Sharp Generalization for Nonparametric Regression by Over-Parameterized Neural Networks: A Distribution-Free Analysis in Spherical Covariate

Yingzhen Yang [1]

Abstract

Sharp generalization bound for neural networks trained by gradient descent (GD) is of central interest in statistical learning theory and deep learning. In this paper, we consider nonparametric regression by an over-parameterized two-layer NN trained by GD. We show that, if the neural network is trained by GD with early stopping, then the trained network renders a sharp rate of the nonparametric regression risk of $\mathcal{O}(\varepsilon_n^2)$, which is the same rate as that for the classical kernel regression trained by GD with early stopping, where ε_n is the critical population rate of the Neural Tangent Kernel (NTK) associated with the network and n is the size of the training data. It is remarked that our result does not require distributional assumptions on the covariate as long as the covariate lies on the unit sphere, in a strong contrast with many existing results which rely on specific distributions such as the spherical uniform data distribution or distributions satisfying certain restrictive conditions. As a special case of our general result, when the eigenvalues of the associated NTK decay at a rate of $\lambda_j \asymp j^{-\frac{d}{d-1}}$ for $j \ge 1$ which happens under certain distributional assumption such as the training features follow the spherical uniform distribution, we immediately obtain the minimax optimal rate of $\mathcal{O}(n^{-\frac{d}{2d-1}})$, which is the major results of several existing works in this direction. The neural network width in our general result is lower bounded by a function of only d and ε_n, and such width does not depend on the minimum eigenvalue of the empirical NTK matrix whose lower bound usually requires additional assumptions on the training data. Our results are built upon two significant technical results which are of independent interest. First, uniform convergence to the NTK is established during the training process by GD, so that we can have a nice decomposition of the neural network function at any step of the GD into a function in the Reproducing Kernel Hilbert Space associated with the NTK and an error function with a small L^∞-norm. Second, local Rademacher complexity is employed to tightly bound the Rademacher complexity of the function class comprising all the possible neural network functions obtained by GD. Our result formally fills the gap between training a classical kernel regression model and training an over-parameterized but finite-width neural network by GD for nonparametric regression without distributional assumptions about the spherical covariate.

1. Introduction

With the stunning success of deep learning in various areas of machine learning (LeCun et al., 2015), generalization analysis for neural networks is of central interest for statistical learning learning and deep learning. Considerable efforts have been made to analyze the optimization of deep neural networks showing that gradient descent (GD) and stochastic gradient descent (SGD) provably achieve vanishing training loss (Du et al., 2019b; Allen-Zhu et al., 2019b; Du et al., 2019a; Arora et al., 2019; Zou & Gu, 2019; Su & Yang, 2019). There are also extensive efforts devoted to generalization analysis of deep neural networks (DNNs) with algorithmic guarantees, that is, the generalization bounds for neural networks trained by gradient descent or its variants. It has been shown that with sufficient over-parameterization, that is, with enough number of neurons in hidden layers, the training dynamics of deep neural networks (DNNs) can be approximated by that of a kernel method with the kernel induced by the neural network architecture, termed the Neural Tangent Kernel (NTK), while other studies such as (Yang & Hu, 2021) show that infinite-width neural networks can still learn features. The key idea of NTK based generalization analysis is that, for highly

[1] School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281, USA. Correspondence to: Yingzhen Yang <yingzhen.yang@asu.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

over-parameterized networks, the network weights almost remain around their random initialization. As a result, one can use the first-order Taylor expansion around initialization to approximate the neural network functions and analyze their generalization capability (Cao & Gu, 2019; Arora et al., 2019; Ghorbani et al., 2021).

Many existing works in generalization analysis of neural networks focus on clean data, but it is a central problem in statistical learning that how neural networks can obtain sharp convergence rates for the risk of nonparametric regression where the observed data are corrupted by noise. Considerable research has been conducted in this direction which shows that various types of DNNs achieve optimal convergence rates for smooth (Yarotsky, 2017; Bauer & Kohler, 2019; Schmidt-Hieber, 2020; Jiao et al., 2023; Zhang & Wang, 2023) or non-smooth (Imaizumi & Fukumizu, 2019) target functions for nonparametric regression. However, most of these works do not have algorithmic guarantees, that is, the DNNs in these works are constructed specially to achieve optimal rates with no guarantees that an optimization algorithm, such as GD or its variants, can obtain such constructed DNNs. To this end, efforts have been made in the literature to study the minimax optimal risk rates for nonparametric regression with over-parameterized neural networks trained by GD with either early stopping (Li et al., 2024) or ℓ^2-regularization (Hu et al., 2021; Suh et al., 2022). However, most existing works either require spherical uniform data distribution on the unit sphere (Hu et al., 2021; Suh et al., 2022) or certain restrictive conditions on the data distribution.

It remains an interesting and important question for the statistical learning and theoretical deep learning literature that if an over-parameterized neural network trained by GD can achieve sharp risk rates for nonparametric regression with milder assumptions or restrictions on the distribution of the covariate, so that theoretical guarantees can be obtained for data in more practical scenarios. In this paper, we give a confirmative answer to this question. We present sharp risk rate for nonparametric regression with an over-parameterized two-layer NN trained by GD with early stopping, which is distribution-free in spherical covariate. Throughout this paper, distribution-free in spherical covariate means that there are no distributional assumptions about the covariate as long as the covariate lies on the unit sphere. Furthermore, our results give confirmative answers to certain open questions or address particular concerns in the literature of training over-parameterized neural networks by GD with early stopping for nonparametric regression with minimax optimal rates, such as the characterization of the stopping time in the early-stopping mechanism, the lower bound for the network width, and the constant learning rate used in GD. Benefiting from our analysis which is distribution-free in spherical covariate, our answers to these open questions or concerns do not require distributional assumptions about spherical covariate. Section 3 summarizes our main results with their significance and comparison to existing works.

We organize this paper as follows. We first introduce the necessary notations in the remainder of this section. We then introduce in Section 2 the problem setup for nonparametric regression. Our main results are summarized in Section 3 and detailed in Section 5. The training algorithm for the over-parameterized two-layer neural network is introduced in Section 4. The roadmap of proofs, the summary of the technical approaches and the novel results in the proofs, and the novel proof strategy of this work are presented in Section 6. The detailed proofs are deferred to Section A-Section C of the appendix, and Section D of the appendix presents the simulation results.

Notations. We use bold letters for matrices and vectors, and regular lower letter 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 the bold letter with subscripts indicates the corresponding rows or elements of a matrix or a vector. We put an arrow on top of a letter with subscript if it denotes a vector, e.g., $\vec{\mathbf{x}}_i$ denotes the i-th training feature. $\|\cdot\|_F$ and $\|\cdot\|_p$ denote the Frobenius norm and the vector ℓ^p-norm or the matrix p-norm. $[m : n]$ denotes all the natural numbers 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 $\mathbb{S}^{d-1} := \{\mathbf{x} \colon \mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\|_2 = 1\}$. 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 $\langle \cdot, \cdot \rangle_\mu$ and $\|\cdot\|_\mu^2$ are defined as $\langle f, g \rangle_{L^2} := \int_{\mathbb{S}^{d-1}} f(x) g(x) \mathrm{d}\mu(x)$ and $\|f\|_{L^2}^2 := \int_{\mathbb{S}^{d-1}} f^2(x) \mathrm{d}\mu(x) < \infty$. $\mathbf{B}(\mathbf{x}; r)$ 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 $\|g\|_\infty := \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} |g(\mathbf{x})|$. L^∞ is the function class whose elements almost surely have bounded L^∞-norm. $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ denote the inner product and the norm in the Hilbert space $\mathcal{H}$. $a = \mathcal{O}(b)$ or $a \lesssim b$ indicates that there exists a constant $c > 0$ such that $a \le cb$. $\tilde{\mathcal{O}}$ indicates there are specific requirements in the constants of the $\mathcal{O}$ notation. $a = o(b)$ and $a = w(b)$ indicate that $\lim |a/b| = 0$ and $\lim |a/b| = \infty$, respectively. $a \asymp b$ or $a = \Theta(b)$ denotes that there exists constants $c_1, c_2 > 0$ such that $c_1 b \le a \le c_2 b$. Throughout this paper we let the input space be $\mathcal{X} = \mathbb{S}^{d-1}$, and $\mathrm{Unif}\left(\mathcal{X}\right)$ 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 $\mathcal{H}$, $\mathcal{H}(\mu_0)$ denotes the ball centered at the origin with radius μ_0 in $\mathcal{H}$. We use $\mathbb{E}_P\left[\cdot\right]$ to denote the expectation with respect to the distribution P.

