (5). The auxiliary sample is independent of the training sample, the noise, and the initialization. For the spherical uniform design, $P = \mathrm{Unif}\,(\mathcal{X})$.

Intuition for the integral kernel $K^{(\mathrm{int})}$. The integral kernel $K^{(\mathrm{int})}$ can be viewed as a spectrally reweighted version of the regular NTK K. It preserves the eigenfunctions of K but transforms the eigenvalues from λ_j to λ_j^{s+2} (Theorem C.2). Since $\lambda_j \in (0, 1)$, this transformation suppresses higher-frequency directions more strongly, yielding a smoother kernel and a smaller associated kernel complexity. This is the spectral reason PGD can achieve a sharper rate when the target function has spectral bias.

In particular, Theorem C.2 in Appendix C shows that the integral operator associated with $K^{(\mathrm{int})}$, $T_{K^{(\mathrm{int})}}$, has the same eigenfunctions $\{e_j\}_{j \geq 1}$ as T_K, and the eigenvalue corresponding to e_j is $\lambda_j^{(\mathrm{int})} = \lambda_j^{s+2} \in (0, 1/8]$ for all $j \geq 1$ and $s \geq 1$. Because $K^{(\mathrm{int})}$ is still a PD kernel, the RKHS associated with $K^{(\mathrm{int})}$ is well-defined, and $\mathcal{H}_{K^{(\mathrm{int})}}(\mu_0)$ indicates a subset of $\mathcal{H}_{K^{(\mathrm{int})}}$ with the RKHS-norm $\|\cdot\|_{K^{(\mathrm{int})}}$ bounded by μ_0. It can be verified that $\mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) = \{f \in L^2(\mathcal{X}, \mu) \colon f = \sum_{j=1}^{\infty} \beta_j e_j, \sum_{j=1}^{\infty} \beta_j^2 / \lambda_j^{s+2} \leq \mu_0^2\}$. It is remarked that $\mathcal{H}_{K^{(\mathrm{int})}}(\mu_0)$ is also the interpolation space $[\mathcal{H}_K]^{s'}$ (with $s' \coloneqq s + 2$) of the RKHS, which characterizes the regularity of the regression function studied in Steinwart and Scovel (2012); Fischer and Steinwart (2020). The interpolation space for general

$s' > 0$ is defined as $[\mathcal{H}_K]^{s'}(\mu_0) := \left\{\sum_{j\geq 1} a_j \lambda_j^{s'/2} e_j \colon \sum_{j\geq 1} a_j^2 \leq \mu_0^2\right\}$. The RKHS associated with $K^{(\mathrm{int})}$ is exactly the interpolation space of $\mathcal{H}_K$ with exponent $s' = s+2$, that is, $\mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) = [\mathcal{H}_K]^{s'}(\mu_0)$. As $\lambda_j \to 0$ with $j \to \infty$ and $\lambda_j^{s+2} < \lambda_j < 1$, we have $\mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) \subseteq \mathcal{H}_K(\mu_0)$. Compared to a function in $\mathcal{H}_K(\mu_0)$, a function in $[\mathcal{H}_K]^{s'}(\mu_0)$ has the expansion coefficients $\{\beta_j\}_{j\geq 1}$ which concentrate more on the leading eigenfunctions with smaller index j. In this sense, we say that the target function $f^* \in \mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) = [\mathcal{H}_K]^{s'}(\mu_0)$ has stronger spectral bias than a function in $\mathcal{H}_K(\mu_0)$. Intuitively, this assumption means the expansion coefficients of f^* in the eigenbasis of K decay faster, concentrating more energy on low-frequency or leading eigenfunctions. In summary, *if* $f^* \in [\mathcal{H}_K]^{s'}(\mu_0) \subseteq \mathcal{H}_K(\mu_0)$, *then we say that* f^* *is a target function in an interpolation space with spectral bias.* The motivation for introducing the interpolation space in Steinwart and Scovel (2012); Fischer and Steinwart (2020) is to quantify how strongly a regression function aligns with leading kernel eigenfunctions, enabling sharper statistical learning rates and connecting RKHS learning theory with classical smoothness spaces such as Sobolev spaces. We note that such target functions with spectral bias have been widely studied in the deep learning literature, and a representative example is given below.

An Example of Functions with Spectral Bias. Under the spherical uniform design, a spherical polynomial with no component in $\ker T_K$ has a finite expansion in positive-eigenvalue spherical harmonics. Ghorbani et al. (2021); Bai and Lee (2020); Cao et al. (2021) study learning low-degree polynomials by linearization or higher-order approximation of neural networks. For a polynomial in the positive spectral subspace, there is a finite index J such that

$$f = \sum_{j=1}^{J} \beta_j e_j.$$

When $\sum_{j=1}^{J} \beta_j^2/\lambda_j^{s'} \leq \mu_0^2$, we have $f \in [\mathcal{H}_K]^{s'}(\mu_0)$.

The Task of Nonparametric Regression. Our goal is to construct an estimator $\widehat{f}$ from the training data $\left\{(\vec{\mathbf{x}}_i, y_i)\right\}_{i=1}^{n}$ such that the risk $\mathbb{E}_P\left[(\widehat{f} - f^*)^2\right]$ converges to 0 at a fast rate, under the target-class assumption $f^* \in [\mathcal{H}_K]^{s'}(\mu_0)$. In this paper, the estimator $\widehat{f}$ is the two-layer neural network (1) trained by PGD with early stopping.

3. Summary of Main Results

This section summarizes our main result and places it in the context of prior work on nonparametric regression with over-parameterized neural networks and algorithmic guarantees. The key result is that, under a covariate distribution that admits a certain polynomial eigenvalue decay rate (EDR) and an interpolation-space target assumption $f^* \in [\mathcal{H}_K]^{s'}(\mu_0)$ with $s' \geq 3$, PGD with early stopping achieves the sharp rate $\mathcal{O}(n^{-\frac{2\alpha s'}{2\alpha s'+1}})$ by inducing an integral kernel $K^{(\mathrm{int})}$ with faster eigenvalue decay than the regular NTK.

It is remarked that all the results and discussions in this paper are for the setting with fixed dimension d, which is a widely adopted setting in the existing works (Hu et al., 2021;

Algorithm 1 Training the Two-Layer NN by PGD

1: $\mathbf{W}(T) \leftarrow$ Training-by-PGD$(T, \mathbf{W}(0), \mathbf{M})$
2: **input:** $T, \mathbf{W}(0), \mathbf{M}$
3: **for** $t = 1, \ldots, T$ **do**
4: Perform the t-th step of PGD by (8)
5: **end for**
6: **return** $\mathbf{W}(T)$

Table 1: Comparisons with existing results on risk rates and assumptions for nonparametric regression by training over-parameterized neural networks with algorithmic guarantees. The target-function assumptions are stated separately because the rates are class-dependent. The responses $\{y_i\}_{i=1}^n$ are corrupted by i.i.d. Gaussian or sub-Gaussian noise. Here P is the distribution of the training features and, for the RKHS-based results, $\tilde{K}$ is the kernel induced by the neural architecture and optimization method in the corresponding work. It is the regular NTK for all such works except ours, for which $\tilde{K} = K^{(\mathrm{int})}$. We use $s' = s + 2$ with $s \geq 1$ and $\alpha = d/(2(d-1))$. The notation for the target class in Li et al. (2024, Proposition 13) is adapted to ours. Its displayed rate is expressed under matching source and spectral parameters. The original result uses a different, biased NTK with EDR $\lambda_j \asymp j^{-(d+1)/d}$. The row substitutes our EDR $\lambda_j \asymp j^{-d/(d-1)}$ for this exponent comparison and does not assert that the original theorem applies unchanged to our architecture. The rates of that result and ours hold with probability at least $1 - \Theta(1/n)$.

Existing Works and Our Result	Distributional Assumptions	Target-Function Assumption	Eigenvalue Decay Rate (EDR)	Rate of Nonparametric Regression Risk
(Hu et al., 2021, Theorem 5.2), (Suh et al., 2022, Theorem 3.11)	P is Unif $(\mathcal{X})$	$f^* \in \mathcal{H}_{\tilde{K}}(\mu_0)$	$\lambda_j \asymp j^{-d/(d-1)}$	$\mathcal{O}(n^{-d/(2d-1)})$
(Li et al., 2024, Proposition 13)	P is sub-Gaussian	$f^* \in [\mathcal{H}_K]^{s'}(\mu_0) \cap L^\infty$ for $s' > 1/(d+1)$	$\lambda_j \asymp j^{-d/(d-1)}$	$\mathcal{O}(n^{-ds'/(ds'+d-1)}) \log^2(n)$
(Yang, 2025, Corollary 5.2)	P supported on $\mathcal{X}$ admits the polynomial EDR $\lambda_j \asymp j^{-d/(d-1)}$	$f^* \in \mathcal{H}_{\tilde{K}}(\mu_0)$	$\lambda_j \asymp j^{-d/(d-1)}$	$\mathcal{O}(n^{-d/(2d-1)})$
Our Result (Theorem 5.1)	P has full support on $\mathcal{X}$ and admits the polynomial EDR $\lambda_j \asymp j^{-d/(d-1)}$	$f^* \in \mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) = [\mathcal{H}_K]^{s'}(\mu_0)$ for $s' = s + 2 \geq 3$	$\lambda_j^{(\mathrm{int})} \asymp j^{-d(s+2)/(d-1)}$	$\mathcal{O}(n^{-2\alpha s'/(2\alpha s'+1)}) = \mathcal{O}(n^{-ds'/(ds'+d-1)})$
Our Result (Corollary 5.4)	$p(\mathbf{x}) \lesssim (1 + \|\mathbf{x}\|_2^2)^{-(d+1)/2}$ for every $\mathbf{x} \in \mathbb{R}^d$	$f^* \in \mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) = [\mathcal{H}_K]^{s'}(\mu_0)$ for $s' = s + 2 \geq 3$	$\lambda_j^{(\mathrm{int})} \asymp j^{-d(s+2)/(d-1)}$	$\mathcal{O}(n^{-2\alpha s'/(2\alpha s'+1)}) = \mathcal{O}(n^{-ds'/(ds'+d-1)})$

Suh et al., 2022; Li et al., 2024; Yang and Li, 2024; Yang, 2025), and we consider $d \geq 5$ in this paper. The main result of this paper is presented in Theorem 5.1 of Section 5.2, which shows that training the two-layer NN (1) by a new Preconditioned Gradient Descent (PGD) method with early stopping, described by Algorithm 1, yields a sharper rate for nonparametric regression risk compared to existing results, under the assumption that the target function lies in the interpolation spaces $[\mathcal{H}_K]^{s'}$ widely analyzed in the statistical learning literature (Steinwart and Scovel, 2012; Fischer and Steinwart, 2020). The comparison with relevant existing works is summarized in Table 1. Prior works such as Raskutti et al. (2014) demonstrate that kernel regression with the kernel $\tilde{K}$, which is the NTK of the neural network considered in each specific work in Table 1 except for this paper, achieves the minimax optimal risk rate of $\mathcal{O}(\varepsilon^2_{\tilde{K},n})$, where $\varepsilon^2_{\tilde{K},n}$ denotes the kernel complexity of $\tilde{K}$ formalized in Section 5.1. When the eigenvalues of $\tilde{K}$ exhibit a polynomial EDR $\lambda_j \asymp j^{-2\alpha}$ for $\alpha > 1/2$, it is well known that $\varepsilon^2_{\tilde{K},n} \asymp n^{-\frac{2\alpha}{2\alpha+1}}$ (Raskutti et al., 2014, Corollary 3). More-

over, for spherical uniform input distributions, Hu et al. (2021, Lemma 3.1) shows that $2\alpha = d/(d-1)$, leading to the minimax rate $\mathcal{O}(n^{-\frac{d}{2d-1}})$. Consequently, nearly all existing works in Table 1 excluding (Li et al., 2024) report the same minimax rate of $\mathcal{O}(n^{-\frac{d}{2d-1}})$ under the spherical uniformly distributed training features when the target function lies in a bounded ball of $\mathcal{H}_{\tilde{K}}$. Such a rate arises because the EDR of the kernel $\tilde{K}$ satisfies $\Theta(j^{-2\alpha}) = \Theta(j^{-\frac{d}{d-1}})$, and the vanilla GD is used during training. Li et al. (2024) improves upon the prior results by obtaining a nearly optimal rate $\mathcal{O}(n^{-\frac{2\alpha s'}{2\alpha s'+1}})\log^2(1/\delta)$ when the target function belongs to the interpolation space $[\mathcal{H}_K]^{s'}(R)$ for $s' > 1/(d+1)$ and $R > 0$, which holds with probability at least $1-\delta$. However, this rate is less sharp than the rate derived in this paper, $\mathcal{O}(n^{-\frac{2\alpha s'}{2\alpha s'+1}}) = \mathcal{O}(n^{-\frac{ds'}{ds'+d-1}})$, for all $s' \geq 3$. In particular, when our PGD is employed to train the two-layer NN (1), the induced kernel is the integral kernel $K^{(\mathrm{int})}$ instead of the regular NTK K (2). Under certain mild conditions on the covariate distribution guaranteeing the polynomial EDR and the source condition that the target function $f^* \in [\mathcal{H}_K]^{s'}(\mu_0)$, the rate derived in this paper is sharper than the prior risk rates, including those in Hu et al. (2021); Suh et al. (2022); Li et al. (2024); Yang and Li (2024); Yang (2025).

While the rate of $\mathcal{O}(n^{-\frac{d}{2d-1}})$ in Hu et al. (2021); Suh et al. (2022); Yang and Li (2024) remains minimax optimal in the context of kernel regression with the regular NTK $\tilde{K}$ when $f^* \in \mathcal{H}_{\tilde{K}}(\mu_0)$, a faster rate is achievable if the target function lies in a stricter space. In particular, if $f^* \in \mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) = [\mathcal{H}_K]^{s'}(\mu_0) \subseteq \mathcal{H}_K(\mu_0)$, then kernel regression using the integral kernel $K^{(\mathrm{int})}$ attains the sharper rate $\mathcal{O}(n^{-\frac{ds'}{ds'+d-1}})$, which, as in Theorem 5.1, is minimax optimal in the sense of regression by the kernel $K^{(\mathrm{int})}$ over the space $\mathcal{H}_{K^{(\mathrm{int})}}(\mu_0)$ (Stone, 1985; Yang and Barron, 1999; Yuan and Zhou, 2016). In fact, Theorem C.2 shows that $K^{(\mathrm{int})}$ has the EDR of $\lambda_j^{(\mathrm{int})} = \lambda_j^{s'} \asymp j^{-2\alpha s'}$ with $s' = s+2$ for all $s \geq 1$ and $j \geq 1$. Accordingly, the associated fixed point of the kernel complexity function associated with the integral kernel $K^{(\mathrm{int})}$ is $\mathcal{O}(n^{-\frac{ds'}{ds'+d-1}})$, yielding the claimed sharper risk rate of $\mathcal{O}(n^{-\frac{ds'}{ds'+d-1}})$.

Our findings show that PGD changes the effective kernel in the reference training dynamics from the regular NTK K to $K^{(\mathrm{int})}$. Under the common source and spectral conditions, the statistical improvement over the displayed vanilla-GD bound in Li et al. (2024, Proposition 13) is removal of the logarithmic factor. The polynomial exponent is unchanged. The faster exponent relative to ordinary-RKHS results reflects the smaller target class. This comparison does not contradict a minimax lower bound or establish that logarithmic losses are unavoidable for regular-NTK methods. See Section 6.4.

4. Training by Gradient Descent and Preconditioned Gradient Descent

We next describe the training procedures used in this paper and highlight the key difference between vanilla GD and the proposed PGD. Both methods optimize the same empirical quadratic loss on the training data. The distinction lies in the preconditioner introduced by PGD which changes the training dynamics of the vanilla GD so as to induce the integral kernel.

In the training process, only $\mathbf{W}$ is optimized with $\boldsymbol{a}$ randomly initialized to ± 1 and then fixed. The following quadratic loss function is minimized during the training process:

Algorithm 2 Algorithm that generates the preconditioner $\mathbf{M} \in \mathbb{R}^{md\times md}$

$\mathbf{M} \in \mathbb{R}^{md\times md} \leftarrow$ Generate-Precondition-Matrix(N)
input: $N, \mathbf{W}(0)$
Generate an i.i.d. sample $\mathbf{Q} = \left\{\vec{\mathbf{q}}_i\right\}_{i=1}^{N}$ such that $\vec{\mathbf{q}}_i \sim P$ for all $i \in [N]$.
Compute the matrix $\mathbf{Z}_{\mathbf{Q}}(0) \in \mathbb{R}^{md\times N}$ according to (9) with $\mathbf{S}$ replaced by $\mathbf{Q}$, using the randomly initialized weight vectors $\mathbf{W}(0) = \left\{\vec{\mathbf{w}}_r(0)\right\}_{r=1}^{m}$.
Compute the preconditioner $\mathbf{M} = \frac{1}{N}\mathbf{Z}_{\mathbf{Q}}(0)\mathbf{K}_N^{(s)}\mathbf{Z}_{\mathbf{Q}}(0)^\top \in \mathbb{R}^{md\times md}$.
return M

$$L(\mathbf{W}) \coloneqq \frac{1}{2n}\sum_{i=1}^{n}\left(f(\mathbf{W}, \vec{\mathbf{x}}_i) - y_i\right)^2. \tag{6}$$

In the $(t+1)$-th step of GD with $t \geq 0$, the weights of the neural network, $\mathbf{W}_{\mathbf{S}}$, are updated by one step of GD through

$$\mathrm{vec}\left(\mathbf{W}_{\mathbf{S}}(t+1)\right) - \mathrm{vec}\left(\mathbf{W}_{\mathbf{S}}(t)\right) = -\frac{\eta}{n}\mathbf{Z}_{\mathbf{S}}(t)(\widehat{\mathbf{y}}(t) - \mathbf{y}). \tag{7}$$

When the target function lies in the interpolation space $[\mathcal{H}_K]^{s'}(\mu_0)$ and the covariate distribution on $\mathcal{X}$ satisfies the spectral assumptions of Section 5, we propose to use a novel Preconditioned Gradient Descent (PGD) instead of GD to train the network, which achieves a provably sharper generalization bound than the regular GD detailed in Section 5.2. Let the preconditioner $\mathbf{M} \in \mathbb{R}^{md\times md}$ be generated by Algorithm 2. $\mathbf{M}$ is computed by the sample $\mathbf{Q}$ of N i.i.d. random variables with distribution P on $\mathcal{X}$ used in the definition of $\widehat{K}^{(\mathrm{int})}$ in (5), and $\mathbf{Q}$ is independent of $\mathbf{W}(0)$. In the $(t+1)$-th step of PGD with $t \geq 0$, the weights of the neural network, $\mathbf{W}_{\mathbf{S}}$, are updated by one step of PGD,

$$\mathrm{vec}\left(\mathbf{W}_{\mathbf{S}}(t+1)\right) - \mathrm{vec}\left(\mathbf{W}_{\mathbf{S}}(t)\right) = -\frac{\eta}{n}\mathbf{M}\mathbf{Z}_{\mathbf{S}}(t)(\widehat{\mathbf{y}}(t) - \mathbf{y}), \tag{8}$$

where $\mathbf{y}_i = y_i$, $\widehat{\mathbf{y}}(t) \in \mathbb{R}^n$ with $[\widehat{\mathbf{y}}(t)]_i = f(\mathbf{W}(t), \vec{\mathbf{x}}_i)$. Compared to the vanilla GD update (7), the PGD update (8) inserts the preconditioner $\mathbf{M}$, which reweights the update directions and thereby alters the effective kernel dynamics induced during training. We also denote $f(\mathbf{W}(t), \cdot)$ by $f_t(\cdot)$ as the neural network function with weight vectors $\mathbf{W}(t)$ obtained after the t-th step of PGD. $\mathbf{Z}_{\mathbf{S}}(t) \in \mathbb{R}^{md\times n}$ is defined as

$$[\mathbf{Z}_{\mathbf{S}}(t)]_{[(r-1)d+1:rd]i} = \frac{1}{\sqrt{m}}\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}}\vec{\mathbf{x}}_i a_r,\ i \in [n], r \in [m], \tag{9}$$

where $[\mathbf{Z}_{\mathbf{S}}(t)]_{[(r-1)d+1:rd]i} \in \mathbb{R}^d$ is a vector with elements in the i-th column of $\mathbf{Z}_{\mathbf{S}}(t)$ with indices in $[(r-1)d+1 : rd]$. We employ the following particular symmetric random initialization so that $\widehat{\mathbf{y}}(0) = \mathbf{0}$, which has been used in existing works such as Chizat et al. (2019); Zhang et al. (2020). In our two-layer NN, m is even, $\left\{\vec{\mathbf{w}}_{2r'}(0)\right\}_{r'=1}^{m/2}$ and $\{a_{2r'}\}_{r'=1}^{m/2}$ are initialized randomly and independently according to

$$\vec{\mathbf{w}}_{2r'}(0) \sim \mathcal{N}(\mathbf{0}, \kappa^2\mathbf{I}_d), a_{2r'} \sim \mathrm{unif}\left(\{-1, 1\}\right), \quad \forall r' \in [m/2], \tag{10}$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, unif $(\{-1, 1\})$ denotes the uniform distribution over $\{1, -1\}$, $\kappa = \Theta(1) \in (0, 1)$ controls the magnitude of initialization. We set $\vec{\mathbf{w}}_{2r'-1}(0) = \vec{\mathbf{w}}_{2r'}(0)$ and $a_{2r'-1} = -a_{2r'}$ for all $r' \in [m/2]$. One can verify that $\widehat{\mathbf{y}}(0) = \mathbf{0}$, that is, the initial output of the two-layer NN is zero. Similar to Du et al. (2019b); Yang and Li (2024); Yang (2025), once randomly initialized, $\boldsymbol{a}$ is fixed during the training. We use $\mathbf{W}(0)$ to denote the set of all the random weight vectors at initialization, that is, $\mathbf{W}(0) = \left\{\vec{\mathbf{w}}_r(0)\right\}_{r=1}^m$. We run Algorithm 1 to train the two-layer NN by PGD, where T is the total number of steps for PGD. Early stopping is enforced in Algorithm 1 through a bounded T.

5. Main Results

We present the definition of kernel complexity in Section 5.1, and then introduce the main results for nonparametric regression of this paper in Section 5.2 for a covariate distribution P supported on $\mathcal{X}$. Furthermore, Section 5.4 presents an extension to a more general covariate distribution P supported on the entire $\mathbb{R}^d$.

5.1 Kernel Complexity

The local kernel complexity has been studied by Bartlett et al. (2005); Koltchinskii (2006); Mendelson (2002). For the PD kernel K, we define the empirical kernel complexity $\widehat{R}_K$ and the population kernel complexity R_K as

$$\widehat{R}_K(\varepsilon) \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \min\left\{\widehat{\lambda}_i, \varepsilon^2\right\}}, \quad R_K(\varepsilon) \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^{\infty} \min\left\{\lambda_i, \varepsilon^2\right\}}. \tag{11}$$

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. Sub-root functions are defined in Definition A.2. For a given noise ratio σ_0, the critical empirical radius $\widehat{\varepsilon}_{K,n} > 0$ is the smallest positive solution to the inequality $\widehat{R}_K(\varepsilon) \le \varepsilon^2/\sigma_0$, where $\widehat{\varepsilon}^2_{K,n}$ is also the fixed point of $\sigma_0 \widehat{R}_K(\varepsilon)$ as a function of ε^2: $\sigma_0 \widehat{R}_K(\widehat{\varepsilon}_{K,n}) = \widehat{\varepsilon}^2_{K,n}$. Similarly, the critical population rate $\varepsilon_{K,n}$ is defined to be the smallest positive solution to the inequality $R_K(\varepsilon) \le \varepsilon^2/\sigma_0$, where $\varepsilon^2_{K,n}$ is the fixed point of $\sigma_0 R_K(\varepsilon)$ as a function of ε^2: $\sigma_0 R_K(\varepsilon_{K,n}) = \varepsilon^2_{K,n}$. Kernel complexity can also be defined for the integral kernel $K^{(\mathrm{int})}$, leading to the empirical kernel complexity $\widehat{R}_{K^{(\mathrm{int})}}$ and the population kernel complexity $R_{K^{(\mathrm{int})}}$ for $K^{(\mathrm{int})}$, with the critical empirical radius $\widehat{\varepsilon}_{K^{(\mathrm{int})},n}$ and the critical population rate $\varepsilon_{K^{(\mathrm{int})},n}$, respectively. For simplicity of notation, we use ε_n and $\widehat{\varepsilon}_n$ to denote $\varepsilon_{K^{(\mathrm{int})},n}$ and $\widehat{\varepsilon}_{K^{(\mathrm{int})},n}$, respectively. In this paper we consider the kernel K such that $\min\{\varepsilon_{K,n}, \varepsilon_n\} \cdot n \to \infty$ as $n \to \infty$, which covers most popular positive semi-definite kernels including the kernel (2) and a broad range of data distributions (Yang et al., 2017).

Let $\eta_t \coloneqq \eta t$ for all $t \ge 0$. We then define the stopping time $\widehat{T}$ as

$$\widehat{T} \coloneqq \min\left\{t \ge 1 \colon \widehat{R}_{K^{(\mathrm{int})}}(\sqrt{1/\eta_t}) > (\sigma_0 \eta_t)^{-1}\right\} - 1. \tag{12}$$

The stopping time in fact limits the number of steps T for Algorithm 1, which enforces the early stopping mechanism. In fact, as will be shown later in this section, we need to have $T \le \widehat{T}$ when training the two-layer NN (1) by PGD with Algorithm 1. The stopping time $\widehat{T}$ is chosen so that optimization progress and statistical complexity are balanced, which is the standard role of early stopping in localized kernel-complexity analyses.

5.2 Nonparametric Regression with Covariate Distribution Supported on $\mathcal{X}$

We present in Theorem 5.1 our main result for nonparametric regression where $f^* \in [\mathcal{H}_K]^{s'}(\mu_0)$ by training the two-layer NN (1) with PGD and early stopping using Algorithm 1, when the covariate distribution P supported on $\mathcal{X}$ admits the polynomial EDR, that is, $\lambda_j \asymp j^{-2\alpha}$ for $j \ge 1$ with $2\alpha = d/(d-1)$. It is shown by existing works such as Hu et al. (2021, Lemma 3.1) that such polynomial EDR holds when P is the spherical uniform distribution, that is, $P = \mathrm{Unif}(\mathcal{X})$. We note that such an assumption about spherical uniform distribution has been widely adopted in the theoretical analysis of neural networks (Bai and Lee, 2020; Ghorbani et al., 2021; Hu et al., 2021).

Theorem 5.1. Suppose the covariate distribution P has full support on $\mathcal{X}$ and admits a polynomial EDR such that $\lambda_j \asymp j^{-2\alpha}$ for $j \ge 1$ with $\alpha = d/(2(d-1))$. Suppose $f^* \in \mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) = [\mathcal{H}_K]^{s+2}(\mu_0)$ for a fixed radius $\mu_0 > 0$. Let $s \ge 1$, $c_T, c_t \in (0,1]$ be positive constants, and $c_T\widehat{T} \le T \le \widehat{T}$. Suppose $\delta \in (0,1)$,

$$m \gtrsim n^{\frac{25\alpha(s+2)}{2\alpha(s+2)+1}} d^{\frac{5}{2}}, \quad N \gtrsim n^{\frac{8\alpha(s+2)}{2\alpha(s+2)+1}} \log(4n/\delta), \tag{13}$$

and the neural network $f_t = f(\mathbf{W}(t), \cdot)$ is trained by PGD in Algorithm 1 with the learning rate $\eta = \Theta(1) \in (0, 8)$ with $T \le \widehat{T}$. Then for every $t \in [c_t T : T]$, with probability (w.p.) at least $1 - \exp(-\Theta(n)) - \delta - 7\exp\left(-\Theta(n^{\frac{1}{2\alpha(s+2)+1}})\right) - 2/n$ over $\mathbf{w}, \mathbf{S}, \mathbf{Q}, \mathbf{W}(0)$, the stopping time satisfies $\widehat{T} \asymp n^{\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}$, and

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] \lesssim n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}. \tag{14}$$

Remark 5.2. It follows from the integral-operator comparison in the proof of Proposition 5.3 that if the Radon-Nikodym derivative of μ with respect to σ, where σ is the uniform probability measure on $\mathcal{X}$, is $\mathrm{d}\mu/\mathrm{d}\sigma = \rho^2$ for a bounded non-zero Riemann-integrable function ρ, then the polynomial EDR holds. As a result, Theorem 5.1 holds with such a probability measure μ of the covariate distribution P whenever its remaining assumptions are satisfied, which covers the spherical uniform distribution as a special case.