2. Problem Setup

We introduce the problem setups for nonparametric regression in this section.

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 feature vector $\vec{\mathbf{x}}_i \in \mathcal{X}$ and its response $y_i \in \mathbb{R}$. Throughout this paper we assume that no two training features coincide, that is, $\vec{\mathbf{x}}_i \neq \vec{\mathbf{x}}_j$ for all $i, j \in [n]$ and $i \neq j$. We denote the training feature vectors by $\mathbf{S} = \left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n$, and denote by P_n the empirical distribution over $\mathbf{S}$. 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 noise with mean 0 and variance proxy σ_0^2, that is, $\mathbb{E}\left[\exp(\lambda w_i)\right] \le \exp(\lambda^2\sigma_0^2/2)$ for any $\lambda \in \mathbb{R}$. f^* is the target function to be detailed later. We define $\mathbf{y} := [y_1, \ldots, y_n]$, $\mathbf{w} := [w_1, \ldots, w_n]^\top$, and use $f^*(\mathbf{S}) := \left[f^*(\vec{\mathbf{x}}_1), \ldots, f^*(\vec{\mathbf{x}}_n)\right]^\top$ to denote the clean target labels. The feature vectors in $\mathbf{S}$ are drawn i.i.d. according to an underlying unknown continuous data distribution P with μ being the probability measure for P. We consider a two-layer NN (NN) in this paper whose 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, $\sigma(\cdot) = \max\left\{\cdot, 0\right\}$ is the ReLU activation function, $\mathbf{W} = \left\{\vec{\mathbf{w}}_r\right\}_{r=1}^m$ with $\vec{\mathbf{w}}_r \in \mathbb{R}^d$ for $r \in [m]$ denotes the weighting vectors in the first layer and m is the number of neurons. $\boldsymbol{a} = [a_1, \ldots, a_m] \in \mathbb{R}^m$ denotes the weights of the second layer. Throughout this paper we also write $\mathbf{W}$ as $\mathbf{W}_{\mathbf{S}}$ so as to indicate that the weighting vectors in $\mathbf{W}$ are trained on the training features $\mathbf{S}$.

2.2. Kernel and Kernel Regression for Nonparametric Regression

We define the kernel function

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

which is in fact the NTK associated with the two-layer NN (1) when only the first layer is trained, and K is a positive-definite (PD) kernel. Let the gram matrix of K over the training data $\mathbf{S}$ 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 $\mathbf{K}_n := \mathbf{K}/n$ is the empirical NTK matrix. Let the eigendecomposition 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 eigenvalues of $\mathbf{K}_n$ and sorted in a non-increasing order. It is proved in existing works, such as (Du et al., 2019b), that $\mathbf{K}_n$ is non-singular, and it can be verified that $\widehat{\lambda}_1 \in (0, 1/2)$. Let $\mathcal{H}_K$ be the Reproducing Kernel Hilbert Space (RKHS) associated with K. Because K is continuous on the compact set $\mathcal{X} \times \mathcal{X}$, the integral operator $T_K \colon L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu), (T_K f)(\mathbf{x}) := \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')\mathrm{d}\mu(\mathbf{x}')$ is a positive, self-adjoint, and compact operator on $L^2(\mathcal{X}, \mu)$. By the spectral theorem, there is a countable orthonormal basis $\{e_j\}_{j \ge 1} \subseteq L^2(\mathcal{X}, \mu)$ and $\{\lambda_j\}_{j \ge 1}$ with $\frac{1}{2} \ge \lambda_1 \ge \lambda_2 \ge \ldots > 0$ such that e_j is the eigenfunction of T_K with λ_j being the corresponding eigenvalue. That is, $T_K e_j = \lambda_j e_j, j \ge 1$. Let $\{\mu_\ell\}_{\ell \ge 1}$ be the distinct eigenvalues associated with T_K, and let m_ℓ be the be the sum of multiplicity of the eigenvalue $\{\mu_{\ell'}\}_{\ell'=1}^{\ell}$. That is, $m_{\ell'} - m_{\ell'-1}$ is the multiplicity of $\mu_{\ell'}$. It is well known that $\left\{v_j = \sqrt{\lambda_j} e_j\right\}_{j \ge 1}$ is an orthonormal basis of $\mathcal{H}_K$. For a positive constant μ_0, we define $\mathcal{H}_K(\mu_0) := \left\{f \in \mathcal{H}_K \colon \|f\|_{\mathcal{H}} \le \mu_0\right\}$ as the closed ball in $\mathcal{H}_K$ centered at 0 with radius μ_0. We note that $\mathcal{H}_K(\mu_0)$ is also specified by $\mathcal{H}_K(\mu_0) = \left\{f \in L^2(\mathcal{X}, \mu) \colon f = \sum_{j=1}^\infty \beta_j e_j, \sum_{j=1}^\infty \beta_j^2/\lambda_j \le \mu_0^2\right\}$.

The Task of Nonparametric Regression. With $f^* \in \mathcal{H}_K(\mu_0)$, the task of the analysis for nonparametric regression is to find an estimator $\widehat{f}$ from the training data $\left\{(\vec{\mathbf{x}}_i, y_i)\right\}_{i=1}^n$ so that the risk $\mathbb{E}_P\left[\left(\widehat{f} - f^*\right)^2\right]$ can converge to 0 with a fast rate. In this work, we aim to establish a sharp rate of the risk where the over-parameterized neural network (1) trained by GD with early stopping serves as the estimator $\widehat{f}$.

Sharp rate of the risk of nonparametric regression using classical kernel regression. The statistical learning literature has established rich results in the sharp convergence rates for the risk of nonparametric kernel regression (Stone, 1985; Yang & Barron, 1999; Raskutti et al., 2014; Yuan & Zhou, 2016), with one representative result in (Raskutti et al., 2014) about kernel regression trained by GD with early stopping. Let ε_n be the critical population rate of the PD kernel K, which is also referred to as the critical radius (Wainwright, 2019) of K. (Raskutti et al., 2014, Theorem 2) shows the following sharp bound for the nonparametric regression risk of a kernel regression model trained by GD with early stopping when $f^* \in \mathcal{H}_K(\mu_0)$. That is, with probability at least $1 - \Theta\left(\exp(-\Theta(n\varepsilon_n^2)\right)$,

$$\mathbb{E}_P\left[\left(f_{\widehat{T}} - f^*\right)^2\right] \lesssim \varepsilon_n^2, \tag{3}$$

where $\widehat{T}$ is the stopping time whose formal definition is deferred to Section 5.1, and $f_{\widehat{T}}$ is the kernel regressor at the $\widehat{T}$-th step of GD for the optimization problem of kernel regression. The risk bound (3) is rather sharp, since it is minimax optimal in several popular learning setups, such as the setup where the eigenvalues $\{\lambda_i\}_{i\ge1}$ exhibit a certain polynomial decay. Such risk bound (3) also holds for a general PD kernel rather than the NTK (2), and the risk bound (3) is also minimax optimal when the PD kernel is low rank. It is also remarked that the risk bound (3) is distribution-free in the bounded covariate, that is, there are no distributional assumptions about the covariate when it is in a bounded input space. Interested readers are referred to (Raskutti et al., 2014) for more details.

The main result of this paper is that the over-parameterized two-layer NN (1) trained by GD with early stopping achieves the same order of risk rate as that in (3) with arbitrary continuous distribution of the spherical covariate, which are summarized in the next section.

3. Summary of Main Results.

Our main results are summarized in this section. Throughout this paper, we consider fixed dimension $d \ge 4$.

First, Theorem 5.1 in Section 5.2 shows that the neural network (1) trained by GD with early stopping using Algorithm 1 enjoys a sharp rate of the nonparametric regression risk, $\mathcal{O}\left(\varepsilon_n^2\right)$, which is the same as that for the classical kernel regression in (3). Such rate of nonparametric regression risk in Theorem 5.1 is distribution-free in spherical covariate, and it immediately leads to minimax optimal rates for certain special cases. For example, when the eigenvalues of the integral operator associated with K has a particular polynomial eigenvalue decay rate (EDR), that is, $\lambda_j \asymp j^{-\frac{d}{d-1}}$ for $j \ge 1$, then in this case $\varepsilon_n^2 \asymp n^{-\frac{d}{2d-1}}$ according to (Raskutti et al., 2014, Corollary 3), and Theorem 5.1 renders the rate of the nonparametric regression risk of $\mathcal{O}(n^{-\frac{d}{2d-1}})$ which is minimax optimal for this special case (Stone, 1985; Yang & Barron, 1999; Yuan & Zhou, 2016). We refer to such EDR the polynomial EDR in the sequel. It is shown in (Bietti & Mairal, 2019; Bietti & Bach, 2021; Li et al., 2024) that the polynomial EDR holds for our NTK in (2) if $P = \mathrm{Unif}\left(\mathcal{X}\right)$, or P satisfies the distributional assumption for (Li et al., 2024, Proposition 13) in Table 1.

We remark that such a minimax optimal rate $\mathcal{O}(n^{-\frac{d}{2d-1}})$ is derived from Theorem 5.1 under the special case of polynomial EDR, and this minimax optimal rate is also the major result of a series of existing works in nonparametric regression by training over-parameterized neural networks (Hu et al., 2021; Suh et al., 2022; Li et al., 2024) when the target function f^* belongs to $\mathcal{H}_{\tilde{K}}$, the RKHS associated with the NTK $\tilde{K}$ of the network in each particular existing work. We note that $\tilde{K}$ is the NTK of the network considered in a particular existing work which may not be the same as our NTK in (2). We also note that one needs to set $s = 1$ in (Li et al., 2024, Proposition 13) so that $f^* \in \mathcal{H}_{\tilde{K}}$, and in this case the risk rate for nonparametric regression in (Li et al., 2024, Proposition 13) is $\mathcal{O}(n^{-\frac{d}{2d-1}})$. To the best of our knowledge, Theorem 5.1 presents the first sharp risk rate for nonparametric regression which is distribution-free in spherical covariate, which is closer to practical scenarios. In contrast, the minimax rates in (Hu et al., 2021; Suh et al., 2022) require spherical uniform data distribution on $\mathcal{X}$. The recent work (Ko & Huo, 2024) also requires certain distributional assumptions for the results about regression convergence rates which does not have algorithmic guarantees. Although the minimax rate in another recent work (Li et al., 2024) does not need the spherical uniform distribution, it still requires a restrictive condition on the data distributions detailed in Table 1, and such condition is met by sub-Gaussian distributions. It is under this condition that (Li et al., 2024) derives the polynomial EDR. Table 1 compares our work to existing works for nonparametric regression with a common setup, that is, $f^* \in \mathcal{H}_{\tilde{K}}$ and the responses $\{y_i\}_{i=1}^n$ are corrupted by i.i.d. Gaussian noise. We further note that although the result in (Kuzborskij & Szepesvári, 2021, Theorem 2) does not require distributional assumptions about the covariate, its risk rate under this common setup is not minimax optimal due to the term σ_0^2 in the risk bound. Furthermore, the other term $\mathcal{O}(n^{\frac{-2}{2+d}})$ in its risk bound suffers from the curse of dimension with a slow rate to 0 for high-dimensional data. We also note that (Kuzborskij & Szepesvári, 2021, Theorem 1) shows the minimax optimal rate of $\mathcal{O}(n^{-\frac{2}{2+d}})$, however, this rate is derived for the noiseless case where the responses are not corrupted by noise.

Second, our results provide confirmative answers to several outstanding open questions or address particular concerns in the existing literature about training over-parameterized neural networks for nonparametric regression by GD with early stopping and sharp risk rates, which are detailed below.

Stopping time in the early-stopping mechanism. An open question raised in (Kuzborskij & Szepesvári, 2021; Hu et al., 2021) is how to characterize the stopping time in the early-stopping mechanism when training the over-parameterized network by GD. Let $\widehat{T}$ be the stopping time, (Li et al., 2024, Proposition 13) shows that the stopping time should satisfy $\widehat{T} \asymp n^{\frac{d}{2d-1}}$ under the distributional as-

Table 1: Comparison between our result and the existing works on the risk rates and assumptions for nonparametric regression by training over-parameterized neural networks with algorithmic guarantees, and the listed results here are under a common and popular setup that $f^* \in \mathcal{H}_{\tilde{K}}$ and the responses $\{y_i\}_{i=1}^n$ are corrupted by i.i.d. Gaussian noise with zero mean and variance σ_0^2.

Existing Works and Our Result	Distributional Assumptions	Eigenvalue Decay Rate (EDR)	Rate of Nonparametric Regression Risk
(Kuzborskij & Szepesvári, 2021, Theorem 2)	No	–	Not minimax optimal, $\sigma_0^2 + \mathcal{O}(n^{\frac{-2}{2+d}})$
(Hu et al., 2021, Theorem 5.2), (Suh et al., 2022, Theorem 3.11)	P is Unif $(\mathcal{X})$	$\lambda_j \asymp j^{-\frac{d}{d-1}}$	minimax optimal, $\mathcal{O}(n^{\frac{-d}{2d-1}})$
(Li et al., 2024, Proposition 13)	P satisfies a restrictive condition: the density $p(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$ satisfies $p(x) \lesssim (1 + \lVert\mathbf{x}\rVert_2^2)^{-(d+2)/2}$.	$\lambda_j \asymp j^{-\frac{d}{d-1}}$	minimax optimal, $\mathcal{O}(n^{\frac{-d}{2d-1}})$
Our Result (Theorem 5.1)	No distributional assumption about P as long as $\mathcal{X} = \mathbb{S}^{d-1}$	No requirement for EDR	$\mathcal{O}\left(\varepsilon_n^2\right)$, which leads to the minimax optimal rate $\mathcal{O}(n^{\frac{-d}{2d-1}})$ claimed in (Hu et al., 2021; Suh et al., 2022) and (Li et al., 2024) as special cases.

sumption in Table 1. In contrast, Theorem 5.1 provides a characterization of $\widehat{T}$ showing that $\widehat{T} \asymp \varepsilon_n^{-2}$, which is distribution-free in spherical covariate. Theorem 5.1 further suggests that for each neural network function f_t obtained at the t-th step of GD with $t \asymp \varepsilon_n^{-2}$, the sharp risk rate of $\mathcal{O}\left(\varepsilon_n^2\right)$ is obtained.

Lower bound for the network width m. Our main result, Theorem 5.1, requires that the network width m, which is the number of neurons in the first layer of the network, satisfies $m \gtrsim d^{\frac{8}{3}}/\varepsilon_n^{\frac{80}{3}}$. Such lower bound for m solely depends on d and ε_n. Under the polynomial EDR, Corollary 5.2, which is a direct consequence of Theorem 5.1, shows that m should satisfy $m \gtrsim n^{\frac{80\alpha}{3(2\alpha+1)}} d^{\frac{8}{3}}$ with $\alpha = d/(2(d-1))$ (see (11)) so that GD with early stopping leads to the minimax rate of $\mathcal{O}(n^{-\frac{d}{2d-1}})$. We remark that this is the first time that the lower bound for the network width m is specified only in terms of n and d under the polynomial EDR with a minimax optimal risk rate for nonparametric regression, which can be easily estimated from the training data. In contrast, under the same polynomial EDR, all the existing works (Hu et al., 2021; Suh et al., 2022; Li et al., 2024) require $m \gtrsim \mathrm{poly}(n, 1/\widehat{\lambda}_n)$. The problem here is that one needs additional assumptions on the training data (Bartlett et al., 2021; Nguyen et al., 2021) to find the lower bound for $\widehat{\lambda}_n$, which is the minimal eigenvalue of the empirical NTK matrix $\mathbf{K}_n$, to further estimate the lower bound for m using the training data.