5.3 Significance of Theorem 5.1

Statistical Significance. Theorem 5.1 establishes the sharp rate $\mathcal{O}(n^{-\frac{ds'}{ds'+d-1}})$ for target functions in $[\mathcal{H}_K]^{s'}(\mu_0)$ with $s' = s + 2 \ge 3$. Since Theorem C.2 shows that $\lambda_j^{(\mathrm{int})} = \lambda_j^{s'} \asymp j^{-2\alpha s'}$, this rate matches the minimax optimal rate associated with kernel regression using $K^{(\mathrm{int})}$ over $\mathcal{H}_{K^{(\mathrm{int})}}$ (Stone, 1985; Yang and Barron, 1999; Yuan and Zhou, 2016). In this sense, the theorem identifies a sharp statistical regime for the neural network under the interpolation-space target assumption.

Reason for Improved Rate. Our sharp rate in Theorem 5.1 is sharper than that in existing works with algorithm guarantees as discussed in Section 3. Our rate arises because PGD changes the kernel induced by training. Vanilla GD leads to the regular NTK (2), whose spectrum yields the standard rate under the usual RKHS target assumption. In contrast, PGD induces the integral kernel $K^{(\mathrm{int})}$, whose eigenvalues decay faster and therefore have lower associated kernel complexity. In fact, the kernel induced by PGD is $K^{(\mathrm{int})}$ instead of K defined in (2), whose eigenvalues satisfy $\lambda_j^{(\mathrm{int})} = \lambda_j^{s'} < \lambda_j$ with $s' \geq 3$. This spectral change is the mechanism that enables the improved rate under spectral bias.

5.4 Extension to Covariate Distribution Supported on $\mathbb{R}^d$

In this section, we consider a covariate distribution P supported on the entire space $\mathbb{R}^d$ such that the distribution of the normalized covariate admits the polynomial EDR.

In order to handle data supported on $\mathbb{R}^d$, we introduce a normalization map $\phi\colon \mathbb{R}^d \to \mathcal{X} = \mathbb{S}^{d-1}, \phi(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$ for $\mathbf{x} \neq \mathbf{0}$. Define $\phi(\mathbf{0})$ arbitrarily. This choice has no effect for covariate distributions with a density. We recall that the probability measure of P is still μ, and the push-forward probability measure on $\mathcal{X}$, denoted as $\phi^*\mu$ which is induced by ϕ, is defined as $(\phi^*\mu)(A) \coloneqq \mu(\phi^{-1}(A))$ for any set $A \subseteq \mathcal{X}$. For simplicity of notation, we let $\bar{\mathbf{x}} = \phi(\mathbf{x})$. As a result, $\phi^*\mu$ is the probability measure of the normalized covariate $\bar{\mathbf{x}}$.

We let $\mu_j(K, \mathcal{X}, \mu')$ be the j-th eigenvalue of the integral operator T_K associated with K under the distribution with probability measure μ' supported on $\mathcal{X}$. Proposition 5.3 shows that when the probability density function p of the covariate distribution P satisfies $p(\mathbf{x}) \lesssim (1 + \|\mathbf{x}\|_2^2)^{-(d+1)/2}$ for all $\mathbf{x} \in \mathbb{R}^d$, then the distribution of the normalized covariate admits the polynomial EDR, that is, $\mu_j(K, \mathcal{X}, \phi^*(\mu)) \asymp j^{-d/(d-1)}$ holds for all $j \geq 1$.

Proposition 5.3. Suppose the covariate distribution P, supported on $\mathbb{R}^d$, has a bounded probability density function p s.t. $p(\mathbf{x}) \lesssim (1 + \|\mathbf{x}\|_2^2)^{-(d+1)/2}$ for all $\mathbf{x} \in \mathbb{R}^d$. Then the polynomial EDR $\mu_j(K, \mathcal{X}, \phi^*(\mu)) \asymp j^{-d/(d-1)}$ holds for all $j \geq 1$.

When the target function, as a function of the normalized covariate, still lies in the interpolation space corresponding to the NTK (2) defined on the normalized covariates, then under the condition on the distribution P in Proposition 5.3, it can be shown as a direct consequence of Theorem 5.1 that the sharp regression risk bound (14) still holds by training the two-layer NN (1) with Algorithm 1. This result is presented as the following corollary. The algorithm is applied to the normalized training features, with the independent auxiliary sample drawn from $\phi^*\mu$.

Corollary 5.4. Suppose $f^* = f^*(\bar{\mathbf{x}}) \in [\mathcal{H}_K]^{s'}(\mu_0)$ where $K(\bar{\mathbf{u}}, \bar{\mathbf{v}}) = \frac{\langle \bar{\mathbf{u}}, \bar{\mathbf{v}} \rangle}{2\pi}(\pi - \arccos\langle \bar{\mathbf{u}}, \bar{\mathbf{v}} \rangle)$ for every $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. Then under the conditions of Theorem 5.1 and Proposition 5.3, (14) still holds with the same high probability as that claimed in Theorem 5.1.

Corollary 5.4 shows that PGD with early stopping still achieves a minimax optimal regression risk rate under a general covariate distribution P supported on $\mathbb{R}^d$ satisfying Proposition 5.3. We remark that Corollary 5.4 includes the sub-Gaussian distribution considered in Li et al. (2024) as a special case. To see this, the density p of the sub-Gaussian distribution considered in Li et al. (2024) is assumed to satisfy $p(\mathbf{x}) \lesssim (1 + \|\mathbf{x}\|_2^2)^{-(d+3)/2}$ for all $\mathbf{x} \in \mathbb{R}^d$, which satisfies the condition in our Proposition 5.3. Our Corollary 5.4 holds

for such a sub-Gaussian distribution whenever the remaining conditions of Theorem 5.1 are satisfied.

5.5 Adaptivity to the Smoothness Parameter s' of the Interpolation Space

In practice, the regularity exponent s' such that $f^* \in [\mathcal{H}_K]^{s'}(\mu_0)$ may be unknown, whereas the preconditioner in Algorithm 2 depends on s. The following nesting property shows that the validity of our guarantees does not require exact knowledge of s', although attaining the rate corresponding to the largest admissible value of s may require such knowledge.

Proposition 5.5. We have $[\mathcal{H}_K]^{s_1}(\mu_0) \subseteq [\mathcal{H}_K]^{s_2}(\mu_0)$ for all $s_1 \geq s_2 > 0$.

Proof Every $f \in [\mathcal{H}_K]^{s_1}(\mu_0)$ such that $f = \sum_{j\geq 1} \beta_j e_j$ satisfies $\sum_{j\geq 1} \beta_j^2/\lambda_j^{s_1} \leq \mu_0^2$. Since $\lambda_j < 1$ for all $j \geq 1$, we know that $\sum_{j\geq 1} \beta_j^2/\lambda_j^{s_2} \leq \mu_0^2$ for $s_1 \geq s_2 > 0$, hence $f \in [\mathcal{H}_K]^{s_2}$. ■

Proposition 5.5 implies robustness of PGD to underspecification of the regularity exponent. In particular, if $f^* \in [\mathcal{H}_K]^{s'}(\mu_0)$ with $s' \geq 3$, then for every $1 \leq s \leq s' - 2$,

$$f^* \in [\mathcal{H}_K]^{s'}(\mu_0) \subseteq [\mathcal{H}_K]^{s+2}(\mu_0).$$

Consequently, Theorem 5.1 and Corollary 5.4 remain valid with effective regularity exponent $s+2$, yielding the risk rate $\mathcal{O}(n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}})$. In particular, the universal choice $s = 1$ is valid for every $s' \geq 3$ and requires no knowledge of s'. It guarantees the rate $\mathcal{O}(n^{-\frac{6\alpha}{6\alpha+1}})$ uniformly over these target classes. If $s' > 3$, however, this choice does not generally attain the faster oracle rate associated with the true exponent s'. Consequently, Proposition 5.5 establishes robustness to underspecification, rather than full rate-adaptivity to unknown smoothness. A practical possibility is to select s from a finite grid using validation data. A theoretical analysis of such a data-dependent selection procedure is left for future work.

6. Roadmap of Proofs

We present the roadmap of proofs for the main result, Theorem 5.1, in this section. We first present the results about the uniform convergence to the NTK (2) during the training process by PGD in Section 6.1, then introduce the basic definitions in Section 6.2. The detailed roadmap and key technical results for the main result are presented in Section 6.3. The difference between this work and the existing kernel learning theory (Caponnetto and De Vito, 2007) is introduced in Section 6.4, and our novel proof strategy is described in Section 6.5.

The proof of Theorem 5.1 is based on the key technical results in Section 6.3, in particular, the uniform convergence of $\widehat{K}^{(\mathrm{int})}(\cdot, \mathbf{x}')$ to $K^{(\mathrm{int})}(\cdot, \mathbf{x}')$ for every fixed $\mathbf{x}' \in \mathcal{X}$ in Theorem 6.2, the generalization result in Theorem 6.5, and the optimization results in Lemma 6.3, Theorem 6.4, and Lemma 6.6.

6.1 Uniform Convergence to the NTK and More

We define the following functions:

$$h(\mathbf{w},\mathbf{x},\mathbf{y}) \coloneqq \mathbf{x}^\top \mathbf{y} \mathbb{I}_{\{\mathbf{w}^\top \mathbf{x}\geq 0\}} \mathbb{I}_{\{\mathbf{w}^\top \mathbf{y}\geq 0\}}, \qquad \widehat{h}(\mathbf{W},\mathbf{x},\mathbf{y}) \coloneqq \frac{1}{m}\sum_{r=1}^{m} h(\vec{\mathbf{w}}_r,\mathbf{x},\mathbf{y}), \tag{15}$$

$$v_R(\mathbf{w},\mathbf{x}) \coloneqq \mathbb{I}_{\{|\mathbf{w}^\top \mathbf{x}|\leq R\}}, \qquad \widehat{v}_R(\mathbf{W},\mathbf{x}) \coloneqq \frac{1}{m}\sum_{r=1}^{m} v_R(\vec{\mathbf{w}}_r,\mathbf{x}). \tag{16}$$

Then we have the following theorem stating the uniform convergence of $\widehat{h}(\mathbf{W}(0),\cdot,\cdot)$ to the NTK K defined in (2) and uniform convergence of $\widehat{v}_R(\mathbf{W}(0),\mathbf{x})$ to 0.

Theorem 6.1 (Yang (2025, Theorem 6.1),Yang and Li (2024, Theorem VI.7, VI.8))**.** The following results hold with $\eta \lesssim 1$, $m \gtrsim \max\left\{n^{2/d}, \Theta(T^{\frac{5}{3}})\right\}$, and $m/\log m \geq d$. Here $R = \eta c_{\mathbf{u}} T/(2\sqrt{m})$, with $c_{\mathbf{u}}$ defined in Section 6.2. The learning-rate condition is used to specialize part (2) to this training-dependent radius. The underlying concentration inequality at a fixed radius does not depend on η.

(1) With probability at least $1-1/n$ over the random initialization $\mathbf{W}(0) = \left\{\vec{\mathbf{w}}_r(0)\right\}_{r=1}^m$,

$$\sup_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{X}} \left|K(\mathbf{x},\mathbf{y}) - \widehat{h}(\mathbf{W}(0),\mathbf{x},\mathbf{y})\right| \leq C_1(m/2,d,1/n) \lesssim \sqrt{\frac{d\log m}{m}}, \tag{17}$$

(2) With probability at least $1-1/n$ over the random initialization $\mathbf{W}(0) = \left\{\vec{\mathbf{w}}_r(0)\right\}_{r=1}^m$,

$$\sup_{\mathbf{x}\in\mathcal{X}} |\widehat{v}_R(\mathbf{W}(0),\mathbf{x})| \leq \frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2,d,1/n) \lesssim \sqrt{d} m^{-\frac{1}{5}} T^{\frac{1}{2}}. \tag{18}$$

6.2 Basic Definitions

We introduce the following definitions for the proof of the main result. Let the gram matrix of $K^{(\text{int})}$ over the training features $\mathbf{S}$ be $\mathbf{K}^{(\text{int})} \in \mathbb{R}^{n\times n}, \mathbf{K}_{ij}^{(\text{int})} = K^{(\text{int})}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j)$ for $i,j\in[n]$, and $\mathbf{K}_n^{(\text{int})} \coloneqq \mathbf{K}^{(\text{int})}/n$. Similarly, $\widehat{\mathbf{K}}^{(\text{int})} \in \mathbb{R}^{n\times n}$ is the gram matrix of $\widehat{K}^{(\text{int})}$ over $\mathbf{S}$, and $\widehat{\mathbf{K}}_n^{(\text{int})} = \widehat{\mathbf{K}}^{(\text{int})}/n$. Let the eigendecomposition of $\mathbf{K}_n^{(\text{int})}$ be $\mathbf{K}_n^{(\text{int})} = \mathbf{U}^{(\text{int})}\mathbf{\Sigma}^{(\text{int})}\mathbf{U}^{(\text{int})\top}$, where $\mathbf{\Sigma}^{(\text{int})}$ is a diagonal matrix with its diagonal elements $\left\{\widehat{\lambda}_i^{(\text{int})}\right\}_{i=1}^n$ being the eigenvalues of $\mathbf{K}_n^{(\text{int})}$ and sorted in a non-increasing order. We have $\widehat{\lambda}_1^{(\text{int})} \in (0, 1/8]$.

We define $\mathbf{u}(t) \coloneqq \widehat{\mathbf{y}}(t) - \mathbf{y}$. Let $\tau \leq 1$ be a positive number. We define $c_{\mathbf{u}} \coloneqq \mu_0/\min\left\{2\sqrt{2}, \sqrt{2e\eta}\right\} + \sigma_0 + \tau + 1$. For $t \geq 0$ and $T \geq 1$, we define

$$R \coloneqq \frac{\eta c_{\mathbf{u}} T}{2\sqrt{m}}, \tag{19}$$

$$\mathcal{V}_t \coloneqq \left\{\mathbf{v}\in\mathbb{R}^n \colon \mathbf{v} = -\left(\mathbf{I}_n - \eta\mathbf{K}_n^{(\text{int})}\right)^t f^*(\mathbf{S})\right\}, \tag{20}$$

$$\mathcal{E}_{t,\tau} \coloneqq \left\{ \mathbf{e} \colon \mathbf{e} = \vec{\mathbf{e}}_1 + \vec{\mathbf{e}}_2 \in \mathbb{R}^n, \vec{\mathbf{e}}_1 = -\left(\mathbf{I}_n - \eta \mathbf{K}_n^{(\mathrm{int})}\right)^t \mathbf{w}, \left\| \vec{\mathbf{e}}_2 \right\|_2 \lesssim \sqrt{n}\tau \right\}, \tag{21}$$

where the upper bound for $\left\| \vec{\mathbf{e}}_2 \right\|_2$ only hides an absolute positive constant independent of t. The set $\mathcal{V}_t$ represents the clean signal trajectory in the training dynamics of PGD. The set $\mathcal{E}_{t,\tau}$ captures the controlled residual and noise components which are added to the clean signal trajectory. We define

$$\mathcal{W}_0 \coloneqq \{\mathbf{W}(0) \colon (17), (18) \text{ hold}\} \tag{22}$$

as the set of all the good random initializations which satisfy (17) and (18) in Theorem 6.1. Theorem 6.1 shows that we have good random initialization with high probability, that is, $\Pr\left[\mathbf{W}(0) \in \mathcal{W}_0\right] \geq 1 - 2/n$. When $\mathbf{W}(0) \in \mathcal{W}_0$, the uniform convergence results, (17) and (18), hold, which guarantees our main result about optimization of the neural network by PGD.

We define the set of neural network weights and the set of functions represented by the neural network during training as follows.

$$\mathcal{W}(\mathbf{S}, \mathbf{Q}, \mathbf{W}(0), T) \coloneqq \Bigg\{ \mathbf{W} \colon \exists t \in [T] \text{ s.t. } \mathrm{vec}\left(\mathbf{W}\right) = \mathrm{vec}\left(\mathbf{W}(0)\right) - \sum_{t'=0}^{t-1} \frac{\eta}{n} \mathbf{M} \mathbf{Z}_{\mathbf{S}}(t') \mathbf{u}(t'),$$
$$\mathbf{u}(t') \in \mathbb{R}^n, \mathbf{u}(t') = \mathbf{v}(t') + \mathbf{e}(t'), \mathbf{v}(t') \in \mathcal{V}_{t'}, \mathbf{e}(t') \in \mathcal{E}_{t',\tau}, \text{ for all } t' \in [0, t-1] \Bigg\}. \tag{23}$$

$\mathcal{W}(\mathbf{S}, \mathbf{Q}, \mathbf{W}(0), T)$ is the set of weights of the neural network trained by PGD on the training features $\mathbf{S}$ and random initialization $\mathbf{W}(0)$ with the preconditioner $\mathbf{M}$ generated by $\mathbf{Q}$ and the steps of PGD not greater than T. The set of functions represented by the two-layer NN with weights in $\mathcal{W}(\mathbf{S}, \mathbf{Q}, \mathbf{W}(0), T)$ is then defined as

$$\mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \mathbf{Q}, \mathbf{W}(0), T) \coloneqq \{f_t = f(\mathbf{W}(t), \cdot) \colon \exists t \in [T], \mathbf{W}(t) \in \mathcal{W}(\mathbf{S}, \mathbf{Q}, \mathbf{W}(0), T)\}. \tag{24}$$

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

$$\mathcal{F}_{\mathrm{ext}}(B, w) \coloneqq \{f \colon f = h + e, h \in \mathcal{H}_{K^{(\mathrm{int})}}(B), \|e\|_\infty \leq w\}. \tag{25}$$

In Theorem 6.4, we will show that the neural network function at each step of PGD, f_t, belongs to the space $\mathcal{F}_{\mathrm{ext}}(B_h, w)$ with high probability where

$$B_h \coloneqq \mu_0 + 1 + \sqrt{2}. \tag{26}$$

6.3 Detailed Roadmap and Key Technical Results

The summary of the approaches and key technical results in the proofs is presented as follows. Our main result is built upon the following three significant technical results of independent interest. In the following text, $\delta \in (0, 1)$ denotes a small probability.

First, we establish the uniform convergence of the empirical integral kernel $\widehat{K}^{(\mathrm{int})}(\cdot, \mathbf{x}')$ to the integral kernel $K^{(\mathrm{int})}(\cdot, \mathbf{x}')$ for every fixed $\mathbf{x}' \in \mathcal{X}$. Such uniform convergence is presented in the following theorem, Theorem 6.2, which is built upon the martingale concentration inequality in Hilbert spaces (Pinelis, 1994, Theorem 3.5).

Theorem 6.2. For every fixed $\mathbf{x}' \in \mathcal{X}$ and every $\delta \in (0,1)$, with probability at least $1-\delta$ over $\mathbf{Q} = \left\{\vec{\mathbf{q}}_i\right\}_{i=1}^{N}$, we have

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{K}^{(\mathrm{int})}(\mathbf{x},\mathbf{x}') - K^{(\mathrm{int})}(\mathbf{x},\mathbf{x}')\right| \lesssim \sqrt{\frac{\log(4/\delta)}{N}}. \tag{27}$$

Theorem 6.2 is the key tool needed to analyze the training dynamics of PGD which involves the integral kernel $K^{(\mathrm{int})}$. Then, using the novel PGD algorithm and the uniform convergence in Theorem 6.2, we have the following optimization result in Lemma 6.3.

Lemma 6.3. Suppose the network width m and N are sufficiently large and finite, and the neural network $f(\mathbf{W}(t),\cdot)$ is trained by PGD using Algorithm 1 with the learning rate $\eta = \Theta(1) \in (0,8)$ on the random initialization $\mathbf{W}(0) \in \mathcal{W}_0$. Then for every $\delta \in (0,1)$, with high probability over the random noise $\mathbf{w}$ and the random sample $\mathbf{Q}$, $\mathbf{W}(t) \in \mathcal{W}(\mathbf{S},\mathbf{Q},\mathbf{W}(0),T)$. Moreover, for all $t \in [0,T]$, $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ where $\mathbf{u}(t) = \widehat{\mathbf{y}}(t) - \mathbf{y}$, $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, and $\|\mathbf{u}(t)\|_2 \le c_{\mathbf{u}}\sqrt{n}$.

Lemma 6.3 formalizes that the PGD residual dynamics $\mathbf{u}(t)$ can be decomposed into a structured clean signal term $\mathbf{v}(t)$ and a controlled error term $\mathbf{e}(t)$, which is the basis for the function-level decomposition to be established next. Based on Lemma 6.3, we obtain a novel and principled decomposition of the neural network function at any step of PGD into a function in $\mathcal{H}_{K^{(\mathrm{int})}}$ with bounded RKHS-norm, and an error function with a small L^∞-norm with high probability.

Theorem 6.4. Suppose the network width m and N are sufficiently large and finite, and the neural network $f_t = f(\mathbf{W}(t),\cdot)$ is trained by PGD using Algorithm 1 with the learning rate $\eta = \Theta(1) \in (0,8)$ and the random initialization $\mathbf{W}(0) \in \mathcal{W}_0$. Then for every $t \in [T]$, with high probability over $\mathbf{w}$ and $\mathbf{Q}$, f_t has the following decomposition on $\mathcal{X}$: $f_t = h_t + e_t \in \mathcal{F}_{\mathrm{ext}}(B_h, w)$, where $h_t \in \mathcal{H}_{K^{(\mathrm{int})}}(B_h)$ with B_h defined in (26), $e_t \in L^\infty$ with sufficiently small magnitude on $\mathcal{X}$, that is, $\|e_t\|_\infty \le w$.

Second, a new technique based on local Rademacher complexity is developed to tightly bound the nonparametric regression risk in Theorem 6.5, which is based on the Rademacher complexity of a localized subset of a much larger function class presented in Lemma B.7 in the appendix. We then use Theorem 6.4 and Lemma B.7 to derive Theorem 6.5.

Theorem 6.5. Suppose $w \in (0,1)$, the network width m and N are sufficiently large and finite, and the neural network $f_t = f(\mathbf{W}(t),\cdot)$ is trained by PGD using Algorithm 1 with the learning rate $\eta = \Theta(1) \in (0,8)$ on the random initialization $\mathbf{W}(0)$, and $T \le \widehat{T}$. Then for every $t \in [T]$, with high probability over $\mathbf{w}$, the random training features $\mathbf{S}$, $\mathbf{Q}$, and $\mathbf{W}(0)$,

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim \min_{0\le Q\le n}\left(\frac{Q}{n} + \left(\frac{\sum_{q=Q+1}^{\infty}\lambda_q^{(\mathrm{int})}}{n}\right)^{1/2}\right) + \varepsilon_n^2 + w, \tag{28}$$

where $\varepsilon_n = \varepsilon_{K^{(\mathrm{int})},n}$ is the population critical radius associated with $K^{(\mathrm{int})}$, namely the smallest positive ε satisfying $R_{K^{(\mathrm{int})}}(\varepsilon) \le \varepsilon^2/\sigma_0$.

Theorem 6.5 converts the spectral decay of $K^{(\mathrm{int})}$ into a sharp population risk control for the neural network trained by PGD via localized Rademacher complexity.

Third, we obtain the following sharp upper bound for the training loss $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$. Lemma 6.6 provides the empirical loss decay needed to complement the local complexity-based generalization bound in Theorem 6.5 and close the final rate.

Lemma 6.6. For every $t \in [T]$, with high probability over $\mathbf{w}$,

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \le \frac{3}{\eta t}\left(\frac{\mu_0^2}{2e} + \frac{1}{\eta T} + 2\right), \tag{29}$$

where μ_0 is the target-class radius so that $f^* \in \mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) = [\mathcal{H}_K]^{s'}(\mu_0)$.

Proof Sketch of Theorem 5.1. We can then prove Theorem 5.1 using the upper bound for the regression risk in (28) of Theorem 6.5 where w is set to $n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}$. With $\widehat{T} \asymp n^{\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}$, it follows from (29) of Lemma 6.6 that $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}$. Then we have $\mathbb{E}_P\left[(f_t - f^*)^2\right] \lesssim n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}$ with $Q \asymp n^{\frac{1}{2\alpha(s+2)+1}}$ in (28). The formal proof of Theorem 5.1 is presented below.

Proof of Theorem 5.1. We apply Theorem B.8 and Lemma B.9 in the appendix to prove this theorem, and Theorem B.8 and Lemma B.9 are the formal versions of Theorem 6.5 and Lemma 6.6, respectively.

First, let $\mathcal{A}_n$ be the critical-radius comparison event used in the proof of Theorem B.8. On this event, the stopping-time estimate below gives $T \le \widehat{T} \lesssim n^{2\alpha(s+2)/(2\alpha(s+2)+1)}$. The width condition implies $m \gtrsim \max\left\{n^{2/d}, T^{5/3}\right\}$ and $m/\log m \ge d$ for all sufficiently large n. Applying Theorem 6.1 conditionally on $\mathbf{S}$ gives $\Pr\left[\mathcal{A}_n \cap \{\mathbf{W}(0) \notin \mathcal{W}_0\}\right] \le 2/n$. It follows from Theorem C.2 in Section C that $\lambda_j^{(\mathrm{int})} = \lambda_j^{s+2} \asymp j^{-2\alpha(s+2)}$ for $j \ge 1$. With $\eta = \Theta(1)$, it follows by Lemma B.9 that with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$,

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

Plugging such bound for $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$ in (113) of Theorem B.8 leads to

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] \lesssim \frac{1}{\eta t} + \left(\frac{Q}{n} + \left(\frac{\sum_{q=Q+1}^{\infty} \lambda_q^{(\mathrm{int})}}{n}\right)^{1/2}\right) + \varepsilon_n^2 + w, \tag{30}$$

where $Q \asymp n^{\frac{1}{2\alpha(s+2)+1}}$ on the RHS of (30). It can be verified that

$$\sum_{q=Q+1}^{\infty} \lambda_q^{(\mathrm{int})} \lesssim \int_Q^{\infty} x^{-2\alpha(s+2)} \mathrm{d}x \lesssim 1/(2\alpha(s+2)-1) \cdot n^{-\frac{2\alpha(s+2)-1}{2\alpha(s+2)+1}}.$$

As a result, we have

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - \frac{6}{\eta t}\left(\frac{\mu_0^2}{2e} + \frac{1}{\eta} + 2\right) \lesssim n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}} + \varepsilon_n^2 + w. \tag{31}$$

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

$$\widehat{\varepsilon}_n^2 \leq \frac{1}{\eta\widehat{T}} \leq \frac{2}{\eta(\widehat{T}+1)} \leq 2\widehat{\varepsilon}_n^2. \tag{32}$$

It is well known, such as Raskutti et al. (2014, Corollary 3), that $\varepsilon_n^2 \asymp n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}$. It follows from Lemma B.14 that $\widehat{\varepsilon}_n^2 \asymp \varepsilon_n^2$ with probability at least $1 - 4\exp(-\Theta(n\varepsilon_n^2))$ over $\mathbf{S}$. In addition, combined with the fact that $T \asymp \widehat{T}$, for any $t \in [c_t T, T]$, we have

$$\frac{1}{\eta t} \asymp \frac{1}{\eta\widehat{T}} \asymp \frac{1}{\eta T} \asymp \widehat{\varepsilon}_n^2 \asymp \varepsilon_n^2 \asymp n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}.$$

Also, we set $w = n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}$ in (31), then with $\eta = \Theta(1)$,

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] \lesssim n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}. \tag{33}$$

With $N \gtrsim n^{\frac{8\alpha(s+2)}{2\alpha(s+2)+1}}\log(4n/\delta)$, the requirement on N, (95) in Theorem B.6 that $N \gtrsim \max\left\{T^2\log(4n/\delta)/w^2, T^4\log(4n/\delta)\right\}$ is satisfied. In addition, with $m \gtrsim n^{\frac{25\alpha(s+2)}{2\alpha(s+2)+1}} d^{\frac{5}{2}}$ and $w = n^{-\frac{2\alpha(s+2)}{2\alpha(s+2)+1}}$, the condition (94) on m in Theorem B.6 that $m \gtrsim \max\{T^{\frac{15}{2}} d^{\frac{5}{2}}/w^5,$ $T^{\frac{25}{2}} d^{\frac{5}{2}}\}$ is satisfied. The joint event constructed in the proof of Theorem B.8 includes the initialization, critical-radius comparison, trajectory, and local-complexity events. On $\mathcal{A}_n$, the additional noise-filter failure probability in Lemma B.9 is bounded by $\exp(-\Theta(n\varepsilon_n^2))$. A union bound adds this term to the six exponential terms in Theorem B.8, while counting the initialization failure $2/n$ and the failure of $\mathcal{A}_n$ only once, yielding the stated probability. ■

6.4 Difference from Existing Kernel Learning Theory

In this section, we compare our PGD guarantee with existing kernel learning theory and vanilla-GD results under matching source and spectral parameters. The comparison distinguishes logarithmic confidence factors from the polynomial risk exponent.