Corollary 5.2 also gives a competitive and smaller lower bound for the network width m than some existing works which give explicit orders of the lower bound for m. For example, under the assumption of uniform spherical distribution, (Suh et al., 2022, Theorem 3.11) requires that $m/\log^3 m \gtrsim L^{20}n^{24}$ where L is the number of layers of the DNN used in that work, and $m/\log^3 m \gtrsim 2^{20}n^{24}$ even with $L = 2$ for the two-layer network (1) used in our work. Furthermore, the proof of (Li et al., 2024, Proposition 13) suggests that $m \gtrsim n^{24}(\log m)^{12}$. Both lower bounds for m in (Suh et al., 2022, Theorem 3.11) and (Li et al., 2024, Proposition 13) are much larger than our lower bound for m, $n^{\frac{80\alpha}{3(2\alpha+1)}} d^{\frac{8}{3}}$, when $n \to \infty$ and d is fixed, which is the setup considered in (Li et al., 2024). It is worthwhile to mention that (Suh et al., 2022; Li et al., 2024) use DNNs with multiple layers for nonparametric regression. As shown in Table 1, through our careful analysis, a shallow two-layer NN (1) exhibits the same minimax risk rate as its deeper counterpart under the same assumptions with much smaller network width. This observation further support the claim in (Bietti & Bach, 2021) that a shallow over-parameterized neural networks with ReLU activations exhibit the same approximation properties as its deeper counterpart, in our nonparametric regression setup.

Training the network with learning rate $\eta = \Theta(1)$. It is also worthwhile to mention that our main result, Theorem 5.1, suggests that a constant learning rate $\eta = \Theta(1)$ can be used for GD when training the two-layer NN (1), which could lead to better empirical optimization performance in practice. Some existing works in fact require an infinitesimal η. For example, (Li et al., 2024, Proposition 13) is obtained by gradient flow where $\eta \to 0$ instead of the practical GD. Furthermore, (Hu et al., 2021, Theorem 5.2) requires the learning rates for both the squared loss and the ℓ^2-regularization term to have the order of $o(n^{-\frac{3d-1}{2d-1}}) \to 0$ as $n \to \infty$. We note that (Nitanda & Suzuki, 2021) also employs constant learning rate in SGD to train neural networks.

More discussion about this work and the relevant literature. We herein provide more discussion about the results of this work and comparison to the existing relevant works with sharp rates for nonparametric regression. While this paper establishes sharp rate which is distribution-free in spherical covariate, such rate still depends on bounded input space ($\mathcal{X} = \mathbb{S}^{d-1}$) and the condition that the target function $f^* \in \mathcal{H}_K(\mu_0)$. Some other existing works consider target function f^* not belonging to the RKHS ball centered at the origin with constant or low radius, such as (Haas et al., 2023; Bordelon et al., 2024). However, the target functions in (Haas et al., 2023; Bordelon et al., 2024) escape the finite norm or low-norm regime of RKHS at the cost of either restriction condition on the density function of the covariate distribution or the training process. In particular, (Haas et al., 2023, Theorem G.5) requires the condition for bounded density function (in its condition (D3)) of the distribution P, which is not required by our result. Moreover, the training process of the model in (Bordelon et al., 2024) requires information about the target function (in its Eq. (4)) and certain distribution P which admits certain polynomial EDR, that is, $\lambda_j \asymp j^{-\alpha}$ with $\alpha > 1$, which happens under certain restrictive conditions on P.

We also note that in this work, only the first layer of an over-parameterized two-layer neural network is trained, while the weights of the second layer are randomly initialized and then fixed in the training process. In existing works such as (Hu et al., 2021; Suh et al., 2022; Allen-Zhu et al., 2019a), all the layers of a deep neural networks with more than two layers are trained by GD or its variants. However, this work shows that only training the first layer still leads to sharp rate for nonparametric regression, which supports the claim in (Bietti & Bach, 2021) that a shallow over-parameterized neural networks with ReLU activations exhibit the same approximation properties as its deeper counterpart.

4. Training by Gradient Descent and Preconditioned Gradient Descent

In the training process of our network (1), 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:

$$L(\mathbf{W}) \coloneqq \frac{1}{2n} \sum_{i=1}^{n} \left(f(\mathbf{W}, \vec{\mathbf{x}}_i) - y_i \right)^2. \tag{4}$$

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

$$\operatorname{vec}\left(\mathbf{W}_{\mathbf{S}}(t+1)\right) - \operatorname{vec}\left(\mathbf{W}_{\mathbf{S}}(t)\right) = -\frac{\eta}{n} \mathbf{Z}_{\mathbf{S}}(t)(\widehat{\mathbf{y}}(t) - \mathbf{y}), \tag{5}$$

where $\mathbf{y}_i = y_i$, $\widehat{\mathbf{y}}(t) \in \mathbb{R}^n$ with $\left[\widehat{\mathbf{y}}(t)\right]_i = f(\mathbf{W}(t), \vec{\mathbf{x}}_i)$. The notations with the subscripts $\mathbf{S}$ indicate the dependence on the training features $\mathbf{S}$. We also denote $f(\mathbf{W}(t), \cdot)$ as $f_t(\cdot)$ as the neural network function with weighting vectors $\mathbf{W}(t)$ obtained after the t-th step of GD.

We define $\mathbf{Z}_{\mathbf{S}}(t) \in \mathbb{R}^{md \times n}$ which is computed by

$$[\mathbf{Z}_{\mathbf{S}}(t)]_{[(r-1)d+1:rd]i} = \frac{1}{\sqrt{m}} \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(t)^\top \vec{\mathbf{x}}_i \ge 0\right\}} \vec{\mathbf{x}}_i a_r \tag{6}$$

for all $i \in [n]$ and $r \in [m]$. 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 $\widehat{\mathbf{y}}(\mathbf{0}) = \mathbf{0}$, 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 $\vec{\mathbf{w}}_{2r'}(0) \sim \mathcal{N}(\mathbf{0}, \kappa^2 \mathbf{I}_d), a_{2r'} \sim \operatorname{unif}\left(\{-1,1\}\right), \forall r' \in [m/2]$, where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $\operatorname{unif}\left(\{-1,1\}\right)$ denotes a uniform distribution over $\{1,-1\}$, $0 < \kappa \le 1$ controls the magnitude of initialization, and $\kappa \asymp 1$. We set $\vec{\mathbf{w}}_{2r'-1}(0) = \vec{\mathbf{w}}_{2r'}(0)$ and $a_{2r'-1} = -a_{2r}$ for all $r' \in [m/2]$. It then can be verified that $\widehat{\mathbf{y}}(0) = \mathbf{0}$. Once randomly initialized, $\boldsymbol{a}$ is fixed during the training. We use $\mathbf{W}(0)$ to denote the set of all the random weighting 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 GD for T steps. Early stopping is enforced in Algorithm 1 through a bounded T via $T \le \widehat{T}$.

Algorithm 1 Training the Two-Layer NN by GD

```
$\mathbf{W}(T) \leftarrow$ Training-by-GD($T, \mathbf{W}(0)$)
input: $T, \mathbf{W}(0)$
for $t = 1, \ldots, T$ do
    Perform the $t$-th step of GD by (5)
end for
return $\mathbf{W}(T)$
```

5. Main Results

We present the definition of kernel complexity in this section, and then introduce the main results for nonparametric regression of this paper.

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

$$\widehat{R}_K(\varepsilon) \coloneqq \sqrt{\frac{1}{n} \sum_{i=1}^{n} \min\left\{\widehat{\lambda}_i, \varepsilon^2\right\}},$$

$$R_K(\varepsilon) \coloneqq \sqrt{\frac{1}{n} \sum_{i=1}^{\infty} \min\left\{\lambda_i, \varepsilon^2\right\}}. \tag{7}$$

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

Let $\eta_t := \eta t$ for all $t \ge 0$, we then define the stopping time $\widehat{T}$ as

$$\widehat{T} := \min\left\{t \colon \widehat{R}_K(\sqrt{1/\eta_t}) > (\sigma_0 \eta_t)^{-1}\right\} - 1. \tag{8}$$

The stopping time in fact limit the number of steps T in for Algorithm 1 as to be shown in Section 5.2, which in turn enforces the early stopping mechanism.

5.2. Results

Theorem 5.1. Let $c_T, c_t \in (0,1]$ be arbitrary positive constants, and $c_T\widehat{T} \le T \le \widehat{T}$. Suppose $f^* \in \mathcal{H}_K(\mu_0)$, and m satisfies

$$m \gtrsim d^{\frac{8}{3}}/\varepsilon_n^{\frac{80}{3}}, \tag{9}$$

and the neural network $f(\mathbf{W}(t),\cdot)$ is trained by GD using Algorithm 1 with the learning rate $\eta \in [1,2)$ and $T \le \widehat{T}$. Then for every $t \in [c_t T \colon T]$, with probability at least $1 - \exp(-\Theta(n)) - 7\exp\left(-\Theta(n\varepsilon_n^2)\right) - 2/n$ over the random noise $\mathbf{w}$, the random training features $\mathbf{S}$ and the random initialization $\mathbf{W}(0)$, the stopping time satisfies $\widehat{T} \asymp \varepsilon_n^{-2}$, and $f(\mathbf{W}(t),\cdot) = f_t$ satisfies

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] \lesssim \varepsilon_n^2. \tag{10}$$

Significance of Theorem 5.1 and comparison to existing works. To the best of our knowledge, Theorem 5.1 is the first theoretical result which proves that over-parameterized neural network trained by gradient descent with early stopping achieves sharp rate of $\mathcal{O}(\varepsilon_n^2)$, *without distributional assumption on the covariate* as long as the input space $\mathcal{X}$ is $\mathbb{S}^{d-1}$. More discussions about the significance with comparison to existing works are detailed in Section 3.

When the polynomial EDR holds, we can apply Theorem 5.1 to obtain the following corollary.

Corollary 5.2 (Applying Theorem 5.1 to the special case of polynomial EDR)**.** Suppose $\lambda_j \asymp j^{-2\alpha}$ for $j \ge 1$ and $\alpha > 1/2$. Let $c_T, c_t \in (0,1]$ be positive constants, and $c_T\widehat{T} \le T \le \widehat{T}$. Suppose m satisfies

$$m \gtrsim n^{\frac{80\alpha}{3(2\alpha+1)}} d^{\frac{8}{3}}, \tag{11}$$

and the neural network $f(\mathbf{W}(t),\cdot)$ is trained by GD using Algorithm 1 with the learning rate $\eta \in [1,2)$ and $T \le \widehat{T}$. Then for every $t \in [c_t T \colon T]$, with probability at least $1 - \exp(-\Theta(n)) - 7\exp\left(-\Theta(n\varepsilon_n^2)\right) - 2/n$ over the random noise $\mathbf{w}$, the random training features $\mathbf{S}$ and the random initialization $\mathbf{W}(0)$, the stopping time satisfies $\widehat{T} \asymp n^{\frac{d}{2d-1}}$,

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] \lesssim \left(\frac{1}{n}\right)^{\frac{2\alpha}{2\alpha+1}}. \tag{12}$$

The significance of Corollary 5.2 is also detailed in Section 3. Section D of the appendix shows the simulation results with the empirical early stopping time and the theoretically predicted early stopping time, $1/\widehat{\varepsilon}_n^2 \asymp n^{d/(2d-1)}$, for a neural network trained by Algoirthm 1.

6. Roadmap of Proofs

We present the roadmap of our theoretical results which lead to the main result, Theorem 5.1 in Section 5. We first present in the next subsection our results about the uniform convergence to the NTK (2) and more, which are crucial in the analysis of training dynamics by GD.

6.1. Uniform Convergence to the NTK and More

We define functions

$$h(\mathbf{w},\mathbf{x},\mathbf{y}) := \mathbf{x}^\top \mathbf{y} \mathbb{1}_{\{\mathbf{w}^\top \mathbf{x} \ge 0\}} \mathbb{1}_{\{\mathbf{w}^\top \mathbf{y} \ge 0\}},$$

$$\widehat{h}(\mathbf{W},\mathbf{x},\mathbf{y}) := \frac{1}{m}\sum_{r=1}^m h(\overrightarrow{\mathbf{w}}_r,\mathbf{x},\mathbf{y}), \tag{13}$$

$$v_R(\mathbf{w},\mathbf{x}) := \mathbb{1}_{\{|\mathbf{w}^\top \mathbf{x}| \le R\}}, \widehat{v}_R(\mathbf{W},\mathbf{x}) := \frac{1}{m}\sum_{r=1}^m v_R(\overrightarrow{\mathbf{w}}_r,\mathbf{x}). \tag{14}$$

Then we have the following theorem stating the uniform convergence of $\widehat{h}(\mathbf{W}(0),\cdot,\cdot)$ to $K(\cdot,\cdot)$ and uniform convergence of $\widehat{v}_R(\mathbf{W}(0),\mathbf{x})$ to $\frac{2R}{\sqrt{2\pi\kappa}}$ for a positive number $R \lesssim \eta T/\sqrt{m}$. While existing works such as (Li et al., 2024) also has uniform convergence results for over-parameterized neural network, our result does not depend on the Hölder continuity of the NTK.

Theorem 6.1. The following results hold with $\eta \lesssim 1$, $m \gtrsim \max\left\{n^{2/d}, T^{\frac{8}{5}}\right\}$, and $m/\log^{\frac{8}{5}} m \ge d$.

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

$$\sup_{\substack{\mathbf{x}\in\mathcal{X},\\ \mathbf{y}\in\mathcal{X}}} \left|K(\mathbf{x},\mathbf{y}) - \widehat{h}(\mathbf{W}(0),\mathbf{x},\mathbf{y})\right| \le C_1(m/2, d, 1/n)$$

$$\lesssim \sqrt{\frac{d \log m}{m}}. \tag{15}$$

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

$$\sup_{\mathbf{x} \in \mathcal{X}} \left| \widehat{v}_R(\mathbf{W}(0), \mathbf{x}) \right| \le \frac{2R}{\sqrt{2\pi\kappa}} + C_2(m/2, d, 1/n)$$
$$\lesssim \sqrt{d} m^{-\frac{3}{16}} T^{\frac{1}{2}}, \tag{16}$$

where $C_1(m/2, d, 1/n), C_2(m/2, d, 1/n)$ are two positive numbers depending on (m, d, n), with their formal definitions deferred to (39) and (42) in Section C.2 of the appendix.

Proof. This theorem follows from Theorem C.1 and Theorem C.2 in Section C.2 of the appendix. Note that $\widehat{h}(\mathbf{W}, \mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{r=1}^{m} h(\overrightarrow{\mathbf{w}}_r, \mathbf{x}, \mathbf{y}) = \frac{1}{m/2} \sum_{r=1}^{m/2} h(\overrightarrow{\mathbf{w}}_{2r}(0), \mathbf{x}, \mathbf{y})$, then part (1) directly follows from Theorem C.1. Similarly, part (2) directly follows from Theorem C.2. □

6.2. Roadmap of Proofs

Because our main result, Theorem 5.1, is proved by Theorem C.10 and Theorem C.11 deferred to Section C.2, we illustrate in Figure 1, deferred to the appendix, the roadmap containing the intermediate theoretical results which lead to our main result, Theorem 5.1.