Under the source condition $f^* \in \mathcal{H}_{K^{(\mathrm{int})}}(\mu_0) = [\mathcal{H}_K]^{s'}(\mu_0)$ with $s' \geq 3$, and certain spectral, embedding, and noise assumptions, the spectral algorithm in Zhang et al. (2024, Theorem 1) using the regular NTK (2), with qualification at least $s'/2$, achieves the regression risk bound $\mathcal{O}(n^{-\frac{ds'}{ds'+d-1}}\log^2(6/\delta))$ with probability at least $1-\delta$, for every fixed $\delta \in (0,1)$ and sufficiently large $n \geq n_0(\delta)$. This logarithmic confidence dependence $\log^2(6/\delta)$ is consistent with the vanilla-GD upper bound in Li et al. (2024, Proposition 13), expressed under the matching source and spectral parameters in Table 1. Related confidence dependence appears in the finite-sample regularized least-squares upper bound of Caponnetto and De Vito (2007, Theorem 4), which uses the multiplier $32\log^2(6/\delta)$.

Furthermore, Zhang et al. (2024, Theorem 2) establishes the minimax lower bound $\Omega(n^{-\frac{ds'}{ds'+d-1}})$ over the corresponding source ball. Our Theorem 5.1 matches this minimax lower bound without a logarithmic factor in the risk bound, under its stated assumptions and probability guarantee.

In the comparisons above, canonical kernel regression methods (Caponnetto and De Vito, 2007; Yao et al., 2007) are applied to the regular NTK (2) with EDR $\lambda_j \asymp j^{-2\alpha}$. They are not applied to the PGD-induced kernel $K^{(\mathrm{int})}$ in these comparisons. The general kernel frameworks can also be applied to $K^{(\mathrm{int})}$ when their assumptions are satisfied. On the other hand, the two-layer NN, trained by our novel PGD, approximately performs kernel regression with a completely different new kernel, the integral kernel $K^{(\mathrm{int})}$ with the smoother capacity condition ($\lambda_j^{(\mathrm{int})} = \lambda_j^{s'} \asymp j^{-2\alpha s'}$), compared to the original capacity condition. Our key insight is that, the interpolation space $[\mathcal{H}_K]^{s'}$ is in fact the RKHS associated with the integral kernel: $[\mathcal{H}_K]^{s'} = \mathcal{H}_{K^{(\mathrm{int})}}$. Kernel regression with the integral kernel and the target function $f^* \in \mathcal{H}_{K^{(\mathrm{int})}}(\mu_0)$ renders the minimax optimal rate of $\mathcal{O}(n^{-\frac{ds'}{ds'+d-1}})$ according to the analytical results in Stone (1985); Yang and Barron (1999); Yuan and Zhou (2016), which is the same rate as that in Theorem 5.1.

Learning in the Induced RKHS. Our PGD analysis decomposes the trained network as $f_t = h_t + e_t$, with $h_t \in \mathcal{H}_{K^{(\mathrm{int})}}(B_h)$ and $\|e_t\|_\infty \le w$. In the matched-kernel notation of Table 1, existing vanilla-GD methods such as Li et al. (2024, Proposition 13) approximate kernel regression in $\mathcal{H}_K$ and, under the comparable source condition $s' \ge 3$, learn target functions in $\mathcal{H}_K$. The common target class satisfies $\mathcal{H}_{K^{(\mathrm{int})}} = [\mathcal{H}_K]^{s'} \subseteq \mathcal{H}_K$. PGD changes the reference kernel and controls the RKHS component of its iterates in $\mathcal{H}_{K^{(\mathrm{int})}}$. Under the matching source and spectral parameters in Table 1, the existing vanilla-GD bound in Li et al. (2024, Proposition 13) contains the logarithmic confidence factor $\log^2(12/\delta)$, whereas our Theorem 5.1 gives a log-free bound under its stated assumptions and probability guarantee. These statements compare proved upper bounds. It is technically nontrivial to induce such a new kernel $K^{(\mathrm{int})}$ when training the two-layer NN with the novel PGD algorithm, to be detailed by our novel proof strategies in the next subsection. Our results also give a smaller sample-size exponent in the sufficient condition for network width compared with the cited analyses, as detailed in the next subsection. We remark that PGD is used in the training process to change the effective kernel from K to $K^{(\mathrm{int})}$, and the trained two-layer NN approximates kernel regression with this integral kernel.

6.5 Novel Proof Strategy and Insights

Our results are built upon two significantly novel proof strategies. First, uniform convergence to the NTK is established during the training process by PGD, so that we can have a novel decomposition of the neural network function at any step of PGD into a function in the RKHS associated with the integral kernel $K^{(\mathrm{int})}$ and an error function with small L^∞-norm with high probability in Theorem 6.4. We remark that PGD is carefully designed so that the induced NTK of the two-layer NN (1) is the integral kernel $K^{(\mathrm{int})}$. In a strong contrast, the vanilla GD extensively used in the existing works for the analysis of over-parameterized neural networks, such as Suh et al. (2022); Li et al. (2024); Yang and Li (2024); Yang (2025), induces the regular NTK such as (2). The uniform convergence of

$\widehat{K}^{(\mathrm{int})}$ to $K^{(\mathrm{int})}$ is established in this work, using the martingale concentration inequality in Hilbert spaces (Pinelis, 1994, Theorem 3.5). As explained in Section 3 and Section 5, PGD leads to a sharper rate of the nonparametric regression risk due to the lower kernel complexity of $K^{(\mathrm{int})}$ than that of the regular NTK such as (2).

Second, based on such novel decomposition in Theorem 6.4, a new technique based on local Rademacher complexity is developed to tightly bound the Rademacher complexity of the function class comprising all the neural network functions obtained by PGD iterations, leading to the sharp bound for the regression risk in Theorem 6.5.

PGD changes the effective kernel within the reference training dynamics: it induces $K^{(\mathrm{int})}$, whose complexity is lower than that of the regular NTK K. The resulting log-free guarantee is proved under the stated source condition, without asserting that the network must leave the linearized training regime.

Comparison of Sufficient Network Widths. Theorem 5.1 gives the sufficient condition $m \gtrsim d^{5/2} n^{25\alpha(s+2)/(2\alpha(s+2)+1)}$, whose sample-size exponent is less than $25/2$. By comparison, Suh et al. (2022, Theorem 3.11) uses $m/\log^3 m \gtrsim L^{20} n^{24}$. The proof of Li et al. (2024, Proposition 13) imposes $m^{-1/24}\sqrt{\log m} \lesssim n^{-1}$, equivalently $m/(\log m)^{12} \gtrsim n^{24}$, as one of its approximation requirements. The proposition states an unspecified polynomial condition that also depends on the inverse smallest empirical NTK eigenvalue. Consequently, our explicit sufficient width has a smaller sample-size exponent than these particular sufficient requirements. This comparison does not establish a necessary width separation. PGD also requires $N \gtrsim n^{8\alpha(s+2)/(2\alpha(s+2)+1)} \log(4n/\delta)$ auxiliary observations, so the width comparison alone does not establish a reduction in total computational cost.

7. Discussion about Extension to General Multi-Layer Neural Networks

This section gives a conditional roadmap for extending PGD to general multi-layer neural networks. The spectral construction parallels the two-layer argument, but a quantitative finite-width theorem for discrete deep-network PGD or stochastic PGD is not proved here. The main additional task is to establish kernel stability and accumulated approximation-error bounds along the preconditioned trajectory.

We use $f(\cdot, \boldsymbol{\theta})$ to denote the neural network function and $\boldsymbol{\theta}$ is the collection of all the M parameters in the neural network, and we use $\boldsymbol{\theta}_t$ to denote the parameters of the neural network at a continuous time point $t \geq 0$. $\left\{\vec{\mathbf{x}}_i, y_i\right\}_{i=1}^n$ are the training data. Then the gradient flow of the neural network by the vanilla GD is

$$\frac{\mathrm{d}\boldsymbol{\theta}_t}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) = -\frac{1}{n}\nabla_\theta f(\mathbf{X}, \boldsymbol{\theta}_t)\left(f(\mathbf{X}, \boldsymbol{\theta}_t) - \mathbf{y}\right), \quad t \geq 0, \tag{34}$$

where $f(\mathbf{X}, \boldsymbol{\theta}_t) = \left[f(\vec{\mathbf{x}}_1, \boldsymbol{\theta}_t), \ldots, f(\vec{\mathbf{x}}_n, \boldsymbol{\theta}_t)\right] \in \mathbb{R}^n$, $\mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$, and $\nabla_\theta f(\mathbf{X}, \boldsymbol{\theta}_t) \in \mathbb{R}^{M \times n}$. It follows that the gradient flow of the neural network function $f(\cdot, \boldsymbol{\theta}_t)$ is

$$\begin{aligned}\frac{\mathrm{d}f(\cdot, \boldsymbol{\theta}_t)}{\mathrm{d}t} = \nabla_{\boldsymbol{\theta}}^\top f(\cdot, \boldsymbol{\theta}_t)\frac{\mathrm{d}\boldsymbol{\theta}_t}{\mathrm{d}t} &= -\frac{1}{n}\nabla_{\boldsymbol{\theta}}^\top f(\cdot, \boldsymbol{\theta}_t)\nabla_\theta f(\mathbf{X}, \boldsymbol{\theta}_t)\left(f(\mathbf{X}, \boldsymbol{\theta}_t) - \mathbf{y}\right) \\ &= -\frac{1}{n}K_t(\cdot, \mathbf{X})\left(f(\mathbf{X}, \boldsymbol{\theta}_t) - \mathbf{y}\right),\end{aligned} \tag{35}$$

where K_t is the empirical kernel defined by

$$K_t(\mathbf{x}, \mathbf{x}') \coloneqq \nabla_{\boldsymbol{\theta}}^\top f(\mathbf{x}, \boldsymbol{\theta}_t) \nabla_\theta f(\mathbf{x}', \boldsymbol{\theta}_t), \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

Extensive existing works show the uniform convergence of K_t to the NTK, denoted as K^{NT}, of the corresponding multi-layer neural network, such as Li et al. (2024, Proposition 45).

Using PGD instead of the vanilla GD in (34), we have the gradient flow for PGD as follows,

$$\frac{\mathrm{d}\boldsymbol{\theta}_t^{(\mathrm{PGD})}}{\mathrm{d}t} = -\frac{1}{n}\mathbf{M}\nabla_\theta f(\mathbf{X}, \boldsymbol{\theta}_t^{(\mathrm{PGD})}) \left(f(\mathbf{X}, \boldsymbol{\theta}_t^{(\mathrm{PGD})}) - \mathbf{y} \right), \quad t \geq 0 \tag{36}$$

where $\boldsymbol{\theta}_t^{(\mathrm{PGD})}$ denotes the PGD parameters and $\mathbf{M} \in \mathbb{R}^{M \times M}$ is fixed at initialization. Let $\mathbf{Q} = \left\{ \vec{\mathbf{q}}_j \right\}_{j=1}^N$ be an independent auxiliary sample from the covariate distribution, and set $J_0(\mathbf{Q}) = \nabla_\theta f(\mathbf{Q}, \boldsymbol{\theta}_0) \in \mathbb{R}^{M \times N}$. In the PGD formulas below, use the mixed kernel $K_t(\mathbf{x}, \mathbf{q}) = \nabla_\theta f(\mathbf{x}, \boldsymbol{\theta}_t^{(\mathrm{PGD})})^\top \nabla_\theta f(\mathbf{q}, \boldsymbol{\theta}_0)$ and its matrix $\mathbf{K}_t(\mathbf{X}, \mathbf{Q}) \in \mathbb{R}^{n \times N}$. If the limiting NTK has Mercer expansion $K^{(\mathrm{NT})}(\mathbf{x}, \mathbf{x}\prime) = \sum_{j \geq 1} \lambda_j e_j(\mathbf{x}) e_j(\mathbf{x}\prime)$, define

$$K^{(\mathrm{NT},s)}(\mathbf{x}, \mathbf{x}\prime) = \sum_{j \geq 1} \lambda_j^s e_j(\mathbf{x}) e_j(\mathbf{x}\prime), \quad \forall \mathbf{x}, \mathbf{x}\prime \in \mathcal{X},$$

with $s \geq 1$. We also define $\mathbf{K}^{(\mathrm{NT},s)} \in \mathbb{R}^{N \times N}$ as the normalized kernel gram matrix of $K^{(\mathrm{NT},s)}$ specified by $\mathbf{K}_{ij}^{(\mathrm{NT},s)} = K^{(\mathrm{NT},s)}(\vec{\mathbf{q}}_i, \vec{\mathbf{q}}_j)/N$ for all $i, j \in [N]$. Then the preconditioner $\mathbf{M}$, similar to Algorithm 2, is computed by

$$\mathbf{M} = \frac{1}{N} J_0(\mathbf{Q}) \mathbf{K}^{(\mathrm{NT},s)} J_0(\mathbf{Q})^\top.$$

As a result, the gradient flow of the neural network trained by PGD is

$$\begin{aligned} \frac{\mathrm{d}f(\cdot, \boldsymbol{\theta}_t^{(\mathrm{PGD})})}{\mathrm{d}t} &= \nabla_{\boldsymbol{\theta}}^\top f(\cdot, \boldsymbol{\theta}_t^{(\mathrm{PGD})}) \frac{\mathrm{d}\boldsymbol{\theta}_t^{(\mathrm{PGD})}}{\mathrm{d}t} \\ &= -\frac{1}{nN} K_t(\cdot, \mathbf{Q}) \mathbf{K}^{(\mathrm{NT},s)} \mathbf{K}_t(\mathbf{X}, \mathbf{Q})^\top \left(f(\mathbf{X}, \boldsymbol{\theta}_t^{(\mathrm{PGD})}) - \mathbf{y} \right), \end{aligned} \tag{37}$$

If the mixed kernels converge uniformly to $K^{(\mathrm{NT})}$ along the PGD trajectory, and the auxiliary approximation is uniformly controlled, then the effective kernel $N^{-1} K_t(\cdot, \mathbf{Q}) \mathbf{K}^{(\mathrm{NT},s)} \mathbf{K}_t(\cdot, \mathbf{Q})^\top$ approximates the integral kernel

$$K^{(\mathrm{NT,int})}(\mathbf{x}, \mathbf{x}') \coloneqq \int_{\mathcal{X} \times \mathcal{X}} K^{(\mathrm{NT})}(\mathbf{x}, \mathbf{v}) K^{(\mathrm{NT},s)}(\mathbf{v}, \mathbf{v}') K^{(\mathrm{NT})}(\mathbf{v}', \mathbf{x}') \mathrm{d}\mu(\mathbf{v}) \otimes \mu(\mathbf{v}'), \ \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{38}$$

Writing the exact flow as the reference integral-kernel flow plus its discrepancy gives

$$\frac{\mathrm{d}f(\cdot, \boldsymbol{\theta}_t^{(\mathrm{PGD})})}{\mathrm{d}t} = -\frac{1}{n} K^{(\mathrm{NT,int})}(\cdot, \mathbf{X}) \left(f(\mathbf{X}, \boldsymbol{\theta}_t^{(\mathrm{PGD})}) - \mathbf{y} \right) + E_t, \tag{39}$$

where $E_t = -n^{-1}(\widehat{K}_t^{(\mathrm{NT,int})}(\cdot, \mathbf{X}) - K^{(\mathrm{NT,int})}(\cdot, \mathbf{X}))(f(\mathbf{X}, \boldsymbol{\theta}_t^{(\mathrm{PGD})}) - \mathbf{y})$ and $\widehat{K}_t^{(\mathrm{NT,int})} = N^{-1}K_t(\cdot, \mathbf{Q})\mathbf{K}^{(\mathrm{NT},s)}\mathbf{K}_t(\cdot, \mathbf{Q})^\top$. If the kernel discrepancy is at most ε_t uniformly, then $\|E_t\|_\infty \leq \varepsilon_t \|f(\mathbf{X}, \boldsymbol{\theta}_t^{(\mathrm{PGD})}) - \mathbf{y}\|_2/\sqrt{n}$. Controlling this error up to the stopping time requires a trajectory-wise stability theorem for PGD.

If the deep NTK has positive eigenvalues $\lambda_j \asymp j^{-2\alpha}$ and the target satisfies $f^* \in [\mathcal{H}_{K^{(\mathrm{NT})}}]^{s+2}(\mu_0)$, the corresponding integral kernel has eigenvalues λ_j^{s+2}. This identifies the candidate risk rate $n^{-2\alpha(s+2)/(2\alpha(s+2)+1)}$. Obtaining this rate for a finite deep network additionally requires the preceding PGD stability bounds, a function decomposition, and statistical error control. The value $2\alpha = d/(d-1)$ must be verified for the specific deep NTK. The two-layer covariate corollary alone does not establish it.

Relation to finite-width deep GD and SGD. Zhou et al. (2026) provide uniform NTK approximation and polynomial-width analysis for deep ReLU networks trained by discrete GD and SGD. In their source and capacity notation, $\beta = (s+2)/2$ and $\gamma = 1/(2\alpha)$ match the displayed exponents: their Corollary 6 bounds the excess risk by $\mathcal{O}(n^{-2\beta/(2\beta+\gamma)} \log^4(n/\delta))$, under its stated sample-size, width, and model assumptions. This exponent comparison does not identify their deep NTK with our two-layer kernel. Extending their tools to PGD requires simultaneous control of (1) layerwise movement under preconditioned updates, (2) mixed-kernel stability on the training and auxiliary samples, (3) auxiliary-sample approximation, and (4) discrete-time error accumulation. Stochastic PGD also requires a martingale analysis of its stochastic updates. A quantitative theorem with explicit dependence on depth, width, auxiliary-sample size, and stopping time is left for future work.

8. Simulation Results

We present simulation results in this section. The simulations are designed to validate the theoretical results in this paper: (1) improved generalization of PGD over vanilla GD with spectrally biased target functions, and (2) the scaling behavior of the early stopping time. We use spherical uniform inputs to match the theoretical setting, and we present results for linear target function and quadratic target function.

8.1 Linear Target Function

We randomly sample n points $\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n$ distributed uniformly on the unit sphere $\mathbb{S}^{49}$ in $\mathbb{R}^{50}$. The sample size n ranges from 100 to 1000 with a step size of 100. We set the target function to $f^*(\mathbf{x}) = \mathbf{s}^\top \mathbf{x}$, where $\mathbf{x} \in \mathbb{S}^{49}$ and $\mathbf{s} \sim \mathrm{Unif}(\mathcal{X})$ is randomly sampled. The variance of the noise is set to $\sigma_0^2 = 1$. We also uniformly and independently sample 1000 points on the unit sphere in $\mathbb{R}^{50}$ to serve as the test set. We train the two-layer NN (1) by Algorithm 1 with the network width $m = 10000$, $N = 8000$, and the learning rate is set to $\eta = 1$. We also set $s = 1$, so that $K^{(s)} = K$. The training is performed on an NVIDIA **A**100 GPU, and we report the test loss in Figure 1 and Figure 2. It can be observed from Figure 1 that PGD always demonstrates better generalization than the vanilla GD through lower test losses across different training data sizes. Figure 2 illustrates the test loss with respect to different numbers of PGD steps for $n = 100, 300, 500, 1000$. It can be observed from Figure 2 that early stopping consistently improves generalization for neural network training, as the test loss initially decreases and then increases due to overfitting.

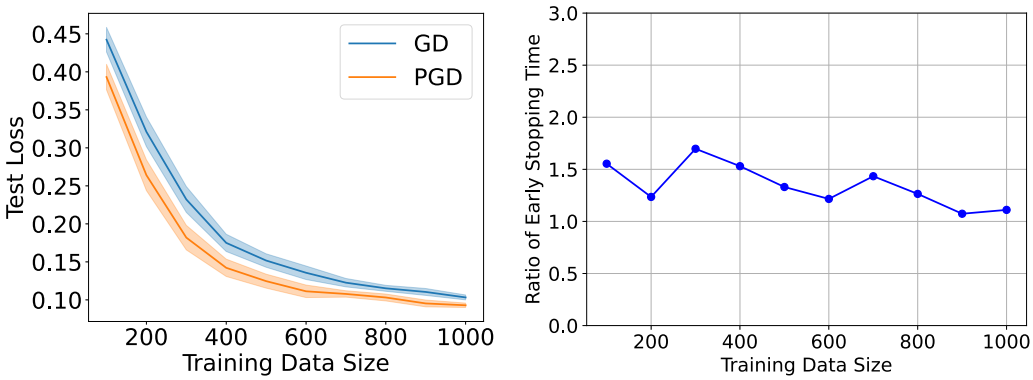


Figure 1: Left: illustration of the test losses by GD and PGD for varying n in $[100, 1000]$ with a step size of 100 for a linear target function. The shaded area in each plot indicates the standard deviation across 10 random initializations of the neural network. Right: illustration of the ratio of early stopping time.

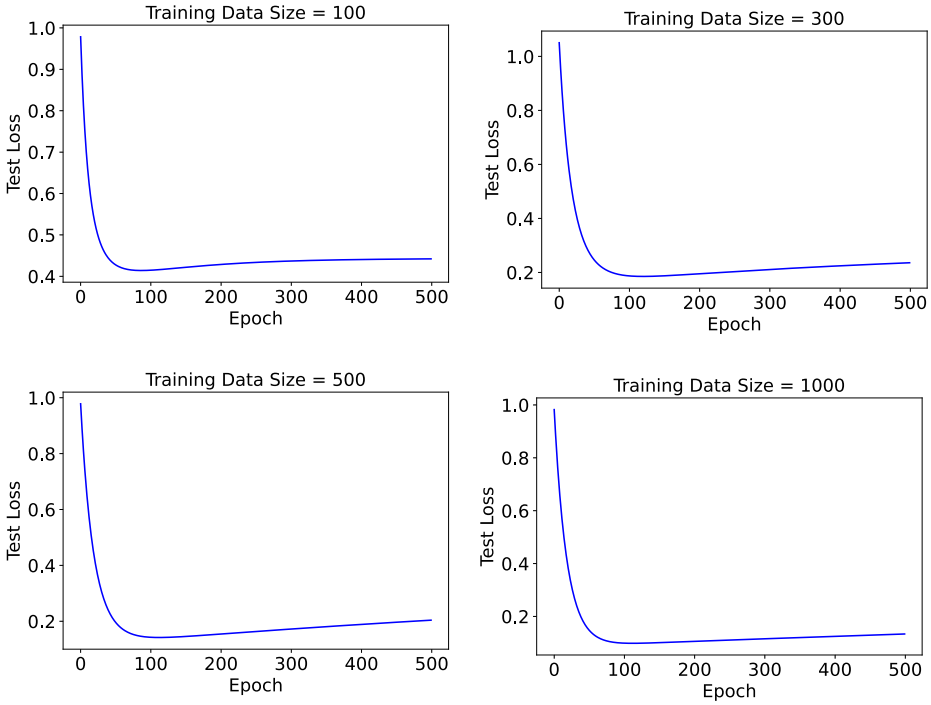


Figure 2: Illustration of the test losses by PGD for a linear target function, which are averaged over 10 random initializations of the neural network.

For each $n \in \{100, 200, \ldots, 1000\}$, we identify the PGD step that achieves the minimum test loss, denoted as $\widehat{t}_n$, which serves as the empirical early stopping time. Theoretically, the early stopping time is predicted to scale as $\widehat{T} \asymp n^{\frac{2\alpha(s+2)}{2\alpha(s+2)+1}} \asymp n^{\frac{3d}{4d-1}}$ with $s = 1$ and $2\alpha = d/(d-1)$. We compute the ratio of early stopping time, $\widehat{t}_n / n^{\frac{3d}{4d-1}}$, averaged over 10 random initializations of the neural network for each n and illustrated in the right plot

of Figure 1. We observe that the ratio of early stopping time is relatively stable and it lies within the interval $[1.0736, 1.6972]$, suggesting that the theoretically predicted early stopping time is empirically proportional to the empirical early stopping time.

8.2 Quadratic Target Function

We present additional simulation results under the same experimental setup as that in Section 8.1 except for the target function, where the target function is a quadratic function specified by $f^*(\mathbf{x}) = (\mathbf{s}^\top \mathbf{x})^2$ with $\mathbf{s} \sim \text{Unif}(\mathcal{X})$. The purpose of this study is to demonstrate the potential of PGD in learning more complex target functions than simple linear functions. Figure 3 Illustrates the test loss by GD and PGD for varying n in $[100, 1000]$ with a step size of 100. Figure 4 illustrates the test loss with respect to different numbers of PGD steps for $n = 100, 300, 500, 1000$. Similar to the case with the linear target function, it can be observed from Figure 3 and Figure 4 that PGD with early stopping consistently improves generalization for neural network training compared with the vanilla GD, and the ratio of early stopping times is stable and lies within the interval $[1.4560, 3.1001]$.

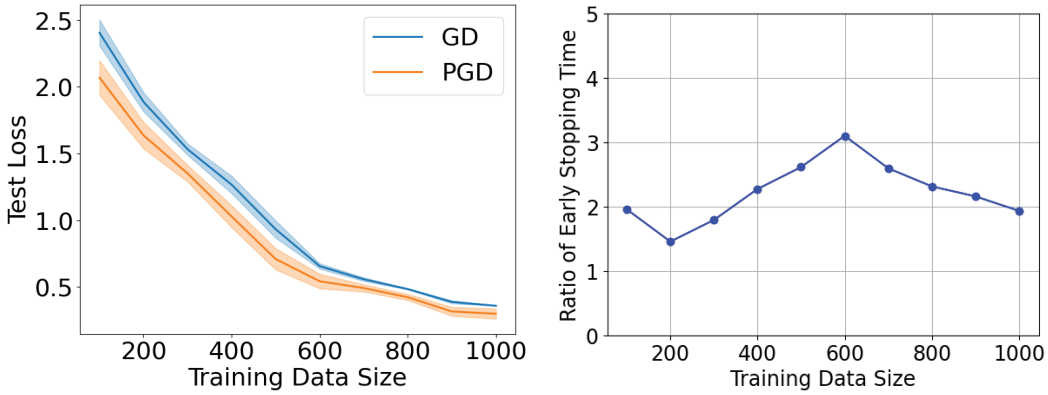


Figure 3: Left: illustration of the test losses by GD and PGD for varying n in $[100, 1000]$ with a step size of 100 for a quadratic target function. The shaded area in each plot indicates the standard deviation across 10 random initializations of the neural network. Right: illustration of the ratio of early stopping time.

8.3 Memory-Efficient PGD

We now describe a memory-efficient implementation of the PGD algorithm described in Algorithm 1. The most computationally expensive step of Algorithm 1 is the computation of $\mathbf{G}(t) := \mathbf{M}\mathbf{Z}_{\mathbf{S}}(t)(\widehat{\mathbf{y}}(t) - \mathbf{y}) \in \mathbb{R}^{md}$ in each PGD iteration (8). A straightforward computation of $\mathbf{G}(t)$ requires that a large-scale matrix $\mathbf{M} \in \mathbb{R}^{md \times md}$ be stored in memory, resulting in a potentially prohibitive memory requirement for PGD with large network width m. To address this computational challenge, we propose a memory efficient PGD algorithm described in Algorithm 3 for each iteration $t \in [T]$ of Algorithm 1. Instead of updating all the m neurons at one step, Algorithm 3 splits all the m neurons into $B = m/S$ batches where S is the batch size and m is divisible by S, and then updates the neurons in the

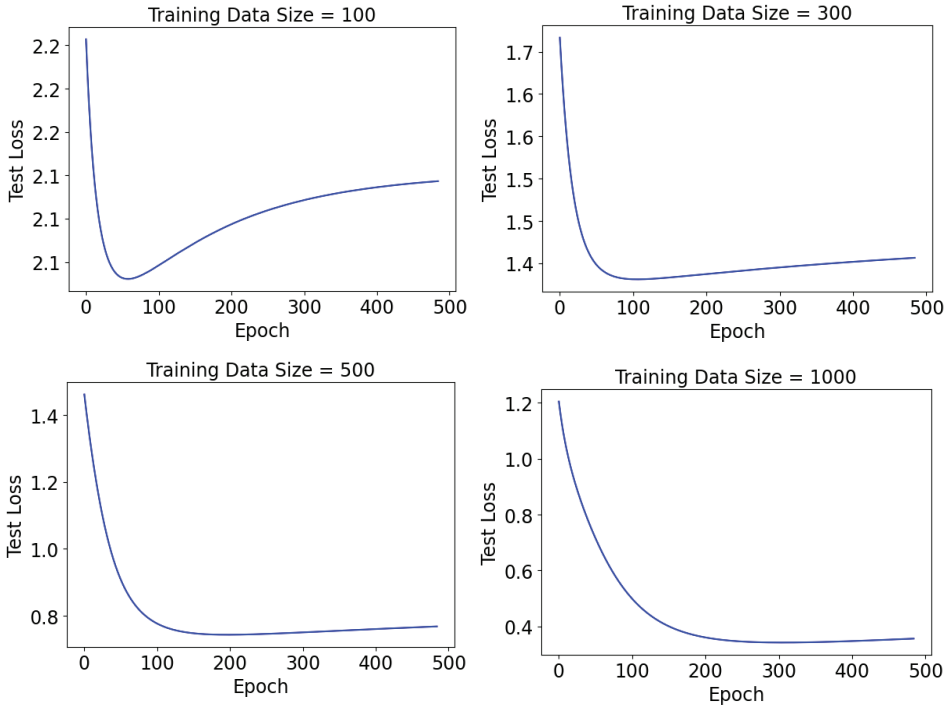


Figure 4: Illustration of the test losses by PGD for a quadratic target function, which are averaged over 10 random initializations of the neural network.