Summary of the technical approaches and novel results in the proofs. Theorem C.8 is the first novel result in the proofs of this work, showing that with high probability, the neural network function $f(\mathbf{W}(t), \cdot)$ at step t of GD can be decomposed into two functions by $f(\mathbf{W}(t), \cdot) = f_t = h + e$, where $h \in \mathcal{H}_K$ is a function in the RKHS associated with K with bounded $\mathcal{H}_K$-norm. The error function e has a small L^∞-norm, that is, $\|e\|_\infty \le w$ with w being a small number controlled by the network width m, that is, larger m leads to smaller w. Theorem C.10 is the second novel result in the proofs, where we derive sharp and novel bound for the nonparametric regression risk of the neural network function $f(\mathbf{W}(t), \cdot)$ in Theorem C.10, that is, $\mathbb{E}_P \left[(f_t - f^*)^2 \right] - 2\mathbb{E}_{P_n} \left[(f_t - f^*)^2 \right] \lesssim \varepsilon_n^2 + w$. To the best of our knowledge, Theorem C.10 is among the first in the literature to employ local Rademacher complexity so as to obtain sharp rate for the risk of nonparametric regression which is distribution-free in spherical covariate, and local Rademacher complexity is employed to tightly bound the Rademacher complexity of the function class comprising all the possible neural network functions obtained by GD.

Novel proof strategy of this work. We remark that the proof strategy of our main result, Theorem 5.1, is significantly novel and different from the existing works in training over-parameterized neural networks for nonparametric regression with minimax rates (Hu et al., 2021; Suh et al., 2022; Li et al., 2024). In particular, the common proof strategy in these works uses the decomposition $f_t - f^* = (f_t - \widehat{f}_t^{\text{(NTK)}}) + (\widehat{f}_t^{\text{(NTK)}} - f^*)$ and then show that both $\left\| f_t - \widehat{f}_t^{\text{(NTK)}} \right\|_{L^2}$ and $\left\| \widehat{f}_t^{\text{(NTK)}} - f^* \right\|_{L^2}$ are bounded by certain minimax optimal rate, where $\widehat{f}_t^{\text{(NTK)}}$ is the kernel regressor obtained by either kernel ridge regression (Hu et al., 2021; Suh et al., 2022) or GD with early stopping (Li et al., 2024). The remark after Theorem C.8 details a formulation of $\widehat{f}_t^{\text{(NTK)}}$. $\left\| \widehat{f}_t^{\text{(NTK)}} - f^* \right\|_{L^2}$ is bounded by the minimax optimal rate under certain distributional assumptions in the covariate, and this is one reason for the distributional assumptions about the covariate in existing works such as (Hu et al., 2021; Suh et al., 2022; Li et al., 2024). In a strong contrast, our analysis does not rely on such decomposition of $f_t - f^*$. Instead of approximating f_t by $\widehat{f}_t^{\text{(NTK)}}$, we have a new decomposition of f_t by $f_t = h_t + e_t$ where f_t is approximated by h_t with e_t being the approximation error. As suggested by the remark after Theorem C.8, we have $h_t = \widehat{f}_t^{\text{(NTK)}} + \widehat{e}_2(\cdot, t)$ so that $f_t = \widehat{f}_t^{\text{(NTK)}} + \widehat{e}_2(\cdot, t) + e_t$. Our analysis only requires the network width m to be suitably large so that the $\mathcal{H}_K$-norm of $\widehat{e}_2(\cdot, t)$ is bounded by a positive constant and $\|e_t\|_\infty \le w$, while the common proof strategy in(Hu et al., 2021; Suh et al., 2022; Li et al., 2024) needs m to be sufficiently large so that both $\|\widehat{e}_2(\cdot, t)\|_\infty$ and $\|e_t\|_\infty$ are bounded by an infinitesimal number (a minimax optimal rate such as $\mathcal{O}(n^{-\frac{d}{2d-1}})$ and then $\left\| f_t - \widehat{f}_t^{\text{(NTK)}} \right\|_{L^2}$ is bounded by such minimax optimal rate. Detailed in Section 3, such novel proof strategy leads to our sharp analysis, rendering a smaller lower bound for m in our main result compared to some existing works.

7. Conclusion

In this paper, we show that an over-parameterized two-layer neural network trained by gradient descent (GD) with early stopping renders a sharp rate of the nonparametric regression risk with the order of $\Theta(\varepsilon_n^2)$ with ε_n being the critical population rate or the critical radius of the NTK, which is distribution-free in spherical covariate. We compare our results to the current state-of-the-art with a detailed roadmap of our technical approaches and results in our proofs.

Acknowledgments

This work is supported by the 2023 Mayo Clinic and Arizona State University Alliance for Health Care Collaborative Research Seed Grant Program under Grant Award Number AWD00038846.

Impact Statement

This paper presents work whose goal is to advance the theoretical understanding of the generalization capability of over-parameterized neural networks trained by gradient descent.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 6155–6166, 2019a.

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

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. 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*, pp. 322–332. PMLR, 2019.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 08 2005.

Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. doi: 10.1017/S0962492921000027.

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

Bietti, A. and Bach, F. R. 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.

Bietti, A. and Mairal, J. On the inductive bias of neural tangent kernels. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, pp. 12873–12884, 2019.

Bordelon, B., Atanasov, A. B., and Pehlevan, C. How feature learning can improve neural scaling laws. *CoRR*, abs/2409.17858, 2024.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, pp. 10835–10845, 2019.

Chizat, L., Oyallon, E., and Bach, F. *On lazy training in differentiable programming*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, 2019a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *Ann. Statist.*, 49(2):1029 – 1054, 2021.

Haas, M., Holzmüller, D., von Luxburg, U., and Steinwart, I. Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Hu, T., Wang, W., Lin, C., and Cheng, G. Regularization matters: A nonparametric perspective on overparametrized neural network. In Banerjee, A. and Fukumizu, K. (eds.), *International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 829–837. PMLR, 2021.

Imaizumi, M. and Fukumizu, K. Deep neural networks learn non-smooth functions effectively. In Chaudhuri, K. and Sugiyama, M. (eds.), *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 869–878. PMLR, 2019.

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

Ko, H. and Huo, X. Universal consistency of wide and deep relu neural networks and minimax optimal convergence rates for kolmogorov-donoho optimal function classes. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

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

Kuzborskij, I. and Szepesvári, C. Nonparametric regression with shallow overparameterized neural networks trained by GD with early stopping. In Belkin, M. and Kpotufe, S. (eds.), *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2853–2890. PMLR, 2021.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521:436–444, 2015.

Ledoux, M. *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.

Li, Y., Yu, Z., Chen, G., and Lin, Q. 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.

Mendelson, S. Geometric parameters of kernel machines. In Kivinen, J. and Sloan, R. H. (eds.), *Conference on Computational Learning Theory*, volume 2375 of *Lecture Notes in Computer Science*, pp. 29–43. Springer, 2002.

Nguyen, Q., Mondelli, M., and Montúfar, G. F. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In Meila, M. and Zhang, T. (eds.), *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8119–8129. PMLR, 2021.

Nitanda, A. and Suzuki, T. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

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

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

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

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

Suh, N., Ko, H., and Huo, X. 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.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Wright, F. T. 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.

Yang, G. and Hu, E. J. Tensor programs IV: feature learning in infinite-width neural networks. In Meila, M. and Zhang, T. (eds.), *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*, pp. 11727–11737. PMLR, 2021.

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

Yang, Y. and Li, P. Gradient descent finds over-parameterized neural networks with sharp generalization for nonparametric regression, 2025.

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

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

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

Zhang, Y., Xu, Z. J., Luo, T., and Ma, Z. A type of generalization error induced by initialization in deep neural networks. In Lu, J. and Ward, R. A. (eds.), *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*, pp. 144–164. PMLR, 2020.

Zou, D. and Gu, Q. An improved analysis of training over-parameterized deep neural networks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, pp. 2053–2062, 2019.

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

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

$$\mathfrak{R}(\mathcal{F}) = \mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n, \{\sigma_i\}_{i=1}^n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i)\right]. \tag{17}$$

The empirical Rademacher complexity is defined as

$$\widehat{\mathfrak{R}}(\mathcal{F}) = \mathbb{E}_{\{\sigma_i\}_{i=1}^n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\vec{\mathbf{x}}_i)\right], \tag{18}$$

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

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 $f \in \mathcal{F}$,Var $\left[f(\vec{\mathbf{x}}_i)\right] \le r$. Then, for every $x > 0$, with probability at least $1 - e^{-x}$,

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}_P[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim P_n}[f(\mathbf{x})]\right) \le \inf_{\alpha > 0}\left(2(1+\alpha)\mathbb{E}_{\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n, \{\sigma_i\}_{i=1}^n}[R_n\mathcal{F}] + \sqrt{\frac{2rx}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{x}{n}\right), \tag{19}$$

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

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}_P[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim P_n}[f(\mathbf{x})]\right) \le \inf_{\alpha \in (0,1)}\left(\frac{2(1+\alpha)}{1-\alpha}\mathbb{E}_{\{\sigma_i\}_{i=1}^n}[R_n\mathcal{F}] + \sqrt{\frac{2rx}{n}} + (b-a)\left(\frac{1}{3} + \frac{1}{\alpha} + \frac{1+\alpha}{2\alpha(1-\alpha)}\right)\frac{x}{n}\right). \tag{20}$$

P_n is the empirical distribution over $\left\{\vec{\mathbf{x}}_i\right\}_{i=1}^n$ with $\mathbb{E}_{\mathbf{x} \sim P_n}[f(\mathbf{x})] = \frac{1}{n} \sum\limits_{i=1}^n f(\vec{\mathbf{x}}_i)$. 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, $|\phi(x) - \phi(y)| \le \mu\,|x - y|$ for $\mu > 0$. Then, for every function class $\mathcal{F}$,

$$\mathbb{E}_{\{\sigma_i\}_{i=1}^n}[R_n\phi \circ \mathcal{F}] \le \mu \mathbb{E}_{\{\sigma_i\}_{i=1}^n}[R_n\mathcal{F}], \tag{21}$$

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

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

Theorem 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 $T \colon \mathcal{F} \to \mathbb{R}+$ and some constant $\bar{B}$ such that for every $f \in \mathcal{F}$, Var $[f] \le T(f) \le \bar{B}P(f)$.

Let ψ be a sub-root function and let r^* be the fixed point of ψ. Assume that ψ satisfies, for any $r \ge r^*$, $\psi(r) \ge \bar{B}\mathfrak{R}(\{f \in \mathcal{F} \colon T(f) \le r\})$. Fix $x > 0$, then for any $K_0 > 1$, with probability at least $1 - e^{-x}$,

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

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

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

Proposition A.4. Let $\mathcal{F}$ be a class of functions with ranges in $[0, b]$ for some positive constant b. Let ψ be a sub-root function such that for all $r \ge 0$, $\mathfrak{R}(\{f \in \mathcal{F} \colon \mathbb{E}_P\left[f(\mathbf{x})\right] \le r\}) \le \psi(r)$, and let r^* be the fixed point of ψ. Then for any $K_0 > 1$, with probability $1 - \exp(-x)$, every $f \in \mathcal{F}$ satisfies

$$\mathbb{E}_P\left[f\right] \le \frac{K_0}{K_0 - 1}\mathbb{E}_{P_n}\left[f\right] + \frac{704K_0}{b}r^* + \frac{x\left(11(b-a) + 26bK_0\right)}{n}. \tag{22}$$

B. Proofs for Theorem 5.1 and Corollary 5.2

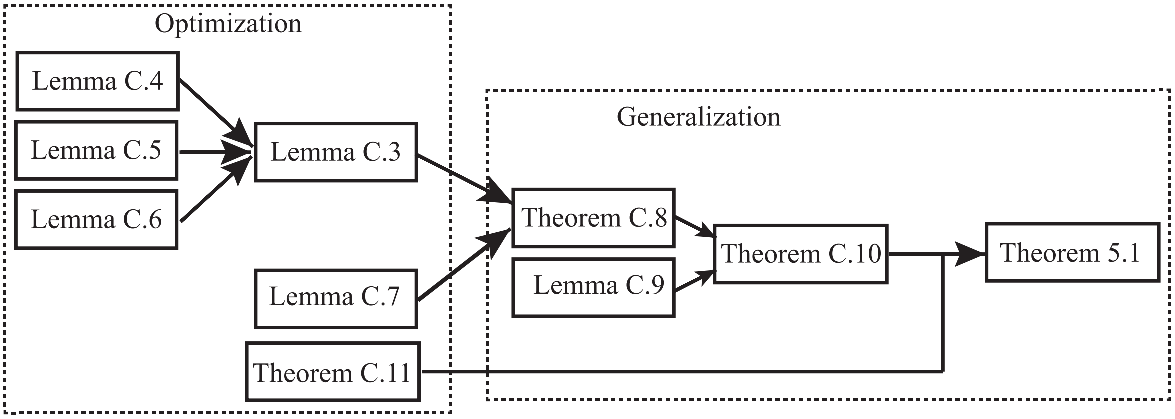

Figure 1: Roadmap of major results leading to the main result, Theorem 5.1. The uniform convergence results in Theorem 6.1 are used in all the optimization results and Theorem C.8.

Proof of Theorem 5.1. We use Theorem C.10 and Theorem C.11 to prove this theorem.

First of all, with the condition on m, d in this theorem, Theorem 6.1 hold, and $\Pr\left[\mathbf{W}(0) \in \mathcal{W}_0\right] \ge 1 - 2/n$. It follows by Theorem C.11 that with probability at least $1 - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$,

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

Plugging such bound for $\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right]$ in (118) of Theorem C.10 leads to

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - \frac{6}{\eta t}\left(\frac{\mu_0^2}{2e} + 3\right) \le c_0'(\varepsilon_n^2 + w). \tag{23}$$

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

$$\widehat{\varepsilon}_n^2 \le \frac{1}{\eta\widehat{T}} \le \frac{2}{\eta(\widehat{T} + 1)} \le 2\widehat{\varepsilon}_n^2. \tag{24}$$

Lemma C.14 suggests that with probability at least $1 - 4\exp(-\Theta(n\varepsilon_n^2))$ over $\mathbf{S}$, $\widehat{\varepsilon}_n^2 \asymp \varepsilon_n^2$. Since $T \asymp \widehat{T}$, for any $t \in [c_t T, T]$, we have

$$\frac{1}{\eta t} \asymp \frac{1}{\eta T} \asymp \frac{1}{\eta \widehat{T}} \asymp \widehat{\varepsilon}_n^2 \asymp \varepsilon_n^2. \tag{25}$$

We have $\Pr[\mathcal{W}_0] \ge 1 - 2/n$. Let $w = \varepsilon_n^2$, we now verify that $w \in (0,1)$. Due to the definition of the fixed point, $w > 0$. Since $\sum\limits_{i \ge 1} \lambda_i = \int_{\mathcal{X}} K(\mathbf{x},\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) = 1/2$, we have

$$0 < w = \frac{1}{n} \sum_{i \ge 1} \min\left\{\lambda_i, \varepsilon_n^2\right\} \le \frac{1}{n} \sum_{i \ge 1} \lambda_i \le \frac{1}{2n} < 1.$$

(10) then follows from (23) with $w = \varepsilon_n^2$, (25) and the union bound. The condition on m in (87) in Theorem C.10, together with $w = \varepsilon_n^2$ and (25) leads to the condition on m in (9). Furthermore, $\widehat{T} \asymp \varepsilon_n^{-2}$ follows from (25) and $\eta = \Theta(1)$.

□

Proof of Corollary 5.2. We apply Theorem 5.1 to prove this corollary.

It is well known, such as (Raskutti et al., 2014, Corollary 3), that $\varepsilon_n^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}}$. It then can be verified by direct calculations that the condition on m, (9) in Theorem 5.1, is satisfied with the given condition (11). It then follows from (10) in Theorem 5.1 that $\mathbb{E}_P\left[(f_{\widehat{T}} - f^*)^2\right] \lesssim n^{-\frac{2\alpha}{2\alpha+1}}$.