Algorithm 3 Memory-Efficient PGD

1: $\mathbf{W_S}(t+1) \leftarrow$ Memory-Efficient-PGD(t)
2: **input:** $t \in [T]$
3: **for** $b = 1, \ldots, B$ **do**
4: Update the b-th batch of the neurons by

$$[\mathrm{vec}\,(\mathbf{W_S}(t+1)) - \mathrm{vec}\,(\mathbf{W_S}(t))]_{[(b-1)Sd+1:bSd]} = -\frac{\eta}{n}\mathbf{M}_{[(b-1)Sd+1:bSd]}\mathbf{Z_S}(t)(\widehat{\mathbf{y}}(t) - \mathbf{y}) \tag{40}$$

5: **end for**
6: **return** $\mathbf{W_S}(t+1)$

b-th batch by (40) for every $b \in [B]$. In every batch update, only a submatrix of $\mathbf{Z_Q}(0)$, $\mathbf{Z_Q}(0)_{[(b-1)Sd+1:bSd]}$ formed by rows of $\mathbf{Z_Q}(0)$ with indices in $[(b-1)Sd+1 : bSd]$, is in memory, and such memory is released before the execution of the next batch update. Such batch updates of all the neurons significantly reduce the memory need. In particular, for each iteration $t \in [T]$, Algorithm 3 has the same time complexity, $\Theta(mdN)$, under $N \geq n$ and $N \ll md$, while its memory complexity is only $\Theta(SdN + N^2)$, in contrast to the significantly higher memory complexity of $\Theta(mdN)$ of a straightforward implementation of Algorithm 1 with $m \gg S$. We note that the same idea of using neuron batches is also used in

the computation of $\mathbf{Z_Q}(0)^\top \mathbf{Z_S}(t)$ in (40). On the NVIDIA A100 GPU, our memory-efficient PGD algorithm only needs 1.12 GB memory with a training speed of 9.2 seconds for each PGD iteration for the simulation results in Section 8.2. We note that the memory cost and the training speed are the same for simulation results in Section 8.1 and Section 8.2. It is also worthwhile to mention that, although each iteration of our PGD algorithm requires a higher time complexity, $\Theta(mdN)$, than that of the vanilla GD, $\Theta(mdn)$, this increased computational cost yields sharper and optimal regression rate, as discussed in Section 3 and Section 5.

9. Conclusion

We study nonparametric regression using an over-parameterized two-layer neural network when the target function lies in an interpolation space exhibiting spectral bias, that is, $[\mathcal{H}_K]^{s'}(\mu_0)$ where $s' \geq 3$ is the smoothness parameter and K is the Neural Tangent Kernel (NTK) of the network. We show that, when the network is trained using a novel and carefully designed Preconditioned Gradient Descent (PGD) algorithm with early stopping, the resulting estimator achieves a sharper and minimax optimal nonparametric regression risk rate of $\mathcal{O}(n^{-\frac{ds'}{ds'+d-1}})$. This result holds under general covariate distributions supported either on the unit sphere $\mathbb{S}^{d-1}$ or the entire $\mathbb{R}^d$ satisfying certain mild conditions, extending beyond the commonly assumed spherical uniform distribution. Our result improves upon current state-of-the-art risk bounds achieved by training neural networks for nonparametric regression. We also provide a detailed comparison with prior works and outline the key technical ideas underlying our analysis, highlighting how the proposed PGD training dynamics approximate kernel regression with the integral kernel $K^{(\mathrm{int})}$, rather than the regular NTK K, and achieve the minimax optimal risk rate over $\mathcal{H}_{K^{(\mathrm{int})}}(\mu_0)$.

We present the basic mathematical results required in our proofs in Section A, then present proofs in the subsequent sections.

Appendix A. Mathematical Tools

We introduce the basic definitions and mathematical results as the basic tools for the subsequent results in the next sections of this appendix.

Definition A.1. Let $\{\sigma_i\}_{i=1}^n$ 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

$$\mathfrak{R}(\mathcal{F}) = \mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n, \{\sigma_i\}_{i=1}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i) \right]. \tag{41}$$

The empirical Rademacher complexity is defined as

$$\widehat{\mathfrak{R}}(\mathcal{F}) = \mathbb{E}_{\{\sigma_i\}_{i=1}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i) \right], \tag{42}$$

For simplicity of notation, 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 $\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i)\right]$ 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 A.1 (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 $f \in \mathcal{F}$,$\mathrm{Var}\left[f(\vec{\mathbf{x}}_i)\right] \leq r$. Then, for every $x > 0$, with probability at least $1 - e^{-x}$,

$$\sup_{f\in\mathcal{F}} \left(\mathbb{E}_P[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_n}[f(\mathbf{x})]\right) \leq \inf_{\alpha>0} \left(2(1+\alpha)\mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n, \{\sigma_i\}_{i=1}^n}[R_n\mathcal{F}] + \sqrt{\frac{2rx}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{x}{n} \right), \tag{43}$$

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

$$\sup_{f\in\mathcal{F}} \left(\mathbb{E}_P[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_n}[f(\mathbf{x})]\right) \leq \inf_{\alpha\in(0,1)} \left(\frac{2(1+\alpha)}{1-\alpha}\mathbb{E}_{\{\sigma_i\}_{i=1}^n}[R_n\mathcal{F}] + \sqrt{\frac{2rx}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha} + \frac{1+\alpha}{2\alpha(1-\alpha)}\right)\frac{x}{n} \right). \tag{44}$$

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

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

Theorem A.2. Let ϕ be a contraction,that is, $|\phi(x) - \phi(y)| \leq \mu |x - y|$ for $\mu > 0$. Then, for every function class $\mathcal{F}$,

$$\mathbb{E}_{\{\sigma_i\}_{i=1}^n} [R_n \phi \circ \mathcal{F}] \leq \mu \mathbb{E}_{\{\sigma_i\}_{i=1}^n} [R_n \mathcal{F}], \tag{45}$$

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

Definition A.2 (Sub-root function, (Bartlett et al., 2005, Definition 3.1)). A function $\psi : [0, \infty) \to [0, \infty)$ is sub-root if it is nonnegative, nondecreasing and if $\frac{\psi(r)}{\sqrt{r}}$ is nonincreasing for $r > 0$.

Theorem A.3 (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 $T : \mathcal{F} \to \mathbb{R}+$ and some constant $\bar{B}$ such that for every $f \in \mathcal{F}$, $\mathrm{Var}[f] \leq T(f) \leq \bar{B} P(f)$. Let ψ be a sub-root function and let r^* be the fixed point of ψ. Assume that ψ satisfies, for any $r \geq r^*$, $\psi(r) \geq \bar{B} \mathfrak{R}(\{f \in \mathcal{F} : T(f) \leq r\})$. Fix $x > 0$. Then for any $K_0 > 1$, with probability at least $1 - e^{-x}$,

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

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

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

Appendix B. Detailed Proofs

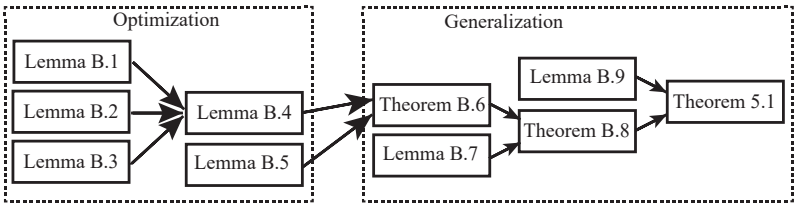


Figure 5: Roadmap of Major Results Leading to Theorem 5.1.

Because Theorem 5.1 is proved by Theorem B.8 and Lemma B.9, we illustrate in Figure 5 the roadmap containing the intermediate theoretical results which lead to Theorem B.8. Theorem B.6 is the formal version of Theorem 6.4, Theorem B.8 is the formal version of Theorem 6.5, and Lemma B.9 is the formal version of Lemma 6.6. Lemma B.4 is the formal version of Lemma 6.3.

B.1 Theorem B.8, Lemma B.9, and related theoretical results

Relation to the proof framework of Yang (2025). The high-level proof architecture in this subsection builds on Yang (2025): first decompose a finite-width neural-network iterate into an RKHS component and a uniformly small residual, and then control the resulting localized function class using local Rademacher complexity. Some general-purpose arguments are therefore inherited. Most notably, Lemma B.7 is adapted from Yang (2025, Lemma C.9). After replacing $(K, \mathcal{H}_K, \{\lambda_j\}_{j\geq 1})$ by $(K^{(\mathrm{int})}, \mathcal{H}_{K^{(\mathrm{int})}}, \{\lambda_j^{(\mathrm{int})}\}_{j\geq 1})$, its localized-Rademacher-complexity calculation is unchanged.

The arguments preceding Lemma B.7 require substantial new elements and modifications. The previous work analyzes vanilla GD, whose reference dynamics are governed by the regular NTK K. In the present paper, PGD uses a random preconditioner constructed from an independent auxiliary sample $\mathbf{Q}$, and its reference dynamics are governed by the integral kernel $K^{(\mathrm{int})}$. Consequently, the residual recursion contains not only the finite-width error caused by variation of the empirical NTK along the training trajectory, as in the previous work, but also a new approximation error associated with the auxiliary sample and its interaction with the time-dependent empirical NTK. Lemmas B.1–B.5 modify the earlier GD analysis to control these coupled errors uniformly up to the early-stopping time. In particular, the spectral filters, critical radii, and stopping rule must now be analyzed using $\mathbf{K}_n^{(\mathrm{int})}$ and the spectrum of $K^{(\mathrm{int})}$, rather than $\mathbf{K}_n$ and the spectrum of K.

The main results specific to the present setting of this paper are the uniform concentration of the empirical composite kernel around $K^{(\mathrm{int})}$ in Theorem B.10, the PGD function decomposition in Theorem B.6, and the $K^{(\mathrm{int})}$-based empirical-loss and population-risk bounds in Lemma B.9 and Theorem B.8. Theorem B.6 shows that every relevant PGD iterate can be written as a function in $\mathcal{H}_{K^{(\mathrm{int})}}$ with controlled RKHS norm plus a residual with small L^∞-norm. Lemma B.7 is then applied to this new function class. Finally, the identity $\lambda_j^{(\mathrm{int})} = \lambda_j^{s+2}$ converts the eigenvalue-tail term in Lemma B.7 into the smaller local complexity associated with $K^{(\mathrm{int})}$, which is the key spectral input leading to the sharp rate in Theorem 5.1. Therefore, compared to Yang (2025), the reusable component is the general localized Rademacher-complexity calculation, whereas the PGD dynamics, the auxiliary-sample approximation, the integral-kernel decomposition, and the resulting sharp-rate analysis are specific to the present paper and novel.

In this subsection, we first present The signal-and-noise norm bound in Lemma B.1, followed by the neuron-movement bound in Lemma B.2, the one-step perturbed PGD recursion in Lemma B.3, and finally the inductive convergence result in Lemma B.4.

Lemma B.1. Let $t \in [0\colon T]$, $\mathbf{v} = -\left(\mathbf{I} - \eta\mathbf{K}_n^{(\mathrm{int})}\right)^t f^*(\mathbf{S})$, $\mathbf{e} = -\left(\mathbf{I} - \eta\mathbf{K}_n^{(\mathrm{int})}\right)^t \mathbf{w}$, and $\eta \in (0, 8)$. Then with probability at least $1 - \exp(-\Theta(n))$ over $\mathbf{w}$,

$$\|\mathbf{v}\|_2 + \|\mathbf{e}\|_2 \leq \left(\frac{\mu_0}{\min\left\{2\sqrt{2}, \sqrt{2e\eta}\right\}} + \sigma_0 + 1\right)\sqrt{n}. \tag{46}$$

Lemma B.2. Suppose that $t \in [0 : T-1]$ for $T \geq 1$, and $\|\widehat{\mathbf{y}}(t') - \mathbf{y}\|_2 \leq \sqrt{n}c_{\mathbf{u}}$ holds for all $0 \leq t' \leq t$. Then

$$\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t') - \vec{\mathbf{w}}_r(0)\right\|_2 \leq R, \quad \forall\, 0 \leq t' \leq t+1. \tag{47}$$

Lemma B.3. Let $\eta = \Theta(1) \in (0, 8)$, $0 \le t \le T-1$ for $T \ge 1$, and suppose that $\|\widehat{\mathbf{y}}(t') - \mathbf{y}\|_2 \le c_{\mathbf{u}}\sqrt{n}$ holds for all $0 \le t' \le t$ and the random initialization $\mathbf{W}(0) \in \mathcal{W}_0$. Then

$$\widehat{\mathbf{y}}(t+1) - \mathbf{y} = \left(\mathbf{I} - \eta \mathbf{K}_n^{(\mathrm{int})}\right)(\widehat{\mathbf{y}}(t) - \mathbf{y}) + \mathbf{E}^{(\mathrm{int})}(t)\,(\widehat{\mathbf{y}}(t) - \mathbf{y}) + \mathbf{E}(t+1), \tag{48}$$

where

$$\|\mathbf{E}(t+1)\|_2 \lesssim \mathbf{E}_{m,\eta,\delta}, \tag{49}$$

and $\mathbf{E}_{m,\eta,\delta}$ is defined by

$$\begin{aligned}\mathbf{E}_{m,\eta,\delta} &\coloneqq \eta c_{\mathbf{u}}\sqrt{n}\left(2\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right) + C_1(m/2, d, 1/n)\right) \\ &\lesssim \sqrt{dn}\, m^{-\frac{1}{5}} T^{\frac{1}{2}}\end{aligned} \tag{50}$$

for $\delta \in (0,1)$. In addition, for every $\delta \in (0,1)$, with probability at least $1-\delta$ over $\mathbf{Q}$,

$$\left\|\mathbf{E}^{(\mathrm{int})}(t)\right\|_2 \lesssim \eta\sqrt{\frac{\log(4n/\delta)}{N}}. \tag{51}$$

Lemma B.4. Suppose $N \gtrsim T^2 \log(4n/\delta)/\tau^2$, and

$$m \gtrsim T^{\frac{15}{2}} d^{\frac{5}{2}}/\tau^5. \tag{52}$$

Suppose the neural network $f(\mathbf{W}(t), \cdot)$ is trained by PGD using Algorithm 1 with the learning rate $\eta = \Theta(1) \in (0, 8)$ on the random initialization $\mathbf{W}(0) \in \mathcal{W}_0$. Then for every $\delta \in (0,1)$, with probability at least $1 - \exp(-\Theta(n)) - \delta$ over $\mathbf{w}$ and $\mathbf{Q}$, $\mathbf{W}(t) \in \mathcal{W}(\mathbf{S}, \mathbf{Q}, \mathbf{W}(0), T)$. Moreover, for all $t \in [0, T]$, $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ where $\mathbf{u}(t) = \widehat{\mathbf{y}}(t) - \mathbf{y}$, $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, and $\|\mathbf{u}(t)\|_2 \le c_{\mathbf{u}}\sqrt{n}$.

Proof of Lemma B.4. First, when $m \gtrsim T^{\frac{15}{2}} d^{\frac{5}{2}}/\tau^5$ with a proper constant, it can be verified that $\mathbf{E}_{m,\eta,\delta} \le \tau\sqrt{n}/(2T)$ where $\mathbf{E}_{m,\eta,\delta}$ is defined by (50) of Lemma B.3. We then use mathematical induction to prove the lemma.

When $t = 0$, we have

$$\mathbf{u}(0) = -\mathbf{y} = \mathbf{v}(0) + \mathbf{e}(0), \tag{53}$$

where $\mathbf{v}(0) \coloneqq -f^*(\mathbf{S}) = -\left(\mathbf{I} - \eta \mathbf{K}_n^{(\mathrm{int})}\right)^0 f^*(\mathbf{S})$, $\mathbf{e}(0) = -\mathbf{w} = \vec{\mathbf{e}}_1(0) + \vec{\mathbf{e}}_2(0)$ with $\vec{\mathbf{e}}_1(0) = -\left(\mathbf{I} - \eta \mathbf{K}_n^{(\mathrm{int})}\right)^0 \mathbf{w}$ and $\vec{\mathbf{e}}_2(0) = \mathbf{0}$. 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 B.1 that $\|\mathbf{u}(0)\|_2 \le c_{\mathbf{u}}\sqrt{n}$ w.p. at least $1 - \exp(-\Theta(n))$ over $\mathbf{w}$.

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 $\mathbf{v}(t_1) \in \mathcal{V}_{t_1}$, and $\mathbf{e}(t_1) = \vec{\mathbf{e}}_1(t_1) + \vec{\mathbf{e}}_2(t_1)$ with $\mathbf{v}(t_1) \in \mathcal{V}_{t_1}$ and $\mathbf{e}(t_1) \in \mathcal{E}_{t_1,\tau}$, and $\|\mathbf{u}(t_1)\|_2 \le c_{\mathbf{u}}\sqrt{n}$ for all $t_1 \in [0, t]$. Then it follows from Lemma B.3 that the recursion $\mathbf{u}(t'+1) = \left(\mathbf{I} - \eta \mathbf{K}_n^{(\mathrm{int})}\right)\mathbf{u}(t') + \mathbf{E}^{(\mathrm{int})}(t')\mathbf{u}(t') + \mathbf{E}(t'+1)$ holds for all $t' \in [0, t]$. As a result, we have

$$\mathbf{u}(t+1) = \left(\mathbf{I} - \eta \mathbf{K}_n^{(\mathrm{int})}\right)\mathbf{u}(t) + \mathbf{E}^{(\mathrm{int})}(t)\mathbf{u}(t) + \mathbf{E}(t+1)$$

$$
\begin{aligned}
&= -\left(\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right)^{t+1} f^*(\mathbf{S}) - \left(\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right)^{t+1}\mathbf{w} \\
&\quad + \sum_{t'=1}^{t+1}\left(\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right)^{t+1-t'}\mathbf{E}(t') + \sum_{t'=1}^{t+1}\left(\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right)^{t+1-t'}\mathbf{E}^{(\mathrm{int})}(t'-1)\mathbf{u}(t'-1) \\
&= \mathbf{v}(t+1)+\mathbf{e}(t+1), \qquad (54)
\end{aligned}
$$

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

$$
\mathbf{v}(t+1) \coloneqq -\left(\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right)^{t+1} f^*(\mathbf{S}) \in \mathcal{V}_{t+1}, \qquad (55)
$$

$$
\begin{aligned}
\mathbf{e}(t+1) &\coloneqq \underbrace{-\left(\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right)^{t+1}\mathbf{w}}_{\vec{\mathbf{e}}_1(t+1)} \\
&\quad + \underbrace{\sum_{t'=1}^{t+1}\left(\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right)^{t+1-t'}\mathbf{E}(t') + \sum_{t'=1}^{t+1}\left(\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right)^{t+1-t'}\mathbf{E}^{(\mathrm{int})}(t'-1)\mathbf{u}(t'-1)}_{\vec{\mathbf{e}}_2(t+1)}. \qquad (56)
\end{aligned}
$$

We now prove the upper bound for $\vec{\mathbf{e}}_2(t+1)$. With $\eta \in (0,8)$, we have $\left\|\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right\|_2 \in (0,1]$. It follows that

$$
\begin{aligned}
&\left\|\vec{\mathbf{e}}_2(t+1)\right\|_2 \\
&\le \sum_{t'=1}^{t+1}\left\|\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right\|_2^{t+1-t'}\left\|\mathbf{E}(t')\right\|_2 + \sum_{t'=1}^{t+1}\left\|\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right\|_2^{t+1-t'}\left\|\mathbf{E}^{(\mathrm{int})}(t'-1)\right\|_2\left\|\mathbf{u}(t'-1)\right\|_2 \\
&\overset{①}{\le} \frac{\tau\sqrt{n}}{2} + c_{\mathbf{u}}\sqrt{n}\cdot T\cdot\max_{t'\in[T]}\left\|\mathbf{E}^{(\mathrm{int})}(t'-1)\right\|_2 \overset{②}{\le} \sqrt{n}\tau, \qquad (57)
\end{aligned}
$$

where ① follows from the fact that $\|\mathbf{E}(t)\|_2 \le \mathbf{E}_{m,\eta,\delta} \le \tau\sqrt{n}/(2T)$ for all $t \in [T]$, (51) in Lemma B.3, and the induction hypothesis. ② follows from $N \gtrsim T^2\log(4n/\delta)/\tau^2$ with $\max_{t'\in[T]}\left\|\mathbf{E}^{(\mathrm{int})}(t'-1)\right\|_2 \lesssim \eta\sqrt{\frac{\log(4n/\delta)}{N}}$ by Lemma B.3 and $c_{\mathbf{u}}\eta = \Theta(1)$. It follows from (57) that $\mathbf{e}(t+1) \in \mathcal{E}_{t+1,\tau}$. Also, it follows from Lemma B.1 that

$$
\begin{aligned}
\|\mathbf{u}(t+1)\|_2 &\le \|\mathbf{v}(t+1)\|_2 + \left\|\vec{\mathbf{e}}_1(t+1)\right\|_2 + \left\|\vec{\mathbf{e}}_2(t+1)\right\|_2 \\
&\le \left(\frac{\mu_0}{\sqrt{2e\eta}} + \sigma_0 + \tau + 1\right)\sqrt{n} \le c_{\mathbf{u}}\sqrt{n},
\end{aligned}
$$

which completes the induction step and the entire proof. ■

Proof of Lemma B.1. When $t \ge 1$, we have $\mathbf{v} = -\left(\mathbf{I}-\eta\mathbf{K}_n^{(\mathrm{int})}\right)^t f^*(\mathbf{S})$, and

$$
\|\mathbf{v}(t)\|_2^2 = \sum_{i=1}^{n}\left(1-\eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^{2t}\left[\mathbf{U}^{(\mathrm{int})\top} f^*(\mathbf{S})\right]_i^2
$$

$$\overset{①}{\leq} \sum_{i=1}^{\operatorname{rank}(\mathbf{K}_n^{(\mathrm{int})})} \frac{1}{2e\eta\widehat{\lambda}_i^{(\mathrm{int})}t}\left[\mathbf{U}^{(\mathrm{int})\top} f^*(\mathbf{S})\right]_i^2 \overset{②}{\leq} \frac{n\mu_0^2}{2e\eta t}. \tag{58}$$

Here ① follows from Lemma B.12, which shows that the signal has zero components in the empirical null space, and Lemma B.13 applied to the positive empirical eigenvalues. ② follows by Lemma B.12. Moreover, it follows from the concentration inequality about quadratic forms of sub-Gaussian random variables in Wright (1973) that

$$\Pr\left[\|\mathbf{w}\|_2^2 - \mathbb{E}\left[\|\mathbf{w}\|_2^2\right] > n\right] \leq \exp\left(-\Theta(n)\right), \tag{59}$$

and $\mathbb{E}\left[\|\mathbf{w}\|_2\right] \leq \sqrt{\mathbb{E}\left[\|\mathbf{w}\|_2^2\right]} \leq \sqrt{n}\sigma_0$. Therefore, $\Pr\left[\|\mathbf{w}\|_2 - \sqrt{n}\sigma_0 > \sqrt{n}\right] \leq \exp\left(-\Theta(n)\right)$. As a result, we have

$$\|\mathbf{v}\|_2 + \|\mathbf{e}\|_2 \leq \sqrt{\frac{n\mu_0^2}{2e\eta}} + \|\mathbf{w}\|_2 \leq \left(\frac{\mu_0}{\sqrt{2e\eta}} + \sigma_0 + 1\right)\sqrt{n}, \quad \forall t \in [T].$$

When $t = 0$, $\|\mathbf{v}\|_2 = \|f^*(\mathbf{S})\|_2 \leq \mu_0/(2\sqrt{2}) \cdot \sqrt{n}$, which completes the proof of (46). ■

Proof of Lemma B.3. Because $\|\widehat{\mathbf{y}}(t') - \mathbf{y}\|_2 \leq \sqrt{n}c_{\mathbf{u}}$ holds for all $t' \in [0, t]$, by Lemma B.2, we have

$$\left\|\vec{\mathbf{w}}_r(t') - \vec{\mathbf{w}}_r(0)\right\|_2 \leq R, \quad \forall\, 0 \leq t' \leq t+1. \tag{60}$$

Define two sets of indices

$$E_{i,R} \coloneqq \left\{r \in [m]\colon \left|\mathbf{w}_r(0)^\top \vec{\mathbf{x}}_i\right| > R\right\}, \quad \bar{E}_{i,R} \coloneqq [m] \setminus E_{i,R}.$$

We have

$$\begin{aligned}
\widehat{\mathbf{y}}_i(t+1) - \widehat{\mathbf{y}}_i(t) &= \frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r\left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t)\vec{\mathbf{x}}_i\right)\right) \\
&= \underbrace{\frac{1}{\sqrt{m}}\sum_{r\in E_{i,R}} a_r\left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t)\vec{\mathbf{x}}_i\right)\right)}_{\coloneqq \mathbf{D}_i^{(1)}} \\
&\quad + \underbrace{\frac{1}{\sqrt{m}}\sum_{r\in \bar{E}_{i,R}} a_r\left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t)\vec{\mathbf{x}}_i\right)\right)}_{\coloneqq \mathbf{E}_i^{(1)}} = \mathbf{D}_i^{(1)} + \mathbf{E}_i^{(1)},
\end{aligned} \tag{61}$$

and $\mathbf{D}^{(1)}, \mathbf{E}^{(1)} \in \mathbb{R}^n$ are vectors with their i-th elements being $\mathbf{D}_i^{(1)}$ and $\mathbf{E}_i^{(1)}$ defined on the RHS of (61). Now we derive the upper bound for $\mathbf{E}_i^{(1)}$. For all $i \in [n]$ we have

$$\left|\mathbf{E}_i^{(1)}\right| = \left|\frac{1}{\sqrt{m}}\sum_{r\in \bar{E}_{i,R}} a_r\left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}(t+1)^\top\vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_i\right)\right)\right|$$

$$
\begin{aligned}
&\leq \frac{1}{\sqrt{m}} \sum_{r\in\bar{E}_{i,R}} \left|\vec{\mathbf{w}}_{\mathbf{S},r}(t+1)^\top \vec{\mathbf{x}}_i - \vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i\right| \leq \frac{1}{\sqrt{m}} \sum_{r\in\bar{E}_{i,R}} \left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t)\right\|_2 \\
&\overset{①}{=} \frac{1}{\sqrt{m}} \sum_{r\in\bar{E}_{i,R}} \left\|\frac{\eta}{nN} \left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]} \mathbf{K}_N^{(s)} \mathbf{Z}_{\mathbf{Q}}(0)^\top \mathbf{Z}_{\mathbf{S}}(t) \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)\right\|_2 \\
&\overset{②}{\leq} \frac{c_{\mathbf{u}}}{\sqrt{m}} \sum_{r\in\bar{E}_{i,R}} \frac{\eta}{2\sqrt{m}} \leq \eta c_{\mathbf{u}} \cdot \frac{\left|\bar{E}_{i,R}\right|}{2m}.
\end{aligned} \tag{62}
$$

Here ①, ② follow from (81) and (82) in the proof of Lemma B.2, as well as the fact that $\left\|\mathbf{K}_N^{(s)}\right\|_2 \leq 1/2$. Since $\mathbf{W}(0) \in \mathcal{W}_0$, we have

$$
\sup_{\mathbf{x}\in\mathcal{X}} \widehat{v}_R(\mathbf{W}(0), \mathbf{x}) \leq \frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n), \tag{63}
$$

where $\widehat{v}_R(\mathbf{W}(0), \mathbf{x}) = \frac{1}{m}\sum_{r=1}^m \mathbb{1}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \mathbf{x}\right| \leq R\right\}}$, so that $\widehat{v}_R(\mathbf{W}(0), \vec{\mathbf{x}}_i) = \left|\bar{E}_{i,R}\right|/m$. It follows from (62), (63) and the induction hypothesis that

$$
\left|\mathbf{E}_i^{(1)}\right| \leq \frac{\eta c_{\mathbf{u}}}{2} \left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right). \tag{64}
$$

It follows from (64) that $\left\|\mathbf{E}^{(1)}\right\|_2$ can be bounded by

$$
\left\|\mathbf{E}^{(1)}\right\|_2 \leq \frac{\eta c_{\mathbf{u}}\sqrt{n}}{2} \left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)\right). \tag{65}
$$

$\mathbf{D}_i^{(1)}$ on the RHS of (61) is expressed by

$$
\begin{aligned}
\mathbf{D}_i^{(1)} &= \frac{1}{\sqrt{m}} \sum_{r\in E_{i,R}} a_r \left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\vec{\mathbf{x}}_i\right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t)\vec{\mathbf{x}}_i\right)\right) \\
&= \frac{1}{\sqrt{m}} \sum_{r\in E_{i,R}} a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(\vec{\mathbf{w}}_{\mathbf{S},r}(t+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t)\right)^\top \vec{\mathbf{x}}_i \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(-\frac{\eta}{n} \left[\mathbf{M}\right]_{[(r-1)d+1:rd]} \mathbf{Z}_{\mathbf{S}}(t) \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)\right)^\top \vec{\mathbf{x}}_i \\
&\quad + \frac{1}{\sqrt{m}} \sum_{r\in\bar{E}_{i,R}} a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(\frac{\eta}{n} \left[\mathbf{M}\right]_{[(r-1)d+1:rd]} \mathbf{Z}_{\mathbf{S}}(t) \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)\right)^\top \vec{\mathbf{x}}_i \\
&= \underbrace{-\frac{\eta}{nN} \left[\mathbf{H}(t)\right]_i \mathbf{K}_N^{(s)} \mathbf{H}(t)^\top \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)}_{:=\mathbf{D}_i^{(2)}} \\
&\quad \underbrace{+ \frac{1}{\sqrt{m}} \sum_{r\in\bar{E}_{i,R}} a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(\frac{\eta}{n} \left[\mathbf{M}\right]_{[(r-1)d+1:rd]} \mathbf{Z}_{\mathbf{S}}(t) \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)\right)^\top \vec{\mathbf{x}}_i}_{:=\mathbf{E}_i^{(2)}},
\end{aligned} \tag{66}
$$

where $\mathbf{H}(t) \in \mathbb{R}^{n\times N}$ is a matrix specified by

$$\mathbf{H}_{pq}(t) = \frac{\vec{\mathbf{x}}_p^\top \vec{\mathbf{q}}_q}{m} \sum_{r=1}^{m} \mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_p \geq 0\right\}} \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{q}}_q \geq 0\right\}}, \quad \forall\, p \in [n], q \in [N].$$

Let $\mathbf{D}^{(2)}, \mathbf{E}^{(2)} \in \mathbb{R}^n$ be vectors with their i-th elements being $\mathbf{D}_i^{(2)}$ and $\mathbf{E}_i^{(2)}$ defined on the RHS of (66). $\mathbf{E}^{(2)}$ can be expressed by $\mathbf{E}^{(2)} = \frac{\eta}{nN}\tilde{\mathbf{E}}^{(2)}\mathbf{K}_N^{(s)}\mathbf{H}(t)^\top(\widehat{\mathbf{y}}(t) - \mathbf{y})$ with $\tilde{\mathbf{E}}^{(2)} \in \mathbb{R}^{n\times N}$ and

$$\tilde{\mathbf{E}}_{pq}^{(2)} = \frac{1}{m} \sum_{r\in \bar{E}_{p,R}} \mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_p \geq 0\right\}} \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{q}}_q \geq 0\right\}} \vec{\mathbf{x}}_p^\top \vec{\mathbf{q}}_q, \qquad \left|\tilde{\mathbf{E}}_{pq}^{(2)}\right| \leq \frac{\left|\bar{E}_{p,R}\right|}{m}$$

for all $p \in [n], q \in [N]$. The spectral norm of $\tilde{\mathbf{E}}^{(2)}$ is bounded by

$$\left\|\tilde{\mathbf{E}}^{(2)}\right\|_2 \leq \left\|\tilde{\mathbf{E}}^{(2)}\right\|_{\mathrm{F}} \leq \sqrt{nN} \max_{p\in[n]} \frac{\left|\bar{E}_{p,R}\right|}{m} \overset{①}{\leq} \sqrt{nN}\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right), \tag{67}$$

where ① follows from (63). Also, $\|\mathbf{H}(t)\|_2 \leq \|\mathbf{H}(t)\|_{\mathrm{F}} \leq \sqrt{nN}$ for all $t \geq 0$. It follows from (67) that $\left\|\mathbf{E}^{(2)}\right\|_2$ can be bounded by

$$\left\|\mathbf{E}^{(2)}\right\|_2 \leq \frac{\eta}{nN}\left\|\tilde{\mathbf{E}}^{(2)}\right\|_2 \left\|\mathbf{K}_N^{(s)}\right\|_2 \|\mathbf{H}(t)\|_2 \|\widehat{\mathbf{y}}(t) - \mathbf{y}\|_2 \leq \frac{\eta c_{\mathbf{u}}\sqrt{n}}{2}\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right). \tag{68}$$

$\mathbf{D}_i^{(2)}$ on the RHS of (66) is expressed by

$$\begin{aligned}
\mathbf{D}^{(2)} &= -\frac{\eta}{nN}\mathbf{H}(t)\mathbf{K}_N^{(s)}\mathbf{H}(t)^\top(\widehat{\mathbf{y}}(t) - \mathbf{y}) \\
&= \underbrace{-\frac{\eta}{nN}\mathbf{H}(0)\mathbf{K}_N^{(s)}\mathbf{H}(0)^\top(\widehat{\mathbf{y}}(t) - \mathbf{y})}_{:=\mathbf{D}^{(3)}} + \underbrace{\frac{\eta}{nN}\left(\mathbf{H}(0) - \mathbf{H}(t)\right)\mathbf{K}_N^{(s)}\mathbf{H}(0)^\top(\widehat{\mathbf{y}}(t) - \mathbf{y})}_{:=\mathbf{E}^{(3)}} \\
&\quad + \underbrace{\frac{\eta}{nN}\mathbf{H}(t)\mathbf{K}_N^{(s)}\left(\mathbf{H}(0) - \mathbf{H}(t)\right)^\top(\widehat{\mathbf{y}}(t) - \mathbf{y})}_{:=\mathbf{E}^{(4)}} = \mathbf{D}^{(3)} + \mathbf{E}^{(3)} + \mathbf{E}^{(4)}.
\end{aligned} \tag{69}$$

On the RHS of (69), $\mathbf{D}^{(3)}, \mathbf{E}^{(3)}, \mathbf{E}^{(4)} \in \mathbb{R}^n$ are vectors which are analyzed as follows. In order to bound the spectral norm of $\mathbf{E}^{(3)}$ and $\mathbf{E}^{(4)}$, we first estimate the upper bound for $|\mathbf{H}_{ij}(t) - \mathbf{H}_{ij}(0)|$ for all $i \in [n]$ and $j \in [N]$. We note that

$$\mathbb{I}_{\left\{\mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \neq \mathbb{I}_{\left\{\mathbf{w}_r(0)^\top \vec{\mathbf{x}}_i \geq 0\right\}}\right\}} \leq \mathbb{I}_{\left\{\left|\mathbf{w}_r(0)^\top \vec{\mathbf{x}}_i\right| \leq R\right\}} + \mathbb{I}_{\left\{\left\|\mathbf{w}_{\mathbf{S},r}(t) - \vec{\mathbf{w}}_r(0)\right\|_2 > R\right\}}. \tag{70}$$

It follows from (70) that

$$\begin{aligned}
&|\mathbf{H}_{ij}(t) - \mathbf{H}_{ij}(0)| \\
&= \left|\frac{\vec{\mathbf{x}}_i^\top \vec{\mathbf{q}}_j}{m} \sum_{r=1}^{m}\left(\mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{q}}_j \geq 0\right\}} - \mathbb{I}_{\left\{\mathbf{w}_r(0)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \mathbb{I}_{\left\{\mathbf{w}_r(0)^\top \vec{\mathbf{q}}_j \geq 0\right\}}\right)\right|
\end{aligned}$$

$$\leq \frac{1}{m}\sum_{r=1}^{m} \mathbb{1}_{\left\{\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \neq \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_i \geq 0\right\}}\right\}}$$
$$\leq \frac{1}{m}\sum_{r=1}^{m}\left(\mathbb{1}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_i\right| \leq R\right\}} + \mathbb{1}_{\left\{\left\|\mathbf{w}_{\mathbf{S},r}(t) - \vec{\mathbf{w}}_r(0)\right\|_2 > R\right\}}\right)$$
$$\leq \widehat{v}_R(\mathbf{W}(0), \vec{\mathbf{x}}_i) \overset{①}{\leq} \frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n), \tag{71}$$

where ① follows from (63). It follows from (71) that $\left\|\mathbf{E}^{(3)}\right\|_2, \left\|\mathbf{E}^{(4)}\right\|_2$ are bounded by

$$\left\|\mathbf{E}^{(3)}\right\|_2 \leq \frac{\eta}{2nN} \cdot \sqrt{nN} \cdot \sqrt{nN}\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right) \|\widehat{\mathbf{y}}(t) - \mathbf{y}\|_2$$
$$\leq \frac{\eta c_{\mathbf{u}}\sqrt{n}}{2}\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right), \tag{72}$$
$$\left\|\mathbf{E}^{(4)}\right\|_2 \leq \frac{\eta c_{\mathbf{u}}\sqrt{n}}{2}\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right). \tag{73}$$

We now approximate $\mathbf{D}^{(3)}$ in terms of $\mathbf{K}_{\mathbf{S},\mathbf{Q}} \in \mathbb{R}^{n\times N}$ with $[\mathbf{K}_{\mathbf{S},\mathbf{Q}}]_{ij} = K(\vec{\mathbf{x}}_i, \vec{\mathbf{q}}_j)$ for $i \in [n], j \in [N]$. We have

$$\mathbf{D}^{(3)} = -\frac{\eta}{nN}\mathbf{H}(0)\mathbf{K}_N^{(s)}\mathbf{H}(0)^\top(\widehat{\mathbf{y}}(t) - \mathbf{y})$$
$$= \underbrace{-\frac{\eta}{nN}\mathbf{K}_{\mathbf{S},\mathbf{Q}}\mathbf{K}_N^{(s)}\mathbf{K}_{\mathbf{S},\mathbf{Q}}^\top(\widehat{\mathbf{y}}(t) - \mathbf{y})}_{:=\mathbf{D}^{(4)}} + \underbrace{\frac{\eta}{nN}(\mathbf{K}_{\mathbf{S},\mathbf{Q}} - \mathbf{H}(0))\mathbf{K}_N^{(s)}\mathbf{K}_{\mathbf{S},\mathbf{Q}}^\top(\widehat{\mathbf{y}}(t) - \mathbf{y})}_{:=\mathbf{E}^{(5)}}$$
$$+ \underbrace{\frac{\eta}{nN}\mathbf{H}(0)\mathbf{K}_N^{(s)}(\mathbf{K}_{\mathbf{S},\mathbf{Q}} - \mathbf{H}(0))^\top(\widehat{\mathbf{y}}(t) - \mathbf{y})}_{:=\mathbf{E}^{(6)}}. \tag{74}$$

We now bound the spectral norm of $\mathbf{E}^{(5)}$ and $\mathbf{E}^{(6)}$. Since $\mathbf{W}(0) \in \mathcal{W}_0$, we have

$$\|\mathbf{K}_{\mathbf{S},\mathbf{Q}} - \mathbf{H}(0)\|_2 \leq \|\mathbf{K}_{\mathbf{S},\mathbf{Q}} - \mathbf{H}(0)\|_{\mathrm{F}} \leq \sqrt{nN}C_1(m/2, d, 1/n), \tag{75}$$

Also, $\|\mathbf{K}_{\mathbf{S},\mathbf{Q}}\|_2 \leq \|\mathbf{K}_{\mathbf{S},\mathbf{Q}}\|_{\mathrm{F}} \leq \sqrt{nN}$. As a result,

$$\left\|\mathbf{E}^{(5)}\right\|_2 \leq \frac{\eta}{2nN} \cdot \sqrt{nN} \cdot \sqrt{nN}C_1(m/2, d, 1/n)\|\widehat{\mathbf{y}}(t) - \mathbf{y}\|_2 \leq \frac{\eta c_{\mathbf{u}}\sqrt{n}}{2} \cdot C_1(m/2, d, 1/n), \tag{76}$$

$$\left\|\mathbf{E}^{(6)}\right\|_2 \leq \frac{\eta c_{\mathbf{u}}\sqrt{n}}{2} \cdot C_1(m/2, d, 1/n). \tag{77}$$

It follows from (66), (69), and (74) that

$$\mathbf{D}_i^{(1)} = \mathbf{D}_i^{(4)} + \mathbf{E}_i^{(2)} + \mathbf{E}_i^{(3)} + \mathbf{E}_i^{(4)} + \mathbf{E}_i^{(5)} + \mathbf{E}_i^{(6)}. \tag{78}$$

It then follows from (61) that

$$\widehat{\mathbf{y}}_i(t+1) - \widehat{\mathbf{y}}_i(t) = \mathbf{D}_i^{(1)} + \mathbf{E}_i^{(1)} = \mathbf{D}_i^{(4)} + \underbrace{\mathbf{E}_i^{(1)} + \mathbf{E}_i^{(2)} + \mathbf{E}_i^{(3)} + \mathbf{E}_i^{(4)} + \mathbf{E}_i^{(5)} + \mathbf{E}_i^{(6)}}_{:=\mathbf{E}_i}$$

$$= -\frac{\eta}{nN}\left[\mathbf{K}_{\mathbf{S},\mathbf{Q}}\right]_i \mathbf{K}_N^{(s)} \mathbf{K}_{\mathbf{S},\mathbf{Q}}^\top \left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right) + \left[\mathbf{E}(t+1)\right]_i = -\frac{\eta}{n}\left[\widehat{\mathbf{K}}^{(\mathrm{int})}\right]_i \left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right) + \left[\mathbf{E}(t+1)\right]_i, \tag{79}$$

where $\mathbf{E}(t+1) \in \mathbb{R}^n$ with its i-th element being $\mathbf{E}_i$, and $\mathbf{E}(t+1) = \mathbf{E}^{(1)} + \mathbf{E}^{(2)} + \mathbf{E}^{(3)} + \mathbf{E}^{(4)} + \mathbf{E}^{(5)} + \mathbf{E}^{(6)}$. It follows from (65), (68), (72), (73), (76), and (77) that

$$\left\|\mathbf{E}(t+1)\right\|_2 \lesssim \eta c_{\mathbf{u}} \sqrt{n}\left(2\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2,d,1/n)\right) + C_1(m/2,d,1/n)\right). \tag{80}$$

The upper bound for $\left\|\mathbf{E}(t+1)\right\|_2$ in (50) follows from (80), Theorem 6.1, and noting that $\eta c_{\mathbf{u}} \leq \Theta(1)$. Finally, we have

$$\begin{aligned}
\widehat{\mathbf{y}}(t+1) - \mathbf{y} &\overset{①}{=} \left(\mathbf{I} - \frac{\eta}{n}\widehat{\mathbf{K}}^{(\mathrm{int})}\right)\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right) + \mathbf{E}(t+1) \\
&= \left(\mathbf{I} - \frac{\eta}{n}\mathbf{K}^{(\mathrm{int})}\right)\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right) + \mathbf{E}^{(\mathrm{int})}(t)\left(\widehat{\mathbf{y}}(t)-\mathbf{y}\right) + \mathbf{E}(t+1),
\end{aligned}$$

where $\mathbf{E}^{(\mathrm{int})}(t) = \eta/n \cdot \left(\mathbf{K}^{(\mathrm{int})} - \widehat{\mathbf{K}}^{(\mathrm{int})}\right)$, and ① follows from (79). It follows from (133) of Theorem B.10 that

$$\left\|\mathbf{E}^{(\mathrm{int})}(t)\right\|_2 \lesssim \eta\sqrt{\frac{\log(4n/\delta)}{N}}$$

holds with probability at least $1-\delta$ over $\mathbf{Q}$, which proves (51). ■

Proof of Lemma B.2. Let $[\mathbf{Z}_{\mathbf{Q}}(0)]_{[(r-1)d+1:rd]}$ denote the submatrix of $\mathbf{Z}_{\mathbf{Q}}(0)$ formed by the rows of $\mathbf{Z}_{\mathbf{Q}}(0)$ with row indices in $[(r-1)d+1:rd]$. By the PGD update rule (8), for every $t'' \in [0:T-1]$, we have

$$\vec{\mathbf{w}}_{\mathbf{S},r}(t''+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t'') = -\frac{\eta}{nN}\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]} \mathbf{K}_N^{(s)} \mathbf{Z}_{\mathbf{Q}}(0)^\top \mathbf{Z}_{\mathbf{S}}(t'')\left(\widehat{\mathbf{y}}(t'')-\mathbf{y}\right). \tag{81}$$

Since $\left\|\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\right\|_2 \leq \left\|\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\right\|_{\mathrm{F}} \leq \sqrt{N/m}$, and $\left\|\mathbf{Z}_{\mathbf{Q}}(0)^\top \mathbf{Z}_{\mathbf{S}}(t)\right\|_2 \leq \sqrt{nN}$, it then follows from (81) that

$$\begin{aligned}
&\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t''+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t'')\right\|_2 \\
&\leq \frac{\eta}{nN}\left\|\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\right\|_2 \left\|\mathbf{K}_N^{(s)}\right\|_2 \left\|\mathbf{Z}_{\mathbf{Q}}(0)^\top \mathbf{Z}_{\mathbf{S}}(t'')\right\|_2 \left\|\widehat{\mathbf{y}}(t'')-\mathbf{y}\right\|_2 \leq \frac{\eta c_{\mathbf{u}}}{2\sqrt{m}},\ \forall t'' \in [0:t].
\end{aligned} \tag{82}$$

Note that (47) trivially holds for $t'=0$. For $t' \in [1, t+1]$, it follows from (82) that

$$\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t') - \vec{\mathbf{w}}_r(0)\right\|_2 \leq \sum_{t''=0}^{t'-1}\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t''+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t'')\right\|_2 \leq \frac{\eta}{2\sqrt{m}}\sum_{t''=0}^{t'-1} c_{\mathbf{u}} \leq \frac{\eta c_{\mathbf{u}} T}{2\sqrt{m}} = R, \tag{83}$$

which completes the proof. ■

Lemma B.5. Let $\eta \in (0, 8)$ and $h(\cdot) = \sum_{t'=0}^{t-1} h(\cdot, t')$ for $t \in [T]$, $T \leq \widehat{T}$ where

$$h(\cdot, t') = v(\cdot, t') + \widehat{e}(\cdot, t'),$$
$$v(\cdot, t') = \frac{\eta}{n} \sum_{j=1}^{n} K^{(\mathrm{int})}(\vec{\mathbf{x}}_j, \mathbf{x}) \mathbf{v}_j(t'),$$
$$\widehat{e}(\cdot, t') = \frac{\eta}{n} \sum_{j=1}^{n} K^{(\mathrm{int})}(\vec{\mathbf{x}}_j, \mathbf{x}) \vec{\mathbf{e}}_j(t'),$$

and $\mathbf{v}(t') \in \mathcal{V}_{t'}$, $\mathbf{e}(t') \in \mathcal{E}_{t',\tau}$ for all $0 \leq t' \leq t-1$. Suppose that

$$\tau \lesssim 1/(\eta T). \tag{84}$$

Then with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over $\mathbf{w}$,

$$\|h\|_{\mathcal{H}_{K^{(\mathrm{int})}}} \leq B_h = \mu_0 + 1 + \sqrt{2}, \tag{85}$$

where B_h is also defined in (26).

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

$$\widehat{e}_1(\cdot, t) = \frac{\eta}{n} \sum_{j=1}^{n} K^{(\mathrm{int})}(\vec{\mathbf{x}}_j, \mathbf{x}) \left[\vec{\mathbf{e}}_1(t)\right]_j, \quad \widehat{e}_2(\cdot, t) = \frac{\eta}{n} \sum_{j=1}^{n} K^{(\mathrm{int})}(\vec{\mathbf{x}}_j, \mathbf{x}) \left[\vec{\mathbf{e}}_2(t)\right]_j,$$

we then have

$$\begin{aligned}
\sum_{t'=0}^{t-1} v(\mathbf{x}, t') &= -\frac{\eta}{n} \sum_{j=1}^{n} \sum_{t'=0}^{t-1} \left[\left(\mathbf{I} - \eta \mathbf{K}_n^{(\mathrm{int})}\right)^{t'} f^*(\mathbf{S})\right]_j K^{(\mathrm{int})}(\vec{\mathbf{x}}_j, \mathbf{x}) \\
&= -\frac{\eta}{n} \sum_{j=1}^{n} \sum_{t'=0}^{t-1} \left[\mathbf{U}^{(\mathrm{int})} \left(\mathbf{I} - \eta \mathbf{\Sigma}^{(\mathrm{int})}\right)^{t'} {\mathbf{U}^{(\mathrm{int})}}^\top f^*(\mathbf{S})\right]_j K^{(\mathrm{int})}(\vec{\mathbf{x}}_j, \mathbf{x}).
\end{aligned} \tag{86}$$

We let $r = \mathrm{rank}(\mathbf{K}_n^{(\mathrm{int})}) \in [n]$. It follows from (86) that

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

$$\leq \frac{1}{n}\sum_{i=1}^{r}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^2}{\widehat{\lambda}_i^{(\mathrm{int})}}\left[\mathbf{U}^{(\mathrm{int})\top}f^*(\mathbf{S})\right]_i^2 \leq \mu_0^2, \tag{87}$$

where the last inequality follows from Lemma B.12.

Similarly, we have

$$\left\|\sum_{t'=0}^{t-1}\widehat{e}_1(\cdot,t')\right\|_{\mathcal{H}_{K^{(\mathrm{int})}}}^2 \leq \frac{1}{n}\sum_{i=1}^{r}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^2}{\widehat{\lambda}_i^{(\mathrm{int})}}\left[\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\right]_i^2. \tag{88}$$

It then follows from the argument in the proof of (Raskutti et al., 2014, Lemma 9) that the RHS of (88) is bounded with high probability. We define a diagonal matrix $\mathbf{R}\in\mathbb{R}^{n\times n}$ with $\mathbf{R}_{ii}=\big(1-(1-\eta\widehat{\lambda}_i^{(\mathrm{int})})^t\big)^2/\widehat{\lambda}_i^{(\mathrm{int})}$ for $i\in[r]$ and $\mathbf{R}_{ii}=0$ for $i\in[r+1,n]$. Then the RHS of (88) is $1/n\cdot\mathrm{tr}\left(\mathbf{U}^{(\mathrm{int})}\mathbf{R}\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\mathbf{w}^\top\right)$. It follows from (Wright, 1973) that

$$\begin{aligned}&\Pr\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}^{(\mathrm{int})}\mathbf{R}\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\mathbf{w}^\top\right)-\mathbb{E}\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}^{(\mathrm{int})}\mathbf{R}\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\mathbf{w}^\top\right)\right]\geq u\right]\\&\leq\exp\left(-c\min\left\{nu/\|\mathbf{R}\|_2,n^2u^2/\|\mathbf{R}\|_{\mathrm{F}}^2\right\}\right)\end{aligned} \tag{89}$$

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

$$\begin{aligned}&\mathbb{E}\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}^{(\mathrm{int})}\mathbf{R}\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\mathbf{w}^\top\right)\right]\leq\frac{\sigma_0^2}{n}\sum_{i=1}^{r}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^2}{\widehat{\lambda}_i^{(\mathrm{int})}}\\&\overset{①}{\leq}\frac{\sigma_0^2}{n}\sum_{i=1}^{r}\min\left\{\frac{1}{\widehat{\lambda}_i^{(\mathrm{int})}},\eta_t^2\widehat{\lambda}_i^{(\mathrm{int})}\right\}\leq\frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^{r}\min\left\{\frac{1}{\eta_t\widehat{\lambda}_i^{(\mathrm{int})}},\eta_t\widehat{\lambda}_i^{(\mathrm{int})}\right\}\\&\overset{②}{\leq}\frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^{r}\min\left\{1,\eta_t\widehat{\lambda}_i^{(\mathrm{int})}\right\}=\frac{\sigma_0^2\eta_t^2}{n}\sum_{i=1}^{r}\min\left\{\eta_t^{-1},\widehat{\lambda}_i^{(\mathrm{int})}\right\}\leq\sigma_0^2\eta_t^2\widehat{R}_{K^{(\mathrm{int})}}^2(\sqrt{1/\eta_t})\leq 1.\end{aligned} \tag{90}$$

Here ① follows from the fact that $(1-\eta\widehat{\lambda}_i^{(\mathrm{int})})^t\geq\max\left\{0,1-t\eta\widehat{\lambda}_i^{(\mathrm{int})}\right\}$, 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 $\widehat{R}_{K^{(\mathrm{int})}}(\sqrt{1/\eta_t})\leq 1/(\sigma_0\eta_t)$, 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

$$\|\mathbf{R}\|_2\leq\max_{i\in[r]}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^2}{\widehat{\lambda}_i^{(\mathrm{int})}}\leq\max_{i\in[r]}\min\left\{\frac{1}{\widehat{\lambda}_i^{(\mathrm{int})}},\eta_t^2\widehat{\lambda}_i^{(\mathrm{int})}\right\}\leq\eta_t. \tag{91}$$

We also have

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

$$\leq \frac{\eta_t^3}{n}\sum_{i=1}^{r}\min\left\{\widehat{\lambda}_i^{(\mathrm{int})},\frac{1}{\eta_t}\right\} \leq \eta_t^3 \widehat{R}^2_{K^{(\mathrm{int})}}(\sqrt{1/\eta_t}) \leq \frac{\eta_t}{\sigma_0^2}. \tag{92}$$

Combining (88)-(92) with $u=1$ in (89), we have

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

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 w.p. at least $1-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$, $\left\|\sum_{t'=0}^{t-1}\widehat{e}_1(\cdot,t')\right\|^2_{\mathcal{H}_{K^{(\mathrm{int})}}}\leq 2$. This event can be chosen simultaneously for $t\in[T]$: the nonnegative diagonal filter $[1-(1-\eta\widehat{\lambda}_i^{(\mathrm{int})})^t]^2/\widehat{\lambda}_i^{(\mathrm{int})}$ is nondecreasing in t on the positive spectrum, so the quadratic form at T dominates all earlier forms.

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

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

so that

$$\left\|\sum_{t'=0}^{t-1}\widehat{e}_2(\cdot,t')\right\|_{\mathcal{H}_{K^{(\mathrm{int})}}}\leq\sum_{t'=0}^{t-1}\left\|\widehat{e}_2(\cdot,t')\right\|_{\mathcal{H}_{K^{(\mathrm{int})}}}\lesssim\sqrt{\widehat{\lambda}_1^{(\mathrm{int})}}\eta T\tau\leq 1, \tag{93}$$

since $\tau\lesssim 1/(\eta T)$. Finally, we have

$$\begin{aligned}\|h\|_{\mathcal{H}_{K^{(\mathrm{int})}}}&\leq\left\|\sum_{t'=0}^{t-1}v(\cdot,t')\right\|_{\mathcal{H}_{K^{(\mathrm{int})}}}+\left\|\sum_{t'=0}^{t-1}\widehat{e}_1(\cdot,t')\right\|_{\mathcal{H}_{K^{(\mathrm{int})}}}+\left\|\sum_{t'=0}^{t-1}\widehat{e}_2(\cdot,t')\right\|_{\mathcal{H}_{K^{(\mathrm{int})}}}\\ &\leq\mu_0+1+\sqrt{2}.\end{aligned}$$

■

Theorem B.6. Suppose $w\in(0,1)$,

$$m\gtrsim\max\left\{T^{\frac{15}{2}}d^{\frac{5}{2}}/w^5, T^{\frac{25}{2}}d^{\frac{5}{2}}\right\}, \tag{94}$$

$$N\gtrsim\max\left\{T^2\log(4n/\delta)/w^2, T^4\log(4n/\delta)\right\}, \tag{95}$$

and the neural network $f_t(\cdot)=f(\mathbf{W}(t),\cdot)$ is trained by PGD in Algorithm 1 with the learning rate $\eta=\Theta(1)\in(0,8)$ and the random initialization $\mathbf{W}(0)\in\mathcal{W}_0$. Then for every $t\in[T]$ with $T\leq\widehat{T}$ and every $\delta\in(0,1)$, w.p. at least $1-\exp(-\Theta(n))-\delta-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$ and the random sample $\mathbf{Q}$, $f_t\in\mathcal{F}_{\mathrm{NN}}(\mathbf{S},\mathbf{Q},\mathbf{W}(0),T)$, and f_t can be decomposed by

$$f_t=h+e\in\mathcal{F}_{\mathrm{ext}}(B_h,w), \tag{96}$$

where $h\in\mathcal{H}_{K^{(\mathrm{int})}}(B_h)$ with B_h defined in (26), $e\in L^\infty$ with $\|e\|_\infty\leq w$. In addition, $\|f_t\|_\infty\leq\frac{B_h}{2\sqrt{2}}+w$.

Proof It follows from Lemma B.4 and its proof that conditioned on an event with probability at least $1-\exp(-\Theta(n))-\delta$ over $\mathbf{w}$ and $\mathbf{Q}$, $f \in \mathcal{F}_{\mathrm{NN}}(\mathbf{S},\mathbf{Q},\mathbf{W}(0),T)$ with $\mathbf{W}(0)\in\mathcal{W}_0$. Moreover, $f(\cdot)=f(\mathbf{W},\cdot)$ with $\mathbf{W}=\left\{\vec{\mathbf{w}}_r\right\}_{r=1}^m \in \mathcal{W}(\mathbf{S},\mathbf{Q},\mathbf{W}(0),T)$, and $\mathrm{vec}(\mathbf{W})=\mathrm{vec}(\mathbf{W}_{\mathbf{S}})=\mathrm{vec}(\mathbf{W}(0))-\sum_{t'=0}^{t-1}\eta/n\cdot\mathbf{M}\mathbf{Z}_{\mathbf{S}}(t')\mathbf{u}(t')$ for some $t\in[T]$, 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 $t'\in[0,t-1]$.

We note that $\vec{\mathbf{w}}_r$ is expressed as

$$\vec{\mathbf{w}}_r=\vec{\mathbf{w}}_{\mathbf{S},r}(t)=\vec{\mathbf{w}}_r(0)-\sum_{t'=0}^{t-1}\frac{\eta}{nN}\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\mathbf{K}_N^{(s)}\mathbf{Z}_{\mathbf{Q}}(0)^\top\mathbf{Z}_{\mathbf{S}}(t')\mathbf{u}(t'), \tag{97}$$

where the notation $\vec{\mathbf{w}}_{\mathbf{S},r}$ emphasizes that $\vec{\mathbf{w}}_r$ depends on the training features $\mathbf{S}$. We define the event

$$E_r(R):=\left\{\left|\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\right|\le R\right\},\quad r\in[m].$$

We now approximate $f(\mathbf{W},\mathbf{x})$ by $g(\mathbf{x}):=\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\ge0\right\}}\vec{\mathbf{w}}_r^\top\mathbf{x}$. We have

$$\begin{aligned}
&|f(\mathbf{W},\mathbf{x})-g(\mathbf{x})|\\
&=\frac{1}{\sqrt{m}}\left|\sum_{r=1}^m a_r\sigma\left(\vec{\mathbf{w}}_r^\top\mathbf{x}\right)-\sum_{r=1}^m a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\ge0\right\}}\vec{\mathbf{w}}_r^\top\mathbf{x}\right|\\
&\le\frac{1}{\sqrt{m}}\sum_{r=1}^m\left|a_r\left(\mathbb{I}_{\{E_r(R)\}}+\mathbb{I}_{\left\{\bar{E}_r(R)\right\}}\right)\left(\sigma\left(\vec{\mathbf{w}}_r^\top\mathbf{x}\right)-\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\ge0\right\}}\vec{\mathbf{w}}_r^\top\mathbf{x}\right)\right|\\
&=\frac{1}{\sqrt{m}}\sum_{r=1}^m\mathbb{I}_{\{E_r(R)\}}\left|\sigma\left(\vec{\mathbf{w}}_r^\top\mathbf{x}\right)-\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\ge0\right\}}\vec{\mathbf{w}}_r^\top\mathbf{x}\right|\\
&=\frac{1}{\sqrt{m}}\sum_{r=1}^m\mathbb{I}_{\{E_r(R)\}}\left|\sigma\left(\vec{\mathbf{w}}_r^\top\mathbf{x}\right)-\sigma\left(\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\right)-\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\ge0\right\}}(\vec{\mathbf{w}}_r-\vec{\mathbf{w}}_r(0))^\top\mathbf{x}\right|\\
&\le\frac{2R}{\sqrt{m}}\sum_{r=1}^m\mathbb{I}_{\{E_r(R)\}}.
\end{aligned}\tag{98}$$

Plugging $R=\frac{\eta c_{\mathbf{u}}T}{2\sqrt{m}}$ in (98), we have

$$\begin{aligned}
|f(\mathbf{W},\mathbf{x})-g(\mathbf{x})|&\le\frac{2R}{\sqrt{m}}\sum_{r=1}^m\mathbb{I}_{\{E_r(R)\}}=\eta c_{\mathbf{u}}T\cdot\frac{1}{m}\sum_{r=1}^m\mathbb{I}_{\{E_r(R)\}}\\
&=\eta c_{\mathbf{u}}T\cdot\widehat{v}_R(\mathbf{W}(0),\mathbf{x})\le\eta c_{\mathbf{u}}T\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right).
\end{aligned}\tag{99}$$

Using (97), we can express $g(\mathbf{x})$ as

$$g(\mathbf{x})=\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\ge0\right\}}\vec{\mathbf{w}}_r(0)^\top\mathbf{x}$$

$$-\sum_{t'=0}^{t-1}\underbrace{\frac{1}{\sqrt{m}}\sum_{r=1}^{m}a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\left(\frac{\eta}{n}\left[\mathbf{M}\right]_{[(r-1)d+1:rd]}\mathbf{Z}_{\mathbf{S}}(t')\mathbf{u}(t')\right)^\top\mathbf{x}}_{:=G_{t'}(\mathbf{x})}\overset{①}{=}-\sum_{t'=0}^{t-1}G_{t'}(\mathbf{x}),\quad(100)$$

where ① follows from the fact that $\frac{1}{\sqrt{m}}\sum_{r=1}^{m}a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\vec{\mathbf{w}}_r(0)^\top\mathbf{x}=f(\mathbf{W}(0),\mathbf{x})=0$ due to the symmetric initialization of the two-layer NN. For each $G_{t'}$ on the RHS of (100),

$$\begin{aligned}
G_{t'}(\mathbf{x}) &\overset{②}{=} \frac{\eta}{nN\sqrt{m}}\sum_{r=1}^{m}a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\left(\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\mathbf{K}_N^{(s)}\mathbf{H}(0)^\top\mathbf{u}(t')\right)^\top\mathbf{x}\\
&\quad+\frac{\eta}{nN\sqrt{m}}\sum_{r=1}^{m}a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\left(\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\mathbf{K}_N^{(s)}\left(\mathbf{H}(t')-\mathbf{H}(0)\right)^\top\mathbf{u}(t')\right)^\top\mathbf{x}\\
&\overset{③}{=} \frac{\eta}{nN\sqrt{m}}\sum_{r=1}^{m}a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\left(\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\mathbf{K}_N^{(s)}\mathbf{K}_{\mathbf{S},\mathbf{Q}}^\top\mathbf{u}(t')\right)^\top\mathbf{x}+\\
&\quad+\frac{\eta}{nN\sqrt{m}}\sum_{r=1}^{m}a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\left(\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\mathbf{K}_N^{(s)}\left(\mathbf{H}(0)-\mathbf{K}_{\mathbf{S},\mathbf{Q}}\right)^\top\mathbf{u}(t')\right)^\top\mathbf{x}\\
&\quad+\frac{\eta}{nN\sqrt{m}}\sum_{r=1}^{m}a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\left(\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\mathbf{K}_N^{(s)}\left(\mathbf{H}(t')-\mathbf{H}(0)\right)^\top\mathbf{u}(t')\right)^\top\mathbf{x}\\
&\quad\underbrace{\overset{④}{=}\frac{\eta}{nN}\sum_{j=1}^{n}\sum_{j',p=1}^{N}K(\mathbf{x},\vec{\mathbf{q}}_{j'})\left[\mathbf{K}_N^{(s)}\right]_{j'p}\left[\mathbf{K}_{\mathbf{S},\mathbf{Q}}^\top\right]_{pj}\mathbf{u}_j(t')}_{:=D(\mathbf{x})}\\
&\quad\underbrace{+\frac{\eta}{nN}\sum_{j=1}^{n}\sum_{j',p=1}^{N}q_{j'}\left[\mathbf{K}_N^{(s)}\right]_{j'p}\left[\mathbf{K}_{\mathbf{S},\mathbf{Q}}^\top\right]_{pj}\mathbf{u}_j(t')}_{:=E_1(\mathbf{x})}\\
&\quad\underbrace{+\frac{\eta}{nN\sqrt{m}}\sum_{r=1}^{m}a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\left(\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\mathbf{K}_N^{(s)}\left(\mathbf{H}(0)-\mathbf{K}_{\mathbf{S},\mathbf{Q}}\right)^\top\mathbf{u}(t')\right)^\top\mathbf{x}}_{:=E_2(\mathbf{x})}\\
&\quad\underbrace{+\frac{\eta}{nN\sqrt{m}}\sum_{r=1}^{m}a_r\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\mathbf{x}\geq 0\right\}}\left(\left[\mathbf{Z}_{\mathbf{Q}}(0)\right]_{[(r-1)d+1:rd]}\mathbf{K}_N^{(s)}\left(\mathbf{H}(t')-\mathbf{H}(0)\right)^\top\mathbf{u}(t')\right)^\top\mathbf{x}}_{:=E_3(\mathbf{x})}
\end{aligned}\quad(101)$$

where $\mathbf{H}(t)\in\mathbb{R}^{n\times N}$ in ② is a matrix defined by

$$\mathbf{H}_{pq}(t)=\frac{\vec{\mathbf{x}}_p^\top\vec{\mathbf{q}}_q}{m}\sum_{r=1}^{m}\mathbb{I}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top\vec{\mathbf{x}}_p\geq 0\right\}}\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top\vec{\mathbf{q}}_q\geq 0\right\}},\quad\forall\, p\in[n],q\in[N].$$

$\mathbf{K}_{\mathbf{S},\mathbf{Q}}\in\mathbb{R}^{n\times N}$ in ③ is specified by $[\mathbf{K}_{\mathbf{S},\mathbf{Q}}]_{ij}=K(\vec{\mathbf{x}}_i,\vec{\mathbf{q}}_j)$ for $i\in[n],j\in[N]$, and $q_{j'}:=\widehat{h}(\mathbf{W}(0),\vec{\mathbf{q}}_{j'},\mathbf{x})-K(\vec{\mathbf{q}}_{j'},\mathbf{x})$ for all $j'\in[N]$ in ④.

We now analyze each term on the RHS of (101). Let $h(\cdot, t') \colon \mathcal{X} \to \mathbb{R}$ be defined by

$$h(\mathbf{x}, t') \coloneqq \frac{\eta}{n} \sum_{j=1}^{n} K^{(\mathrm{int})}(\mathbf{x}, \vec{\mathbf{x}}_j) \mathbf{u}_j(t'),$$

then $h(\cdot, t')$ is an element in the RKHS $\mathcal{H}_{K^{(\mathrm{int})}}$ for each $t' \in [0, t-1]$. It follows from (132) in Theorem B.10 that with probability at least $1-\delta$ over $\mathbf{Q}$,

$$\sup_{\mathbf{x} \in \mathcal{X}, i \in [n]} \left| \widehat{K}^{(\mathrm{int})}(\mathbf{x}, \vec{\mathbf{x}}_i) - K^{(\mathrm{int})}(\mathbf{x}, \vec{\mathbf{x}}_i) \right| \lesssim \sqrt{\frac{\log(4n/\delta)}{N}}. \tag{102}$$

As a result, we have

$$\begin{aligned} \left\| D(\mathbf{x}) - h(\mathbf{x}, t') \right\|_\infty &= \frac{\eta}{n} \sup_{\mathbf{x} \in \mathcal{X}} \left| \sum_{j=1}^{n} \widehat{K}^{(\mathrm{int})}(\cdot, \vec{\mathbf{x}}_j) \mathbf{u}_j(t') - \sum_{j=1}^{n} K^{(\mathrm{int})}(\cdot, \vec{\mathbf{x}}_j) \mathbf{u}_j(t') \right| \\ &\lesssim \eta \sqrt{\frac{\log(4n/\delta)}{N}}. \end{aligned} \tag{103}$$

We define

$$h(\cdot) \coloneqq - \sum_{t'=0}^{t-1} h(\cdot, t'), \tag{104}$$

and $e(\cdot) \coloneqq f(\mathbf{W}, \cdot) - h(\cdot)$. We now derive the L^∞-norm of e. To this end, we find the upper L^∞ bounds for E_1, E_2, E_3 on the RHS of (101). Since $\mathbf{W}(0) \in \mathcal{W}_0$, $\left| q_{j'} \right| \le C_1(m/2, d, 1/n)$ for all $j' \in [N]$. Moreover, $\|\mathbf{K}_{\mathbf{S},\mathbf{Q}}\|_2 \le \sqrt{nN}$, $\|\mathbf{u}(t')\|_2 \le c_{\mathbf{u}} \sqrt{n}$ with high probability, and we have

$$\begin{aligned} \|E_1\|_\infty &= \left\| \frac{\eta}{nN} \sum_{j=1}^{n} \sum_{j',p=1}^{N} q_{j'} \left[\mathbf{K}_N^{(s)} \right]_{j'p} \left[\mathbf{K}_{\mathbf{S},\mathbf{Q}}^\top \right]_{pj} \mathbf{u}_j(t') \right\|_\infty \\ &\le \frac{\eta}{nN} \left\| \mathbf{u}(t') \right\|_2 \left\| \mathbf{K}_N^{(s)} \right\|_2 \|\mathbf{K}_{\mathbf{S},\mathbf{Q}}\|_2 \sqrt{N} C_1(m/2, d, 1/n) \le \frac{\eta c_{\mathbf{u}}}{2} \cdot C_1(m/2, d, 1/n). \end{aligned} \tag{105}$$

We now bound the last term on the RHS of (101). Define $\mathbf{X}' \in \mathbb{R}^{dm}$ with $[\mathbf{X}']_{(r-1)d+1:rd} = \frac{a_r}{\sqrt{m}} \mathbb{I}_{\left\{ \vec{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0 \right\}} \vec{\mathbf{x}}$ for all $r \in [m]$, then $\|\mathbf{X}'\|_2 \le 1$. Moreover, $\|\mathbf{K}_{\mathbf{S},\mathbf{Q}} - \mathbf{H}(0)\|_2 \le \sqrt{nN}$ $C_1(m/2, d, 1/n)$ by (75) in the proof of Lemma B.3, and $\|\mathbf{Z}_{\mathbf{Q}}(0)\|_2 \le \sqrt{N}$, so we have

$$\begin{aligned} \|E_2\|_\infty &\le \frac{\eta}{nN} \left\| \left(\mathbf{X}'^\top \mathbf{Z}_{\mathbf{Q}}(0) \mathbf{K}_N^{(s)} \left(\mathbf{H}(0) - \mathbf{K}_{\mathbf{S},\mathbf{Q}} \right)^\top \mathbf{u}(t') \right) \right\|_2 \\ &\le \frac{\eta}{nN} \left\| \mathbf{X}' \right\|_2 \|\mathbf{Z}_{\mathbf{Q}}(0)\|_2 \left\| \mathbf{K}_N^{(s)} \right\|_2 \|\mathbf{H}(0) - \mathbf{K}_{\mathbf{S},\mathbf{Q}}\|_2 \left\| \mathbf{u}(t') \right\|_2 \\ &\le \frac{\eta}{2nN} \cdot \sqrt{N} \cdot \sqrt{nN} C_1(m/2, d, 1/n) \cdot c_{\mathbf{u}} \sqrt{n} \le \frac{\eta c_{\mathbf{u}}}{2} \cdot C_1(m/2, d, 1/n). \end{aligned} \tag{106}$$

We have $\|\mathbf{H}_{ij}(t) - \mathbf{H}_{ij}(0)\|_2 \le \sqrt{nN} \left(\frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n) \right)$ by (71) in the proof of Lemma B.3. As a result, E_3 can be bounded by

$$\|E_3\|_\infty \le \frac{\eta}{nN} \left\| \mathbf{X}'^\top \mathbf{Z}_{\mathbf{Q}}(0) \mathbf{K}_N^{(s)} \left(\mathbf{H}(t') - \mathbf{H}(0) \right)^\top \mathbf{u}(t') \right\|_2$$

$$\leq \frac{\eta}{nN}\|\mathbf{X}'\|_2\|\mathbf{Z}_{\mathbf{Q}}(0)\|_2\left\|\mathbf{K}_N^{(s)}\right\|_2\|\mathbf{H}(t')-\mathbf{H}(0)\|_2\|\mathbf{u}(t')\|_2$$
$$\leq \frac{\eta}{2nN}\cdot\sqrt{N}\cdot\sqrt{nN}\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right)\cdot c_{\mathbf{u}}\sqrt{n}$$
$$\leq \frac{\eta c_{\mathbf{u}}}{2}\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right). \tag{107}$$

Combining (101), (103), (105), (106), and (107), for any $t'\in[0,t-1]$, we obtain

$$\sup_{\mathbf{x}\in\mathcal{X}}\left|G_{t'}(\mathbf{x})-h(\mathbf{x},t')\right|\leq\sup_{\mathbf{x}\in\mathcal{X}}\left|D(\mathbf{x})-h(\mathbf{x},t')\right|+\|E_1\|_\infty+\|E_2\|_\infty+\|E_3\|_\infty$$
$$\leq \eta c_{\mathbf{u}}\left(\frac{\sqrt{\frac{\log(4n/\delta)}{N}}}{c_{\mathbf{u}}}+C_1(m/2,d,1/n)+\frac{1}{2}\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right)\right). \tag{108}$$

It then follows from (99), (100), and (108) that $\|e\|_\infty$ can be bounded by

$$\|e\|_\infty=\|f(\mathbf{W},\cdot)-h\|_\infty\leq\|f(\mathbf{W},\cdot)-g\|_\infty+\|g-h\|_\infty$$
$$\leq\|f(\mathbf{W},\cdot)-g\|_\infty+\sum_{t'=0}^{t-1}\left\|G_{t'}-h(\cdot,t')\right\|_\infty$$
$$\overset{②}{\leq}\eta c_{\mathbf{u}}T\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right)$$
$$+\eta c_{\mathbf{u}}T\left(\frac{\sqrt{\frac{\log(4n/\delta)}{N}}}{c_{\mathbf{u}}}+C_1(m/2,d,1/n)+\frac{1}{2}\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right)\right)$$
$$\leq\eta c_{\mathbf{u}}T\left(\frac{\sqrt{\frac{\log(4n/\delta)}{N}}}{c_{\mathbf{u}}}+C_1(m/2,d,1/n)+\frac{3}{2}\left(\frac{2R}{\sqrt{2\pi\kappa}}+C_2(m/2,d,1/n)\right)\right)\coloneqq\Delta_{m,n,N,\eta}. \tag{109}$$

We now give estimates for $\Delta_{m,n,N,\eta}$. Since $\mathbf{W}(0)\in\mathcal{W}_0$, it follows from Theorem 6.1 that

$$\Delta_{m,n,N,\eta}\lesssim T\sqrt{\frac{\log(4n/\delta)}{N}}+\sqrt{d}m^{-\frac{1}{5}}T^{\frac{3}{2}},$$

By direct calculations, for any $w>0$, when

$$N\gtrsim T^2\log(4n/\delta)/w^2,\quad m\gtrsim T^{\frac{15}{2}}d^{\frac{5}{2}}/w^5,$$

we have $\Delta_{m,n,N,\eta}\leq w$. It follows from Lemma B.5 that with probability at least $1-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$,

$$\|h\|_{\mathcal{H}_{K^{(\mathrm{int})}}}\leq B_h, \tag{110}$$

where B_h is defined in (26), and τ is required to satisfy $\tau\lesssim 1/(\eta T)$. Let $\tau=\Theta(1/(\eta T))$. Lemma B.4 requires that $m\gtrsim T^{\frac{15}{2}}d^{\frac{5}{2}}/\tau^5$. As a result, we have

$$m\gtrsim T^{\frac{25}{2}}d^{\frac{5}{2}}.$$

We note that the requirement on N, $N \gtrsim T^2 \log(4n/\delta)/\tau^2$, in Lemma B.4 is satisfied with $N \gtrsim T^4 \log(4n/\delta)$. Finally, it follows from the Cauchy-Schwarz inequality that $\|h\|_\infty \leq B_h/(2\sqrt{2})$ since $\sup_{\mathbf{x}\in\mathcal{X}} K^{(\mathrm{int})}(\mathbf{x},\mathbf{x}) \in (0, 1/8]$. It then follows from (109) (110) and the above calculations that

$$\|f(\mathbf{W},\cdot)\|_\infty \leq \|h\|_\infty + \|e(\mathbf{x})\|_\infty \leq \frac{B_h}{2\sqrt{2}} + w,$$

which completes the proof. ■

We then have the following lemma about the Rademacher complexity of a localized function class $\left\{f \in \mathcal{F}_{\mathrm{ext}}(B,w)\colon \mathbb{E}_P\left[f^2\right] \leq r\right\}$.

Lemma B.7 (Adapted from (Yang, 2025, Lemma C.9))**.** For every $B, w > 0$ and every $r > 0$,

$$\mathfrak{R}\left(\left\{f \in \mathcal{F}_{\mathrm{ext}}(B,w)\colon \mathbb{E}_P\left[f^2\right] \leq r\right\}\right) \leq \varphi_{B,w}(r), \tag{111}$$

where

$$\varphi_{B,w}(r) \coloneqq \min_{Q\colon Q\geq 0} \left((\sqrt{r}+w)\sqrt{\frac{Q}{n}} + B\left(\frac{\sum_{q=Q+1}^{\infty} \lambda_q^{(\mathrm{int})}}{n}\right)^{1/2}\right) + w. \tag{112}$$

Theorem B.8. Suppose $w \in (0,1)$ and m, N satisfy (94) and (95), respectively. Suppose also that $m \gtrsim n^{2/d}$ and $m/\log m \geq d$. Suppose the neural network $f_t = f(\mathbf{W}(t),\cdot)$ is trained by PGD in Algorithm 1 with the learning rate $\eta = \Theta(1) \in (0,8)$ on the random initialization $\mathbf{W}(0)$, and $T \leq \widehat{T}$. Then for every $t \in [T]$ and every $\delta \in (0,1)$, with probability at least $1 - 2/n - \exp(-\Theta(n)) - \delta - 6\exp\left(-\Theta(n\varepsilon_n^2)\right)$ over the random noise $\mathbf{w}$, the random training features $\mathbf{S}$, the random sample $\mathbf{Q}$, and the random initialization $\mathbf{W}(0)$,

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim \min_{0\leq Q\leq n} \left(\frac{B_0 Q}{n} + B_h\left(\frac{\sum_{q=Q+1}^{\infty} \lambda_q^{(\mathrm{int})}}{n}\right)^{1/2}\right) + \varepsilon_n^2 + w, \tag{113}$$

where $B_0 \coloneqq B_h/(2\sqrt{2}) + 1 + \mu_0/(2\sqrt{2})$.

Proof We first remark that the conditions on m, N are required by Lemma B.4 and Theorem B.6. Define $\mathcal{I}_0 = \{\mathbf{W}(0) \in \mathcal{W}_0\}$. Since $T \geq 1$, the width assumptions imply the conditions of Theorem 6.1, which gives $\Pr[\mathcal{I}_0^c] \leq 2/n$. If T is chosen from $\mathbf{S}$, this bound follows by applying that theorem conditionally on $\mathbf{S}$ and integrating. Let $0 < c_- \leq c_+$ be fixed constants from Lemma B.14, and define the training-feature event

$$\mathcal{A}_n = \left\{c_-\varepsilon_n^2 \leq \widehat{\varepsilon}_n^2 \leq c_+\varepsilon_n^2\right\}, \qquad \Pr[\mathcal{A}_n^c] \leq 4\exp(-\Theta(n\varepsilon_n^2)).$$

For each fixed realization of $\mathbf{S}$ and $\mathbf{W}(0)$ in $\mathcal{I}_0$, Lemma B.4 and Theorem B.6 give the trajectory and decomposition event Ω with

$$\Pr\left[\Omega^c \mid \mathbf{S}, \mathbf{W}(0)\right] \leq \exp(-\Theta(n)) + \delta + \exp(-\Theta(n\widehat{\varepsilon}_n^2)).$$

We regard Ω as the corresponding event in the joint probability space of $(\mathbf{W}(0), \mathbf{w}, \mathbf{Q}, \mathbf{S})$. Both $\mathcal{A}_n$ and $\mathcal{I}_0$ are measurable with respect to $(\mathbf{S}, \mathbf{W}(0))$. Since $\widehat{\varepsilon}_n^2 \geq c_- \varepsilon_n^2$ on $\mathcal{A}_n$, the tower property gives

$$\begin{aligned}\Pr\left[\mathcal{A}_n \cap \mathcal{I}_0 \cap \Omega^c\right] &= \mathbb{E}\left[\mathbb{1}_{\{\mathcal{A}_n \cap \mathcal{I}_0\}} \Pr\left[\Omega^c \mid \mathbf{S}, \mathbf{W}(0)\right]\right] \\ &\leq \exp(-\Theta(n)) + \delta + \exp(-\Theta(n\varepsilon_n^2)).\end{aligned}$$

On $\mathcal{I}_0 \cap \Omega$, we have $\mathbf{W}(t) \in \mathcal{W}(\mathbf{S}, \mathbf{Q}, \mathbf{W}(0), T)$ and

$$f(\mathbf{W}(t), \cdot) = f_t = h + e \in \mathcal{F}_{\text{ext}}(B_h, w)$$

with $h \in \mathcal{H}_{K^{(\text{int})}}(B_h)$ and $\|e\|_\infty \leq w$.

We then derive the sharp upper bound for $\mathbb{E}_P\left[(f_t - f^*)^2\right]$ by applying Theorem A.3 to the function class

$$\mathcal{F} = \left\{F = (f - f^*)^2 : f \in \mathcal{F}_{\text{ext}}(B_h, w)\right\}.$$

With $B_0 = B_h/(2\sqrt{2}) + 1 + \mu_0/(2\sqrt{2}) \geq B_h/(2\sqrt{2}) + w + \mu_0/(2\sqrt{2})$, we have $\|F\|_\infty \leq B_0^2$ with $F \in \mathcal{F}$, so that $\mathbb{E}_P\left[F^2\right] \leq B_0^2 \mathbb{E}_P[F]$. Let $T(F) = B_0^2 \mathbb{E}_P[F]$ for $F \in \mathcal{F}$. Then $\operatorname{Var}[F] \leq \mathbb{E}_P\left[F^2\right] \leq T(F) = B_0^2 \mathbb{E}_P[F]$. We have

$$\begin{aligned}\mathfrak{R}\left(\{F \in \mathcal{F} \colon T(F) \leq r\}\right) &= \mathfrak{R}\left(\left\{(f - f^*)^2 \colon f \in \mathcal{F}_{\text{ext}}(B_h, w), \mathbb{E}_P\left[(f - f^*)^2\right] \leq \frac{r}{B_0^2}\right\}\right) \\ &\overset{①}{\leq} 2B_0 \mathfrak{R}\left(\left\{f - f^* \colon f \in \mathcal{F}_{\text{ext}}(B_h, w), \mathbb{E}_P\left[(f - f^*)^2\right] \leq \frac{r}{B_0^2}\right\}\right) \\ &\overset{②}{\leq} 4B_0 \mathfrak{R}\left(\left\{f \in \mathcal{F}_{\text{ext}}(B_h, w) \colon \mathbb{E}_P\left[f^2\right] \leq \frac{r}{4B_0^2}\right\}\right), \end{aligned} \tag{114}$$

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

It follows from (114) and Lemma B.7 that

$$\begin{aligned}B_0^2 \mathfrak{R}\left(\{F \in \mathcal{F} \colon T(F) \leq r\}\right) &\leq 4B_0^3 \mathfrak{R}\left(\left\{f \colon f \in \mathcal{F}_{\text{ext}}(B_h, w), \mathbb{E}_P\left[f^2\right] \leq \frac{r}{4B_0^2}\right\}\right) \\ &\leq 4B_0^3 \varphi_{B_h, w}\left(\frac{r}{4B_0^2}\right) \coloneqq \psi(r). \end{aligned} \tag{115}$$

ψ defined as the RHS of (115) 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 $0 \le r \le \psi(r) = 4B_0^3 \varphi\left(\frac{r}{4B_0^2}\right)$. Therefore, by the definition of φ in (112), for every $0 \le Q \le n$, we have

$$\frac{r}{4B_0^3} \le \left(\frac{\sqrt{r}}{2B_0} + w\right)\sqrt{\frac{Q}{n}} + B_h\left(\frac{\sum_{q=Q+1}^{\infty} \lambda_q^{(\mathrm{int})}}{n}\right)^{1/2} + w. \tag{116}$$

Solving the quadratic inequality (116) for r, we have

$$r \le \frac{8B_0^4 Q}{n} + 8B_0^3\left(w\left(\sqrt{\frac{Q}{n}} + 1\right) + B_h\left(\frac{\sum_{q=Q+1}^{\infty} \lambda_q^{(\mathrm{int})}}{n}\right)^{1/2}\right). \tag{117}$$

(117) holds for every $0 \le Q \le n$, so we have

$$r \le 8B_0^3 \min_{0\le Q\le n}\left(\frac{B_0 Q}{n} + w\left(\sqrt{\frac{Q}{n}} + 1\right) + B_h\left(\frac{\sum_{q=Q+1}^{\infty} \lambda_q^{(\mathrm{int})}}{n}\right)^{1/2}\right). \tag{118}$$

It then follows from (115) and Theorem A.3 that there is an event $\mathcal{G}_x$, depending only on $\mathbf{S}$, with probability at least $1-\exp(-x)$, on which the following inequality holds simultaneously for every $f_t \in \mathcal{F}_{\mathrm{ext}}(B_h, w)$:

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - \frac{K_0}{K_0 - 1}\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] - \frac{x\left(11B_0^2 + 26B_0^2 K_0\right)}{n} \le \frac{704K_0}{B_0^2} r^*, \tag{119}$$

or

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \lesssim r^* + \frac{x}{n} \tag{120}$$

with $K_0 = 2$ in (119). It follows from (118) and (120) that

$$\begin{aligned}&\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \\ &\lesssim \min_{0\le Q\le n}\left(\frac{B_0 Q}{n} + w\left(\sqrt{\frac{Q}{n}} + 1\right) + B_h\left(\frac{\sum_{q=Q+1}^{\infty} \lambda_q^{(\mathrm{int})}}{n}\right)^{1/2}\right) + \frac{x}{n}.\end{aligned} \tag{121}$$

Let $x = n\varepsilon_n^2$. On $\mathcal{I}_0 \cap \mathcal{A}_n \cap \mathcal{G}_{n\varepsilon_n^2} \cap \Omega$, (113) holds. The preceding bounds and a union bound give

$$\Pr\left[\left(\mathcal{I}_0 \cap \mathcal{A}_n \cap \mathcal{G}_{n\varepsilon_n^2} \cap \Omega\right)^c\right] \le \Pr\left[\mathcal{I}_0^c\right] + \Pr\left[\mathcal{A}_n^c\right] + \Pr\left[\mathcal{G}_{n\varepsilon_n^2}^c\right] + \Pr\left[\mathcal{A}_n \cap \mathcal{I}_0 \cap \Omega^c\right]$$

$$\leq \frac{2}{n} + \exp(-\Theta(n)) + \delta + 6\exp(-\Theta(n\varepsilon_n^2)).$$

This proves the stated joint probability bound without requiring independence among these events. ■

Lemma B.9. Let $\eta \in (0, 8)$. Suppose the neural network trained after the t-th step of PGD, $f_t = f(\mathbf{W}(t), \cdot)$, satisfies $\mathbf{u}(t) = f_t(\mathbf{S}) - \mathbf{y} = \mathbf{v}(t) + \mathbf{e}(t)$ with $\mathbf{v}(t) \in \mathcal{V}_t$ and $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, and $t \in [T]$ with $T \leq \widehat{T}$. If

$$\tau \lesssim \frac{1}{\eta T}, \tag{122}$$

Then for every $t \in [T]$, with probability at least $1 - \exp\left(-c' n\widehat{\varepsilon}_n^2\right)$ over the random noise $\mathbf{w}$, we have

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \leq \frac{3}{\eta t}\left(\frac{\mu_0^2}{2e} + \frac{1}{\eta T} + 2\right). \tag{123}$$

Proof

We have

$$f_t(\mathbf{S}) = f^*(\mathbf{S}) + \mathbf{w} + \mathbf{v}(t) + \mathbf{e}(t), \tag{124}$$

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 $\vec{\mathbf{e}}_1(t) = -\left(\mathbf{I}_n - \eta\mathbf{K}_n^{(\mathrm{int})}\right)^t \mathbf{w}$ and $\left\|\vec{\mathbf{e}}_2(t)\right\|_2 \lesssim \sqrt{n}\tau$. It follows from (124) that

$$\begin{aligned}
&\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] = \frac{1}{n}\|f_t(\mathbf{S}) - f^*(\mathbf{S})\|_2^2 = \frac{1}{n}\|\mathbf{v}(t) + \mathbf{w} + \mathbf{e}(t)\|_2^2 \\
&= \frac{1}{n}\left\|-\left(\mathbf{I} - \eta\mathbf{K}_n^{(\mathrm{int})}\right)^t f^*(\mathbf{S}) + \left(\mathbf{I}_n - \left(\mathbf{I}_n - \eta\mathbf{K}_n^{(\mathrm{int})}\right)^t\right)\mathbf{w} + \vec{\mathbf{e}}_2(t)\right\|_2^2 \\
&\overset{①}{\leq} \frac{3}{n}\sum_{i=1}^n \left(1 - \eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^{2t}\left[\mathbf{U}^{(\mathrm{int})\top} f^*(\mathbf{S})\right]_i^2 + \frac{3}{n}\sum_{i=1}^n \left(1 - \left(1 - \eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^2\left[\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\right]_i^2 + \frac{3}{n}\left\|\vec{\mathbf{e}}_2(t)\right\|_2^2 \\
&\overset{②}{\leq} \frac{3\mu_0^2}{2e\eta t} + \frac{3}{n}\sum_{i=1}^n \left(1 - \left(1 - \eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^2\left[\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\right]_i^2 + \frac{3}{n}\left\|\vec{\mathbf{e}}_2(t)\right\|_2^2 \\
&\leq \frac{3}{\eta t}\left(\frac{\mu_0^2}{2e} + \frac{1}{\eta T}\right) + 3\cdot\underbrace{\frac{1}{n}\sum_{i=1}^n \left(1 - \left(1 - \eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^2\left[\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\right]_i^2}_{:=E_\varepsilon}
\end{aligned} \tag{125}$$

Here ① follows from the Cauchy-Schwarz inequality, ② follows from (58) in the proof of Lemma B.1. The final inequality follows from (122), with the implicit constant chosen sufficiently small that $\|\vec{\mathbf{e}}_2(t)\|_2 \leq \sqrt{n}/(\eta T)$.

We then derive the upper bound for E_ε on the RHS of (125). We define a diagonal matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ with $\mathbf{R}_{ii} = \left(1-(1-\eta\widehat{\lambda}_i^{(\mathrm{int})})^t\right)^2$ for $i \in [n]$. Then we have $E_\varepsilon = 1/n \cdot \mathrm{tr}\left(\mathbf{U}^{(\mathrm{int})}\mathbf{R}\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\mathbf{w}^\top\right)$. It follows from Wright (1973) that

$$\begin{aligned}
&\Pr\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}^{(\mathrm{int})}\mathbf{R}\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\mathbf{w}^\top\right) - \mathbb{E}\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}^{(\mathrm{int})}\mathbf{R}\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\mathbf{w}^\top\right)\right] \geq u\right] \\
&\leq \exp\left(-c\min\left\{nu/\|\mathbf{R}\|_2, n^2u^2/\|\mathbf{R}\|_{\mathrm{F}}^2\right\}\right) \qquad (126)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}^{(\mathrm{int})}\mathbf{R}\mathbf{U}^{(\mathrm{int})\top}\mathbf{w}\mathbf{w}^\top\right)\right] \leq \frac{\sigma_0^2}{n}\sum_{i=1}^n\left(1-\left(1-\eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^2 \\
&\overset{①}{\leq} \frac{\sigma_0^2}{n}\sum_{i=1}^n \min\left\{1, \eta_t^2(\widehat{\lambda}_i^{(\mathrm{int})})^2\right\} \leq \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^n\min\left\{\frac{1}{\eta_t}, \eta_t(\widehat{\lambda}_i^{(\mathrm{int})})^2\right\} \\
&\overset{②}{\leq} \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^n\min\left\{\frac{1}{\eta_t}, \widehat{\lambda}_i^{(\mathrm{int})}\right\} = \sigma_0^2\eta_t\widehat{R}^2_{K^{(\mathrm{int})}}(\sqrt{1/\eta_t}) \leq \frac{1}{\eta_t}. \qquad (127)
\end{aligned}$$

Here ① follows from the fact that $(1-\eta\widehat{\lambda}_i^{(\mathrm{int})})^t \geq \max\left\{0, 1-t\eta\widehat{\lambda}_i^{(\mathrm{int})}\right\}$, 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 $\widehat{R}_{K^{(\mathrm{int})}}(\sqrt{1/\eta_t}) \leq 1/(\sigma_0\eta_t)$, 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

$$\|\mathbf{R}\|_2 \leq \max_{i\in[n]}\left(1-\left(1-\eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^2 \leq \max_{i\in[n]}\min\left\{1, \eta_t^2(\widehat{\lambda}_i^{(\mathrm{int})})^2\right\} \leq 1. \qquad (128)$$

We also have

$$\begin{aligned}
\frac{1}{n}\|\mathbf{R}\|_{\mathrm{F}}^2 &= \frac{1}{n}\sum_{i=1}^n\left(1-\left(1-\eta\widehat{\lambda}_i^{(\mathrm{int})}\right)^t\right)^4 \leq \frac{\eta_t}{n}\sum_{i=1}^n\min\left\{\frac{1}{\eta_t}, \eta_t^3(\widehat{\lambda}_i^{(\mathrm{int})})^4\right\} \\
&\leq \frac{\eta_t}{n}\sum_{i=1}^n\min\left\{\widehat{\lambda}_i^{(\mathrm{int})}, \frac{1}{\eta_t}\right\} = \eta_t\widehat{R}^2_{K^{(\mathrm{int})}}(\sqrt{1/\eta_t}) \leq \frac{1}{\sigma_0^2\eta_t}. \qquad (129)
\end{aligned}$$

Combining (126)-(129), we have

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

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 $1-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$,

$$E_\varepsilon \leq u + \frac{1}{\eta_t} = \frac{2}{\eta_t}. \qquad (130)$$

It then follows from (125), (126)-(130) that

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \le \frac{3}{\eta t}\left(\frac{\mu_0^2}{2e} + \frac{1}{\eta T} + 2\right)$$

with probability at least $1 - \exp\left(-c' n \widehat{\varepsilon}_n^2\right)$. The same event works for every $t \in [T]$: apply the noise-energy bound at T and use the monotonicity of $[1-(1-\eta\widehat{\lambda}_i^{(\mathrm{int})})^t]^2$, together with $1/(\eta T) \le 1/(\eta t)$. ■

B.2 Proofs for the Approximate Uniform Convergence for the Kernel $K^{(\mathrm{int})}$

In this subsection, we present the main theorem, Theorem B.10, regarding the approximate uniform convergence of $\widehat{K}^{(\mathrm{int})}(\cdot, \mathbf{x}')$ to $K^{(\mathrm{int})}(\cdot, \mathbf{x}')$ for every fixed $\mathbf{x}' \in \mathcal{X}$. Theorem B.10 is the formal version of Theorem 6.2 in Section 6.3.

Theorem B.10. For every fixed $\mathbf{x}' \in \mathcal{X}$ and every $\delta \in (0,1)$, with probability at least $1-\delta$ over $\mathbf{Q} = \left\{\vec{\mathbf{q}}_i\right\}_{i=1}^N$, we have

$$\sup_{\mathbf{x}\in\mathcal{X}} \left|\widehat{K}^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') - K^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}')\right| \lesssim \sqrt{\frac{\log(4/\delta)}{N}}. \tag{131}$$

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

$$\sup_{\mathbf{x}\in\mathcal{X}, i\in[n]} \left|\widehat{K}^{(\mathrm{int})}(\mathbf{x}, \vec{\mathbf{x}}_i) - K^{(\mathrm{int})}(\mathbf{x}, \vec{\mathbf{x}}_i)\right| \lesssim \sqrt{\frac{\log(4n/\delta)}{N}}, \tag{132}$$

$$\left\|\widehat{\mathbf{K}}^{(\mathrm{int})} - \mathbf{K}^{(\mathrm{int})}\right\|_2 \lesssim n\sqrt{\frac{\log(4n/\delta)}{N}}. \tag{133}$$

Proof We define

$$p(\mathbf{q}, \mathbf{x}') \coloneqq \frac{1}{N}\sum_{j=1}^N K^{(s)}(\mathbf{q}, \vec{\mathbf{q}}_j) K(\vec{\mathbf{q}}_j, \mathbf{x}'), \quad \forall \mathbf{q}, \mathbf{x}' \in \mathcal{X}. \tag{134}$$

It follows from Theorem C.1 and the fact that $|K(\mathbf{x}, \mathbf{x}')| \le 1/2$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ that $\sup_{\mathbf{q},\mathbf{x}'\in\mathcal{X}} |p(\mathbf{q}, \mathbf{x}')| \le 1/4$. We now fix $\mathbf{x}' \in \mathcal{X}$ in the following arguments. It follows from (140) of Lemma B.11 that for every $t > 0$ and every $i \in [N]$,

$$\Pr\left[\left\|\frac{1}{N}\sum_{j=1}^N K^{(s)}(\cdot, \vec{\mathbf{q}}_j) K(\vec{\mathbf{q}}_j, \mathbf{x}') - \bar{K}^{(s)}(\cdot, \mathbf{x}')\right\|_{\mathcal{H}_K} < t\right] \ge 1 - 2\exp\left(-\Theta(Nt^2)\right), \tag{135}$$

where $\bar{K}^{(s)}(\cdot, \mathbf{x}') \coloneqq \mathbb{E}_{\mathbf{q}}\left[K^{(s)}(\cdot, \mathbf{q}) K(\mathbf{q}, \mathbf{x}')\right]$. The following arguments are conditioned on the event that (135) holds.

For each $i \in [N]$, we have $K^{(s)}(\cdot, \vec{\mathbf{q}}_i) \in \mathcal{H}_K$, and $\bar{K}^{(s)}(\cdot, \mathbf{x}') \in \mathcal{H}_K$. It follows from (135) that, for all $\mathbf{q} \in \mathcal{X}$,

$$\left|p(\mathbf{q}, \mathbf{x}') - \bar{K}^{(s)}(\mathbf{q}, \mathbf{x}')\right| = \left|\left\langle p(\cdot, \mathbf{x}') - \bar{K}^{(s)}(\cdot, \mathbf{x}'), K(\cdot, \mathbf{q})\right\rangle_{\mathcal{H}_K}\right|$$

$$\leq \left\| \frac{1}{N} \sum_{j=1}^{N} K^{(s)}(\cdot, \vec{\mathbf{q}}_j) K(\vec{\mathbf{q}}_j, \mathbf{x}') - \bar{K}^{(s)}(\cdot, \mathbf{x}') \right\|_{\mathcal{H}_K} \cdot \|K(\cdot, \mathbf{q})\|_{\mathcal{H}_K} \leq \frac{t}{\sqrt{2}}. \tag{136}$$

Define

$$\bar{K}^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') := \frac{1}{N} \sum_{j=1}^{N} K(\mathbf{x}, \vec{\mathbf{q}}_j) \bar{K}^{(s)}(\vec{\mathbf{q}}_j, \mathbf{x}'), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

Then it follows from the definition of $\widehat{K}^{(\mathrm{int})}$ in (5) that for all $\mathbf{x} \in \mathcal{X}$,

$$\left| \widehat{K}^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') - \bar{K}^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') \right| = \left| \frac{1}{N} \sum_{j=1}^{N} K(\mathbf{x}, \vec{\mathbf{q}}_j) p(\vec{\mathbf{q}}_j, \mathbf{x}') - \frac{1}{N} \sum_{j=1}^{N} K(\mathbf{x}, \vec{\mathbf{q}}_j) \bar{K}^{(s)}(\vec{\mathbf{q}}_j, \mathbf{x}') \right|$$

$$\leq \frac{1}{N} \sum_{j=1}^{N} \left| K(\mathbf{x}, \vec{\mathbf{q}}_j) \right| \left| p(\vec{\mathbf{q}}_j, \mathbf{x}') - \bar{K}^{(s)}(\vec{\mathbf{q}}_j, \mathbf{x}') \right| \leq \frac{t}{2\sqrt{2}}, \tag{137}$$

where the last inequality follows from (136).

Given the fixed $\mathbf{x}' \in \mathcal{X}$, we now approximate $K^{(\mathrm{int})}(\cdot, \mathbf{x}')$ by $\bar{K}^{(\mathrm{int})}(\cdot, \mathbf{x}')$. First, it can be verified from the definition of $\bar{K}^{(s)}$ and Theorem C.1 that $\sup_{\mathbf{q}, \mathbf{x}' \in \mathcal{X}} \left| \bar{K}^{(s)}(\mathbf{q}, \mathbf{x}') \right| \leq 1/4$ and $\sup_{\mathbf{q} \in \mathcal{X}} |p(\mathbf{q}, \mathbf{x}')| \leq 1/4$. It then follows from (139) of Lemma B.11 that

$$\Pr \left[\left\| \bar{K}^{(\mathrm{int})}(\cdot, \mathbf{x}') - \mathbb{E}_{\mathbf{q}} \left[K(\cdot, \mathbf{q}) \bar{K}^{(s)}(\mathbf{q}, \mathbf{x}') \right] \right\|_{\mathcal{H}_K} > t \right] \leq 2 \exp\left(-\Theta(Nt^2) \right), \tag{138}$$

and we have $\mathbb{E}_{\mathbf{q}} \left[K(\cdot, \mathbf{q}) \bar{K}^{(s)}(\mathbf{q}, \mathbf{x}') \right] = K^{(\mathrm{int})}(\cdot, \mathbf{x}')$. It follows from (137) and (138) that with probability at least $1 - 4\exp\left(-\Theta(Nt^2)\right)$, for all $\mathbf{x} \in \mathcal{X}$,

$$\left| \widehat{K}^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') - K^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') \right| \leq \left| \widehat{K}^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') - \bar{K}^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') \right| + \left| \bar{K}^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') - K^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') \right|$$

$$\leq \frac{t}{2\sqrt{2}} + \left\| \bar{K}^{(\mathrm{int})}(\cdot, \mathbf{x}') - \mathbb{E}_{\mathbf{q}} \left[K(\cdot, \mathbf{q}) \bar{K}^{(s)}(\mathbf{q}, \mathbf{x}') \right] \right\|_{\mathcal{H}_K} \cdot \|K(\cdot, \mathbf{x}))\|_{\mathcal{H}_K} \leq \frac{3t}{2\sqrt{2}},$$

Choosing $t = C\sqrt{\log(4/\delta)/N}$ with C sufficiently large proves (131). Applying this bound with failure probability δ/n at each training feature and using the union bound proves (132). The spectral-norm bound (133) follows from

$$\left\| \widehat{\mathbf{K}}^{(\mathrm{int})} - \mathbf{K}^{(\mathrm{int})} \right\|_2 \leq n \max_{i,j \in [n]} \left| \widehat{K}^{(\mathrm{int})}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) - K^{(\mathrm{int})}(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) \right|.$$

■

Lemma B.11. Suppose that p is a measurable function defined on $\mathcal{X}$ and $\sup_{\mathbf{x} \in \mathcal{X}} |p(\mathbf{x})| \leq p_0 < \infty$. Then for every $r > 0$,

$$\Pr \left[\left\| \frac{1}{N} \sum_{i=1}^{N} K(\cdot, \vec{\mathbf{q}}_i) p(\vec{\mathbf{q}}_i) - \mathbb{E}_{\mathbf{q}} \left[K(\cdot, \mathbf{q}) p(\mathbf{q}) \right] \right\|_{\mathcal{H}_K} > r \right] \leq 2 \exp\left(-\Theta(Nr^2) \right). \tag{139}$$

Similarly, for every $r > 0$,

$$\Pr\left[\left\|\frac{1}{N}\sum_{i=1}^{N} K^{(s)}(\cdot,\vec{\mathbf{q}}_i)p(\vec{\mathbf{q}}_i) - \mathbb{E}_{\mathbf{q}}\left[K^{(s)}(\cdot,\mathbf{q})p(\mathbf{q})\right]\right\|_{\mathcal{H}_K} > r\right] \le 2\exp\left(-\Theta(Nr^2)\right). \quad (140)$$

Proof We work in $\mathcal{B} = \mathcal{H}_K$ with its RKHS norm, which is used whenever the norm subscript is omitted below. Since K is continuous on the compact space $\mathcal{X}$, $\mathcal{H}_K$ is separable. By the parallelogram identity, $\mathcal{H}_K$ is $(2,1)$-smooth in the sense of Pinelis (1994). If $p \equiv 0$, the conclusion is immediate. Otherwise choose $p_0 > 0$ with $\sup_{\mathbf{x}\in\mathcal{X}} |p(\mathbf{x})| \le p_0$. We then construct the martingale $\{f_k\}_{k\in[N]}$. For each $k \in [N]$, we define

$$f_k \coloneqq \mathbb{E}\left[\frac{1}{p_0\sqrt{2N}}\sum_{i=1}^{N}\left(K(\cdot,\vec{\mathbf{q}}_i)p(\vec{\mathbf{q}}_i) - \mathbb{E}_{\mathbf{q}}\left[K(\cdot,\mathbf{q})p(\mathbf{q})\right]\right)\middle|\,\mathcal{F}_k\right], \forall k \in [N],$$

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

$$f_N = \frac{1}{p_0\sqrt{2N}}\sum_{i=1}^{N}\left(K(\cdot,\vec{\mathbf{q}}_i)p(\vec{\mathbf{q}}_i) - \mathbb{E}_{\mathbf{q}}\left[K(\cdot,\mathbf{q})p(\mathbf{q})\right]\right),$$
$$d_k = f_k - f_{k-1} = \frac{1}{p_0\sqrt{2N}}\left(K(\cdot,\vec{\mathbf{q}}_k)p(\vec{\mathbf{q}}_k) - \mathbb{E}_{\mathbf{q}}\left[K(\cdot,\mathbf{q})p(\mathbf{q})\right]\right), \forall k \in [N],$$

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

$$\begin{aligned}\|d_k\|_{\mathcal{H}_K} &= \left\|\frac{1}{p_0\sqrt{2N}}\left(K(\cdot,\vec{\mathbf{q}}_k)p(\vec{\mathbf{q}}_k) - \mathbb{E}_{\mathbf{q}}\left[K(\cdot,\mathbf{q})p(\mathbf{q})\right]\right)\right\|_{\mathcal{H}_K} \\ &\overset{①}{\le} \frac{1}{p_0\sqrt{2N}}\left(p_0\left\|K(\cdot,\vec{\mathbf{q}}_k)\right\|_{\mathcal{H}_K} + p_0\mathbb{E}_{\mathbf{q}}\left[\|K(\cdot,\mathbf{q})\|_{\mathcal{H}_K}\right]\right) \overset{②}{\le} \frac{1}{\sqrt{N}}, \end{aligned} \quad (141)$$

where ① follows from the triangle inequality and the Jensen's inequality, and ② follows from the fact that $\left\|K(\cdot,\vec{\mathbf{q}}_k)\right\|_{\mathcal{H}_K} \le 1/\sqrt{2}$.

Extend the martingale by $f_k = f_N$ and $\mathcal{F}_k = \mathcal{F}_N$ for $k > N$, so that $d_k = 0$ for $k > N$. The deterministic bound in (141) gives $\sum_{k=1}^{\infty} \operatorname{ess\,sup}_{\mathbf{Q}} \|d_k\|_{\mathcal{H}_K}^2 \le 1$. Applying Pinelis (1994, Theorem 3.5) with smoothness constant $D = 1$ and $b_* = 1$, we have $\Pr\left[f^* = \max_{k\in[N]} \|f_k\| > r\right] \le 2\exp\left(-\frac{r^2}{2}\right)$, and it follows that for every $r > 0$,

$$\Pr\left[\left\|\frac{1}{p_0\sqrt{2N}}\sum_{i=1}^{N}\left(K(\cdot,\vec{\mathbf{q}}_i)p(\vec{\mathbf{q}}_i) - \mathbb{E}_{\mathbf{q}}\left[K(\cdot,\mathbf{q})p(\mathbf{q})\right]\right)\right\|_{\mathcal{H}_K} > r\right] \le 2\exp\left(-\frac{r^2}{2}\right),$$

and it follows that

$$\Pr\left[\left\|\frac{1}{N}\sum_{i=1}^{N} K(\cdot,\vec{\mathbf{q}}_i)p(\vec{\mathbf{q}}_i) - \mathbb{E}_{\mathbf{q}}\left[K(\cdot,\mathbf{q})p(\mathbf{q})\right]\right\|_{\mathcal{H}_K} > r\right] \le 2\exp\left(-\Theta(Nr^2)\right),$$

which completes the proof of (139) and the constant in $\Theta(Nr^2)$ depends on p_0.

For each $i \in [N]$, we have $K^{(s)}(\cdot, \vec{\mathbf{q}}_i) \in \mathcal{H}_K$ according to the definition of $K^{(s)}$ in (4) and noting that $\left\{\sqrt{\lambda_j} e_j\right\}$ is an orthonormal basis of $\mathcal{H}_K$. Moreover,

$$\left\|K^{(s)}(\cdot, \vec{\mathbf{q}}_i)\right\|_{\mathcal{H}_K}^2 = \sum_{j \geq 1} \lambda_j^{2s-1} e_j(\vec{\mathbf{q}}_i)^2 = K^{(2s-1)}(\vec{\mathbf{q}}_i, \vec{\mathbf{q}}_i) \leq \frac{1}{2},$$

because $2s-1 \geq 1$ and Theorem C.1 applies. Repeating the proof of (139) to the martingale

$$f_k^{(s)} = \mathbb{E}\left[\frac{1}{p_0\sqrt{2N}} \sum_{i=1}^{N} \left(K^{(s)}(\cdot, \vec{\mathbf{q}}_i) p(\vec{\mathbf{q}}_i) - \mathbb{E}_{\mathbf{q}}\left[K^{(s)}(\cdot, \mathbf{q}) p(\mathbf{q})\right]\right) \middle| \mathcal{F}_k\right], \forall k \in [N] \cup \{0\},$$

proves (140). ∎

B.3 More Results about Reproducing Kernel Hilbert Spaces

Lemma B.12 (In the proof of Raskutti et al. (2014, Lemma 8))**.** Let $r_K = \operatorname{rank}(\mathbf{K}_n)$ and $r_{\text{int}} = \operatorname{rank}(\mathbf{K}_n^{(\text{int})})$. For every $f \in \mathcal{H}_K(\mu_0)$,

$$\left[\mathbf{U}^\top f(\mathbf{S})\right]_i = 0, \qquad i > r_K,$$

and

$$\frac{1}{n} f(\mathbf{S})^\top (\mathbf{K}_n)^\dagger f(\mathbf{S}) = \frac{1}{n} \sum_{i=1}^{r_K} \frac{\left[\mathbf{U}^\top f(\mathbf{S})\right]_i^2}{\widehat{\lambda}_i} \leq \mu_0^2.$$

Similarly, for every $f \in \mathcal{H}_{K^{(\text{int})}}(\mu_0)$,

$$\left[\mathbf{U}^{(\text{int})\top} f(\mathbf{S})\right]_i = 0, \qquad i > r_{\text{int}},$$

and

$$\frac{1}{n} f(\mathbf{S})^\top \left(\mathbf{K}_n^{(\text{int})}\right)^\dagger f(\mathbf{S}) = \frac{1}{n} \sum_{i=1}^{r_{\text{int}}} \frac{\left[\mathbf{U}^{(\text{int})\top} f(\mathbf{S})\right]_i^2}{\widehat{\lambda}_i^{(\text{int})}} \leq \mu_0^2.$$

Proof We first prove the results for the kernel K. Define the empirical kernel operator $\widehat{T}_{K,n} \colon \mathcal{H}_K \to \mathcal{H}_K$ by

$$\left(\widehat{T}_{K,n} g\right)(\cdot) := \frac{1}{n} \sum_{j=1}^{n} K(\vec{\mathbf{x}}_j, \cdot) g(\vec{\mathbf{x}}_j), \qquad g \in \mathcal{H}_K.$$

Recall that $\mathbf{K}_n = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$, $\mathbf{\Sigma} = \operatorname{diag}\left(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n\right)$, where $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_{r_K} > 0$, $\widehat{\lambda}_{r_K+1} = \cdots = \widehat{\lambda}_n = 0$. For every $i \in [r_K]$, let $\mathbf{U}_i$ denote the i-th column of $\mathbf{U}$, and define

$$\widehat{e}_i(\cdot) := \frac{1}{\sqrt{n\widehat{\lambda}_i}} \sum_{j=1}^{n} [\mathbf{U}_i]_j K(\vec{\mathbf{x}}_j, \cdot).$$

It can be verified that $\{\widehat{e}_i\}_{i=1}^{r_K}$ are the orthonormal eigenfunctions of $\widehat{T}_{K,n}$ associated with its positive eigenvalues $\left\{\widehat{\lambda}_i\right\}_{i=1}^{r_K}$. We next show that the evaluation vector of every function in $\mathcal{H}_K$ has no component in the null space of $\mathbf{K}_n$. For every $i > r_K$, we have $\mathbf{K}_n\mathbf{U}_i = \mathbf{0}$, and hence

$$\left\|\sum_{j=1}^{n}[\mathbf{U}_i]_j\, K(\vec{\mathbf{x}}_j,\cdot)\right\|_{\mathcal{H}_K}^2 = \mathbf{U}_i^\top\mathbf{K}\mathbf{U}_i = n\mathbf{U}_i^\top\mathbf{K}_n\mathbf{U}_i = 0.$$

Consequently,

$$\left[\mathbf{U}^\top f(\mathbf{S})\right]_i = \sum_{j=1}^{n}[\mathbf{U}_i]_j\, f(\vec{\mathbf{x}}_j) = \left\langle f, \sum_{j=1}^{n}[\mathbf{U}_i]_j\, K(\vec{\mathbf{x}}_j,\cdot)\right\rangle_{\mathcal{H}_K} = 0, \qquad i > r_K.$$

For every $i \in [r_K]$, the reproducing property gives

$$\langle f, \widehat{e}_i\rangle_{\mathcal{H}_K} = \frac{1}{\sqrt{n\widehat{\lambda}_i}}\sum_{j=1}^{n}[\mathbf{U}_i]_j\, f(\vec{\mathbf{x}}_j) = \frac{\left[\mathbf{U}^\top f(\mathbf{S})\right]_i}{\sqrt{n\widehat{\lambda}_i}}.$$

Since

$$\mathbf{K}_n^\dagger = \mathbf{U}\mathrm{diag}\left(\widehat{\lambda}_1^{-1},\ldots,\widehat{\lambda}_{r_K}^{-1},0,\ldots,0\right)\mathbf{U}^\top,$$

we have

$$\frac{1}{n}f(\mathbf{S})^\top\mathbf{K}_n^\dagger f(\mathbf{S}) = \frac{1}{n}\sum_{i=1}^{r_K}\frac{\left[\mathbf{U}^\top f(\mathbf{S})\right]_i^2}{\widehat{\lambda}_i} = \sum_{i=1}^{r_K}\langle f,\widehat{e}_i\rangle_{\mathcal{H}_K}^2 \le \|f\|_{\mathcal{H}_K}^2 \le \mu_0^2,$$

where the first inequality follows from Bessel's inequality. The results for $K^{(\mathrm{int})}$ are proved by a similar argument.

■

Lemma B.13. For any positive real number $a \in (0,1)$ and positive integer t, we have

$$(1-a)^t \le e^{-ta} \le \frac{1}{eta}. \tag{142}$$

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 B.14. For all sufficiently large n, with probability at least $1 - 4\exp(-\Theta(n\varepsilon_n^2))$,

$$\varepsilon_n^2 \lesssim \widehat{\varepsilon}_n^2, \quad \widehat{\varepsilon}_n^2 \lesssim \varepsilon_n^2. \tag{143}$$

Similarly, with probability at least $1 - 4\exp(-\Theta(n\varepsilon_{K,n}^2))$,

$$\varepsilon_{K,n}^2 \lesssim \widehat{\varepsilon}_{K,n}^2, \quad \widehat{\varepsilon}_{K,n}^2 \lesssim \varepsilon_{K,n}^2. \tag{144}$$

Proof We verify the hypotheses of Yang and Li (2024, Lemma B.7) for both kernels, K and $K^{(\mathrm{int})}$. The kernel K is continuous and satisfies $K(\mathbf{x},\mathbf{x}) = 1/2$ for every $\mathbf{x} \in \mathcal{X}$. By Theorems C.1 and C.2, $K^{(\mathrm{int})}$ is continuous and

$$K^{(\mathrm{int})}(\mathbf{x},\mathbf{x}) = \sum_{j\geq 1} \lambda_j^{s+2} e_j(\mathbf{x})^2 > 0, \qquad \forall \mathbf{x} \in \mathcal{X}.$$

The strict positivity follows from $\sum_{j\geq 1} \lambda_j e_j(\mathbf{x})^2 = K(\mathbf{x},\mathbf{x}) = 1/2$ and positivity of every λ_j. Compactness of $\mathcal{X}$, together with $\lambda_1 \leq 1/2$ and $s \geq 1$, gives

$$0 < c_{\mathrm{int}} \coloneqq \min_{\mathbf{x}\in\mathcal{X}} K^{(\mathrm{int})}(\mathbf{x},\mathbf{x}) \leq \max_{\mathbf{x}\in\mathcal{X}} K^{(\mathrm{int})}(\mathbf{x},\mathbf{x}) \leq \frac{1}{8}.$$

Here c_{int} may depend on the fixed design distribution and on s, but is independent of n. Both kernels satisfy the required diagonal bounds. Their largest population eigenvalues are $\lambda_1 > 0$ and $\lambda_1^{(\mathrm{int})} = \lambda_1^{s+2} > 0$. For all sufficiently large n, the sample-size requirements

$$n \gtrsim \max\left\{\lambda_1^{-1}, (\lambda_1^{(\mathrm{int})})^{-1}\right\}$$

are satisfied. Applying the cited lemma to K yields (144), and applying it to $K^{(\mathrm{int})}$ yields (143). ∎

Lemma B.15 (Yang and Li (2024, Lemma B.9))**.** Suppose $\psi\colon [0,\infty) \to [0,\infty)$ is a sub-root function with the unique fixed point r^*. Then the following properties hold.

(1) Let $a \geq 0$. Then $\psi(r) + a$ as a function of r is also a sub-root function with fixed point r_a^*, and $r^* \leq r_a^* \leq r^* + 2a$.

(2) Let $b \geq 1$, $c \geq 0$. Then $\psi(br + c)$ as a function of r is also a sub-root function with fixed point r_b^*, and $r_b^* \leq br^* + 2c/b$.

(3) Let $b \geq 1$. Then $\psi_b(r) = b\psi(r)$ is also a sub-root function with fixed point r_b^*, and $r_b^* \leq b^2 r^*$.

Appendix C. More Results about $\mathcal{H}_{K^{(\mathrm{int})}}$

Theorem C.1. The RHS of (3), $\sum_{j\geq 1} \lambda_j^s e_j(\mathbf{x}) e_j(\mathbf{x}')$, converges uniformly on $\mathcal{X}\times\mathcal{X}$ with $s \geq 1$, and $K^{(s)}$ is well-defined. Moreover, $\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \left|K^{(s)}(\mathbf{x},\mathbf{x}')\right| \leq 1/2$.

Proof It follows from the proof of Mercer's theorem, such as that in (Sun, 2005), that the convergence on the RHS of

$$K^{(s)}(\mathbf{x},\mathbf{x}') = \sum_{j\geq 1} \lambda_j^s e_j(\mathbf{x}) e_j(\mathbf{x}'), \quad \forall \mathbf{x},\mathbf{x}' \in \mathcal{X},$$

with $s \geq 1$ is uniform and absolute. Moreover, by the definition of the eigenvalue and eigenfunction, we have $T_K e_1 = \lambda_1 e_1$, so that

$$\begin{aligned}\lambda_1^2 &= \lambda_1^2 \int_{\mathcal{X}} e_1^2(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) = \int_{\mathcal{X}} \left(\lambda_1 e_1(\mathbf{x})\right)^2 \mathrm{d}\mu(\mathbf{x}) = \int_{\mathcal{X}} \left(T_K e_1\right)(\mathbf{x})^2 \mathrm{d}\mu(\mathbf{x}) \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}') e_1(\mathbf{x}') \mathrm{d}\mu(\mathbf{x}') \right)^2 \mathrm{d}\mu(\mathbf{x}) \\ &\leq \int_{\mathcal{X}} \left(\int_{\mathcal{X}} K^2(\mathbf{x}, \mathbf{x}') \mathrm{d}\mu(\mathbf{x}') \cdot \int_{\mathcal{X}} e_1^2(\mathbf{x}') \mathrm{d}\mu(\mathbf{x}') \right) \mathrm{d}\mu(\mathbf{x}) \leq \frac{1}{4},\end{aligned}$$

where the first inequality is due to the Hölder's inequality. It follows that $\lambda_j \leq 1/2 < 1$ for all $j \geq 1$. As a result, for all $\mathbf{x} \in \mathcal{X}$, we have

$$K^{(s)}(\mathbf{x}, \mathbf{x}) = \sum_{j \geq 1} \lambda_j^s e_j^2(\mathbf{x}) \leq \sum_{j \geq 1} \lambda_j e_j^2(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}) = \frac{1}{2}.$$

Also, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, it follows from the Cauchy-Schwarz inequality that

$$\left| K^{(s)}(\mathbf{x}, \mathbf{x}') \right| = \left| \sum_{j \geq 1} \lambda_j^s e_j(\mathbf{x}) e_j(\mathbf{x}') \right| \leq \left(\sum_{j \geq 1} \lambda_j^s e_j^2(\mathbf{x}) \right)^{1/2} \cdot \left(\sum_{j \geq 1} \lambda_j^s e_j^2(\mathbf{x}') \right)^{1/2} \leq \frac{1}{2}.$$

■

Theorem C.2. Let $\{e_j\}_{j \geq 1} \subseteq L^2(\mathcal{X}, \mu)$ be a countable orthonormal basis of $\overline{\mathrm{range}(T_K)} = (\ker T_K)^\perp$ consisting of the eigenfunctions of the integral operator $T_K \colon L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$, $\left(T_K f\right)(\mathbf{x}) \coloneqq \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') \mathrm{d}\mu(\mathbf{x}')$, a positive, self-adjoint, and compact operator on $L^2(\mathcal{X}, \mu)$. Let $\{\lambda_j\}_{j \geq 1}$ be a sequence with $\frac{1}{2} \geq \lambda_1 \geq \lambda_2 \geq \ldots > 0$ such that e_j is the eigenfunction of T_K with λ_j being the corresponding eigenvalue. Then e_j is the eigenfunction of $T_{K^{(\mathrm{int})}}$ with λ_j^{s+2} being the corresponding eigenvalue. That is, $T_{K^{(\mathrm{int})}} e_j = \lambda_j^{s+2} e_j$, so that $\lambda_j^{(\mathrm{int})} = \lambda_j^{s+2}$ for all $j \geq 1$. Moreover, $T_{K^{(\mathrm{int})}}$ vanishes on $\ker T_K$, and $\ker T_{K^{(\mathrm{int})}} = \ker T_K$.

Proof First, it follows from the Mercer's theorem that

$$K(\mathbf{v}, \mathbf{v}') = \sum_{j \geq 1} \lambda_j e_j(\mathbf{v}) e_j(\mathbf{v}'), \quad \forall \mathbf{v}, \mathbf{v}' \in \mathcal{X},$$

and the convergence on the RHS of the above equality is uniform and absolute. Then it follows from the definition of $K^{(\mathrm{int})}$ in (4) that

$$\begin{aligned}K^{(\mathrm{int})}(\mathbf{x}, \mathbf{x}') &= \int_{\mathcal{X} \times \mathcal{X}} K(\mathbf{x}, \mathbf{v}) K^{(s)}(\mathbf{v}, \mathbf{v}') K(\mathbf{v}', \mathbf{x}') \mathrm{d}\mu(\mathbf{v}) \otimes \mu(\mathbf{v}') \\ &\overset{①}{=} \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \sum_{j \geq 1} \lambda_j e_j(\mathbf{x}) e_j(\mathbf{v}) \cdot \sum_{j \geq 1} \lambda_j^s e_j(\mathbf{v}) e_j(\mathbf{v}') \mathrm{d}\mu(\mathbf{v}) \right) \cdot \sum_{j \geq 1} \lambda_j e_j(\mathbf{v}') e_j(\mathbf{x}') \mu(\mathbf{v}')\end{aligned}$$

$$\overset{②}{=} \int_{\mathcal{X}} \left(\sum_{j\geq 1} \lambda_j^{s+1} e_j(\mathbf{x}) e_j(\mathbf{v}') \right) \cdot \sum_{j\geq 1} \lambda_j e_j(\mathbf{v}') e_j(\mathbf{x}') \mu(\mathbf{v}') \overset{③}{=} \sum_{j\geq 1} \lambda_j^{s+2} e_j(\mathbf{x}) e_j(\mathbf{x}') \tag{145}$$

where ① follows from the Fubini's Theorem, and ②,③ follow by the orthogonality of the orthogonal basis $\{e_j\}_{j\geq 1}$.

It follows from (145) that for all $j \geq 1$,

$$\left(T_{K^{(\mathrm{int})}} e_j\right)(\mathbf{x}) = \int_{\mathcal{X}} \left(\sum_{j'\geq 1} \lambda_{j'}^{s+2} e_{j'}(\mathbf{x}) e_{j'}(\mathbf{x}') \right) e_j(\mathbf{x}') \mathrm{d}\mu(\mathbf{x}') = \lambda_j^{s+2} e_j(\mathbf{x}),$$

which proves that $\lambda_j^{(\mathrm{int})} = \lambda_j^{s+2}$ for all $j \geq 1$. If $h \in \ker T_K$, then $\langle h, e_j \rangle_{L^2(\mathcal{X},\mu)} = 0$ for all j, so the same kernel expansion gives $T_{K^{(\mathrm{int})}} h = 0$. Conversely, if $T_{K^{(\mathrm{int})}} h = 0$, then $\lambda_j^{s+2} \langle h, e_j \rangle_{L^2(\mathcal{X},\mu)} = 0$ for every j. Since every λ_j is positive, $h \in (\overline{\mathrm{range}(T_K)})^{\perp} = \ker T_K$. ■

Appendix D. More Result about the Eigenvalue Decay Rate

Proof of Proposition 5.3. We first let $\lambda_j = \mu_j(K, \mathcal{X}, \phi^*(\mu))$, and let $S(d-1) = 2\pi^{d/2}/\Gamma(d/2)$ denote the surface area of $\mathbb{S}^{d-1}$. Let $\widehat{\mu} = \phi^*(\mu)$ with probability density function $\widehat{p}$. For any measurable subset $\mathbf{A} \subseteq \mathcal{X}$, we have

$$\widehat{\mu}(\mathbf{A}) = \int_{\mathbf{x}\in\mathbf{A}} \int_0^{\infty} p(r\mathbf{x}) r^{d-1} \mathrm{d}r \mathrm{d}\tilde{\sigma} = S(d-1) \int_{\mathbf{x}\in\mathbf{A}} \int_0^{\infty} p(r\mathbf{x}) r^{d-1} \mathrm{d}r \mathrm{d}\sigma(\mathbf{x}), \tag{146}$$

where $\tilde{\sigma}$ is the surface measure on $\mathcal{X}$, and $\sigma = \tilde{\sigma}/S(d-1)$ is the uniform probability measure on $\mathcal{X}$. It follows from (146) that

$$\frac{\mathrm{d}\widehat{\mu}}{\mathrm{d}\sigma} = q,$$

where, for σ-almost every $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned} q(\mathbf{x}) = S(d-1) \int_0^{\infty} p(r\mathbf{x}) r^{d-1} \mathrm{d}r &\lesssim S(d-1) \int_0^{\infty} r^{d-1} (1+r^2)^{-(d+1)/2} \mathrm{d}r \\ &\lesssim \frac{S(d-1)}{2} B(d/2, 1/2), \end{aligned} \tag{147}$$

and B denotes the Beta function. The last inequality of (147) follows from Lemma D.1.

We have $B(d/2, 1/2) = \sqrt{\pi} \frac{\Gamma(d/2)}{\Gamma((d+1)/2)} \asymp 1/\sqrt{d}$. It then follows from (147) that $q(\mathbf{x}) \lesssim S(d-1)/\sqrt{d}$. Set $\nu = \widehat{\mu} = q\sigma$ and $Q = \|q\|_{L^\infty(\sigma)}$. The preceding estimate gives $Q < \infty$, and $\int_{\mathcal{X}} q \mathrm{d}\sigma = 1$ implies $Q \geq 1$. We compare the integral operators

$$(T_{K,\xi} f)(\mathbf{x}) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) \mathrm{d}\xi(\mathbf{y}), \qquad \xi \in \{\sigma, \nu\}.$$

These are positive, self-adjoint, compact operators on $L^2(\xi)$. Let $\lambda_j^\sigma = \mu_j(K, \mathcal{X}, \sigma)$ denote the positive eigenvalues of $T_{K,\sigma}$, counted with multiplicity and ordered nonincreasingly. By the spherical uniform EDR in Hu et al. (2021); Bietti and Mairal (2019); Bietti and Bach (2021), there are constants $c_0, C_0 > 0$, depending only on d, such that

$$c_0 j^{-b} \le \lambda_j^\sigma \le C_0 j^{-b}, \qquad b = \frac{d}{d-1}, \quad j \ge 1.$$

For $f \in L^2(\nu)$, define $qf = 0$ on $\{q = 0\}$. Since $0 \le q \le Q$, we have $qf \in L^2(\sigma)$ and

$$\|qf\|_{L^2(\sigma)}^2 = \int_{\mathcal{X}} q^2 f^2 \mathrm{d}\sigma \le Q \int_{\mathcal{X}} qf^2 \mathrm{d}\sigma = Q\|f\|_{L^2(\nu)}^2.$$

Boundedness of K justifies Fubini's theorem. Using $\mathrm{d}\nu = q\mathrm{d}\sigma$ gives

$$\langle T_{K,\nu} f, f\rangle_{L^2(\nu)} = \iint_{\mathcal{X}^2} K(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) f(\mathbf{y}) \mathrm{d}\nu(\mathbf{x}) \mathrm{d}\nu(\mathbf{y}) = \langle T_{K,\sigma}(qf), qf\rangle_{L^2(\sigma)}. \tag{148}$$

Let $e_1, e_2, \ldots$ be real $L^2(\sigma)$-orthonormal eigenfunctions corresponding to $\lambda_1^\sigma, \lambda_2^\sigma, \ldots$. They belong to $L^2(\nu)$ because $q \le Q$. Fix $j \ge 1$. If $f \in L^2(\nu)$ satisfies $\langle f, e_a\rangle_{L^2(\nu)} = 0$ for $1 \le a < j$, then qf is orthogonal to these eigenfunctions in $L^2(\sigma)$. The spectral theorem gives

$$\langle T_{K,\nu} f, f\rangle_{L^2(\nu)} = \langle T_{K,\sigma}(qf), qf\rangle_{L^2(\sigma)} \le \lambda_j^\sigma \|qf\|_{L^2(\sigma)}^2 \le Q\lambda_j^\sigma \|f\|_{L^2(\nu)}^2.$$

Write $\lambda_j = \mu_j(K, \mathcal{X}, \nu)$, listing eigenvalues with multiplicity and appending zeros if the rank is finite. The orthogonality conditions impose at most $j-1$ linear constraints. The min–max principle on $L^2(\nu)$ therefore gives

$$\lambda_j \le Q\lambda_j^\sigma \le QC_0 j^{-b}.$$

We now prove the lower bound $\lambda_j \gtrsim j^{-b}$ for all $j \ge 1$. For $\tau > 0$, let E_τ be the direct sum of all eigenspaces of $T_{K,\sigma}$ whose eigenvalues are at least τ. When $m = \dim E_\tau > 0$, choose a real $L^2(\sigma)$-orthonormal eigenbasis $e_1, \ldots, e_m$ of E_τ, ordered so that $T_{K,\sigma} e_a = \lambda_a^\sigma e_a$. These eigenfunctions have continuous representatives, given by $e_a = (\lambda_a^\sigma)^{-1} T_{K,\sigma} e_a$.

We first claim that

$$\sum_{a=1}^m e_a(\mathbf{x})^2 = m, \qquad \forall \mathbf{x} \in \mathcal{X}. \tag{149}$$

For an orthogonal matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$, we define $(U_\mathbf{R} f)(\mathbf{x}) = f(\mathbf{R}^{-1}\mathbf{x})$. Rotation invariance of σ makes $U_\mathbf{R}$ unitary. The identity $K(\mathbf{Rx}, \mathbf{Ry}) = K(\mathbf{x}, \mathbf{y})$ and a change of variables give $T_{K,\sigma} U_\mathbf{R} = U_\mathbf{R} T_{K,\sigma}$. For every f in an eigenspace with eigenvalue $\lambda \ge \tau$, commutation gives

$$T_{K,\sigma}(U_\mathbf{R} f) = U_\mathbf{R}(T_{K,\sigma} f) = \lambda U_\mathbf{R} f.$$

Therefore, $U_\mathbf{R}$ maps each such eigenspace into itself. Since E_τ is the direct sum of these complete eigenspaces, $U_\mathbf{R} E_\tau \subseteq E_\tau$. The same argument applies to $U_{\mathbf{R}^{-1}} = U_\mathbf{R}^{-1}$, which gives $U_\mathbf{R} E_\tau = E_\tau$. There is an orthogonal matrix $\mathbf{O}_\mathbf{R} \in \mathbb{R}^{m\times m}$ such that

$$\big(e_1(\mathbf{R}^{-1}\mathbf{x}), \ldots, e_m(\mathbf{R}^{-1}\mathbf{x})\big)^\top = \mathbf{O}_\mathbf{R} \big(e_1(\mathbf{x}), \ldots, e_m(\mathbf{x})\big)^\top.$$

This equality holds everywhere because both sides are continuous and σ has full support. Taking squared Euclidean norms shows that $F(\mathbf{x}) = \sum_{a=1}^{m} e_a(\mathbf{x})^2$ is invariant under rotations. Transitivity of the rotation action on $\mathcal{X}$ makes F constant. Its constant value is

$$F(\mathbf{x}) = \int_{\mathcal{X}} F(\mathbf{y}) \mathrm{d}\sigma(\mathbf{y}) = \sum_{a=1}^{m} \int_{\mathcal{X}} e_a(\mathbf{y})^2 \mathrm{d}\sigma(\mathbf{y}) = m,$$

which proves (149).
Fix $j \geq 1$ and we set

$$L = \lceil 2Qj \rceil, \quad \tau = c_0 L^{-b}.$$

Since $\lambda_L^{\sigma} \geq \tau$, the subspace E_τ has dimension $m \geq L$. We define $\mathbf{A} \in \mathbb{R}^{m \times m}$ by

$$\mathbf{A}_{ab} = \int_{\mathcal{X}} q(\mathbf{x}) e_a(\mathbf{x}) e_b(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x}), \qquad a, b \in [m].$$

Orthonormality of $\{e_a\}_{a=1}^{m}$ and $0 \leq q \leq Q$ give $0 \preceq \mathbf{A} \preceq Q\mathbf{I}_m$. By (149) and normalization of q,

$$\mathrm{tr}\,(\mathbf{A}) = \int_{\mathcal{X}} q(\mathbf{x}) \sum_{a=1}^{m} e_a(\mathbf{x})^2 \mathrm{d}\sigma(\mathbf{x}) = m.$$

If k eigenvalues of $\mathbf{A}$ are at least $1/2$, then

$$m \leq kQ + \frac{m-k}{2}, \qquad k \geq \frac{m}{2Q-1} \geq \frac{m}{2Q} \geq j.$$

For every $f \in L^2(\nu)$, the quadratic-form identity (148) and the spectral theorem give

$$\langle T_{K,\nu} f, f \rangle_{L^2(\nu)} = \langle T_{K,\sigma}(qf), qf \rangle_{L^2(\sigma)} \geq \tau \sum_{a=1}^{m} \big| \langle f, e_a \rangle_{L^2(\nu)} \big|^2.$$

Let $F_j \subseteq \mathbb{R}^m$ be spanned by j orthonormal eigenvectors of $\mathbf{A}$ corresponding to the largest j eigenvalues at least $1/2$. For $\mathbf{z} \in F_j$, set $f_{\mathbf{z}} = \sum_{a=1}^{m} z_a e_a$, viewed as an element of $L^2(\nu)$. Then

$$\|f_{\mathbf{z}}\|_{L^2(\nu)}^2 = \mathbf{z}^\top \mathbf{A} \mathbf{z} \geq \frac{1}{2}\|\mathbf{z}\|_2^2, \qquad \sum_{a=1}^{m} \big| \langle f_{\mathbf{z}}, e_a \rangle_{L^2(\nu)} \big|^2 = \mathbf{z}^\top \mathbf{A}^2 \mathbf{z} \geq \frac{1}{2} \mathbf{z}^\top \mathbf{A} \mathbf{z}.$$

Consequently, $\{f_{\mathbf{z}} : \mathbf{z} \in F_j\}$ is a j-dimensional subspace of $L^2(\nu)$ on which

$$\langle T_{K,\nu} f_{\mathbf{z}}, f_{\mathbf{z}} \rangle_{L^2(\nu)} \geq \frac{\tau}{2} \|f_{\mathbf{z}}\|_{L^2(\nu)}^2.$$

The min–max principle on $L^2(\nu)$ gives

$$\lambda_j \geq \frac{\tau}{2} = \frac{c_0}{2} \lceil 2Qj \rceil^{-b} \geq \frac{c_0}{2(3Q)^b} j^{-b},$$

where $Q \geq 1$ gives $\lceil 2Qj \rceil \leq 3Qj$. This bound for every j also proves that $T_{K,\nu}$ has infinitely many positive eigenvalues. Combining the upper and lower bounds yields

$$\frac{c_0}{2(3Q)^b} j^{-b} \leq \mu_j(K, \mathcal{X}, \phi^* \mu) \leq Q C_0 j^{-b}, \qquad j \geq 1. \tag{150}$$

Since d is fixed and $Q < \infty$ follows from the original density assumption, (150) proves $\mu_j(K, \mathcal{X}, \phi^* \mu) \asymp j^{-d/(d-1)}$. ■

Lemma D.1. The beta function defined on the real numbers, $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} \mathrm{d}t$ for all $a, b > 0$, can be written as

$$B(a,b) = \int_0^\infty u^{a-1}(1+u)^{-(a+b)} \mathrm{d}u.$$

Proof This lemma can be verified by the change of variable $t = \frac{u}{1+u}$. ■

References

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.

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.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.*, 47(4):2261 – 2285, 2019.

Alberto Bietti and Francis R. Bach. Deep equals shallow for relu networks in kernel regimes. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

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.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, Jul 2007. ISSN 1615-3383. doi: 10.1007/s10208-006-0196-8.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. *On lazy training in differentiable programming*. Curran Associates Inc., Red Hook, NY, USA, 2019.

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.

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.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020. URL http://jmlr.org/papers/v21/19-734.html.

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

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.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *International*

Conference on Artificial Intelligence and Statistics, volume 89 of *Proceedings of Machine Learning Research*, pages 869–878. PMLR, 2019.

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.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *Ann. Statist.*, 51(2):691 – 716, 2023.

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

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.

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.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

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

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.*, 48(4):1875 – 1897, 2020.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35(3):363–417, Jun 2012. ISSN 1432-0940. doi: 10.1007/s00365-012-9153-3.

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.

Hongwei Sun. Mercer theorem for rkhs on noncompact sets. *J. Complex.*, 21(3):337–349, June 2005. ISSN 0885-064X.

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.

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.

Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38611–38654. PMLR, 23–29 Jul 2023.

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.

Yun Yang, Mert Pilanci, and Martin J. Wainwright. Randomized sketches for kernels: Fast and optimal nonparametric regression. *Ann. Statist.*, 45(3):991 – 1023, 2017.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, Aug 2007. ISSN 1432-0940. doi: 10.1007/s00365-006-0663-2.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

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

Gavin Zhang, Salar Fattahi, and Richard Y. Zhang. Preconditioned gradient descent for overparameterized nonconvex burer–monteiro factorization with global optimality certification. *Journal of Machine Learning Research*, 24(163):1–55, 2023.

Haobo Zhang, Yicheng Li, and Qian Lin. On the optimality of misspecified spectral algorithms. *Journal of Machine Learning Research*, 25(188):1–50, 2024.

Kaiqi Zhang and Yu-Xiang Wang. Deep learning meets nonparametric regression: Are weight-decayed dnns locally adaptive? In *International Conference on Learning Representations*. OpenReview.net, 2023.

Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, and Zheng Ma. A type of generalization error induced by initialization in deep neural networks. In Jianfeng Lu and Rachel A. Ward, editors, *Proceedings of Mathematical and Scientific Machine Learning, MSML 2020, 20-24 July 2020, Virtual Conference / Princeton, NJ, USA*, volume 107 of *Proceedings of Machine Learning Research*, pages 144–164. PMLR, 2020.

Junyu Zhou, Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Optimal rates for generalization of gradient descent methods with deep neural networks, 2026. URL https://arxiv.org/abs/2606.06764.

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.