□

C. Detailed Proofs

Because Theorem 5.1 is proved by Theorem C.10 and Theorem C.11, in this section, we establish and prove all the theoretical results which lead to Theorem C.10 and Theorem C.11, along with the proof of Theorem C.10 and Theorem C.11.

C.1. Basic Definitions

We introduce the following definitions for the proof of Theorem 5.2. We define

$$\mathbf{u}(t) := \widehat{\mathbf{y}}(t) - \mathbf{y}. \tag{26}$$

Let $\tau \le 1$ be a positive number, and $\varepsilon_0 \in (0,1)$ is an arbitrary positive constant. For $t \ge 0$ and $T \ge 1$ we define the following quantities (or recall their definitions if defined before),

$$c_{\mathbf{u}} = \mu_0 / \min\left\{2, \sqrt{2e\eta}\right\} + \sigma_0 + \tau + 1,$$

$$R = \frac{\eta c_{\mathbf{u}} T}{\sqrt{m}}, \tag{27}$$

$$\mathcal{V}_t := \left\{\mathbf{v} \in \mathbb{R}^n \colon \mathbf{v} = -\left(\mathbf{I}_n - \eta \mathbf{K}_n\right)^t f^*(\mathbf{S})\right\}, \tag{28}$$

$$\mathcal{E}_{t,\tau} := \left\{\mathbf{e} \colon \mathbf{e} = \overset{\rightharpoonup}{\mathbf{e}}_1 + \overset{\rightharpoonup}{\mathbf{e}}_2 \in \mathbb{R}^n, \overset{\rightharpoonup}{\mathbf{e}}_1 = -\left(\mathbf{I}_n - \eta \mathbf{K}_n\right)^t \mathbf{w}, \left\|\overset{\rightharpoonup}{\mathbf{e}}_2\right\|_2 \le \sqrt{n}\tau\right\}. \tag{29}$$

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

$$\mathcal{W}(\mathbf{S}, \mathbf{W}(0), T) := \left\{\mathbf{W} \colon \exists t \in [T] \text{ s.t. } \operatorname{vec}\left(\mathbf{W}\right) = \operatorname{vec}\left(\mathbf{W}(0)\right) - \sum_{t'=0}^{t-1} \frac{\eta}{n} \mathbf{Z}_{\mathbf{S}}(t') \mathbf{u}(t'),\right.$$

$$\mathbf{u}(t') \in \mathbb{R}^n, \mathbf{u}(t') = \mathbf{v}(t') + \mathbf{e}(t'), \mathbf{v}(t') \in \mathcal{V}_{t'}, \mathbf{e}(t') \in \mathcal{E}_{t',\tau}, \text{ for all } t' \in [0, t-1] \Big\} . \tag{30}$$

$\mathcal{W}(\mathbf{S}, \mathbf{W}(0), T)$ is the set of weights of the neural network trained by GD on the training data $\mathbf{S}$ and random initialization $\mathbf{W}(0)$ with the preconditioner $\mathbf{M}$ generated by $\mathbf{Q}$ and the steps of GD no greater than T. The set of functions represented by the two-layer NN with weights in $\mathcal{W}(\mathbf{S}, \mathbf{W}(0), T)$ is then defined as

$$\mathcal{F}_{\rm NN}(\mathbf{S}, \mathbf{W}(0), T) \coloneqq \{ f_t = f(\mathbf{W}(t), \cdot) \colon \exists\, t \in [T], \mathbf{W}(t) \in \mathcal{W}(\mathbf{S}, \mathbf{W}(0), T) \} . \tag{31}$$

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

$$\mathcal{F}(B, w) \coloneqq \{ f \colon f = h + e, h \in \mathcal{H}_K(B), \|e\|_\infty \le w \} . \tag{32}$$

We define the constant

$$B_h \coloneqq \mu_0 + 1 + \sqrt{2}. \tag{33}$$

It will be shown in Theorem C.8 that with high probability, the two-layer NN (1) trained by GD lies in the function class $\mathcal{F}(B_h, w)$ where w can be sufficiently small with a sufficiently large network width m.

We define

$$\mathcal{W}_0 \coloneqq \{ \mathbf{W}(0) \colon (15), (16) \text{ hold} \} \tag{34}$$

be the set of all the good random initializations which satisfy (15) and (16) in Theorem 6.1. Theorem 6.1 shows that we have good random initialization with high probability, that is, $\Pr[\mathbf{W}(0) \in \mathcal{W}_0] \ge 1 - 2/n$. When $\mathbf{W}(0) \in \mathcal{W}_0$, the uniform convergence results, (15) and (16), hold with high probability, which is crucial for our main result in Theorem 5.1.

C.2. Theorem C.10, Theorem C.11, and their proofs with related theoretical results

Theorem C.10 (repeat). Suppose $w \in (0,1)$ and m satisfy

$$m \gtrsim \max\left\{ T^8 d^{\frac{8}{3}} / w^{\frac{16}{3}}, T^{\frac{40}{3}} d^{\frac{8}{3}} \right\}, \tag{35}$$

and the neural network $f(\mathbf{W}(t), \cdot)$ is trained by GD in Algorithm 1 with the learning rate $\eta = \Theta(1) \in (0, 1/\widehat{\lambda}_1)$ on random initialization $\mathbf{W}(0) \in \mathcal{W}_0$, and $T \le \widehat{T}$. Then for every $t \in [T]$, with probability at least $1 - \exp(-\Theta(n)) - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right) - \exp\left(-n\varepsilon_n^2\right)$ over the random noise $\mathbf{w}$ and the random training features $\mathbf{S}$,

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \le c_0 \min_{0 \le Q \le n} \left(\frac{B_0 Q}{n} + w\left(\sqrt{\frac{Q}{n}} + 1\right) + B_h \left(\frac{\sum\limits_{q=Q+1}^{\infty} \lambda_q}{n} \right)^{1/2} \right)^2 + c_0 \varepsilon_n^2. \tag{36}$$

Furthermore, with probability at least $1 - \exp(-\Theta(n)) - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right) - \exp\left(-n\varepsilon_n^2\right)$ over the random noise $\mathbf{w}$, the random training features $\mathbf{S}$,

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \le c_0'(\varepsilon_n^2 + w). \tag{37}$$

Here B_0, c_0, c_0' are absolute positive constants depending on μ_0, and c_0' also depends on σ_0.

Theorem C.11 (repeat). Suppose the neural network trained after the t-th step of gradient descent, $f_t = f(\mathbf{W}(t), \cdot)$, 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 $T \le \widehat{T}$. If

$$\eta \in [1, 2), \quad \tau \le \frac{1}{\eta T},$$

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

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

We have the following two theorems regarding the uniform convergence of $\widehat{h}(\mathbf{W}(0), \cdot, \cdot)$ to $K(\cdot, \cdot)$ and the uniform convergence of $\widehat{v}_R(\mathbf{W}(0), \cdot)$ to $\frac{2R}{\sqrt{2\pi}\kappa}$. Noting that $d \ge 4$, Theorem C.1 and Theorem C.2 can be proved by using the proofs of (Yang & Li, 2025, Theorem VI.7,Theorem VI.8).

Theorem C.1 (Adapted from (Yang & Li, 2025, Theorem VI.7) for $d \ge 4$)**.** Let $\mathbf{W}(0) = \left\{ \vec{\mathbf{w}}_r(0) \right\}_{r=1}^m$, where each $\vec{\mathbf{w}}_r(0) \sim \mathcal{N}(\mathbf{0}, \kappa^2 \mathbf{I}_d)$ for $r \in [m]$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over $\mathbf{W}(0)$,

$$\sup_{\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{X}} \left| K(\mathbf{x}, \mathbf{y}) - \widehat{h}(\mathbf{W}(0), \mathbf{x}, \mathbf{y}) \right| \le C_1(m, d, \delta), \tag{38}$$

where

$$C_1(m,d,\delta) := \frac{1}{\sqrt{m}} \left(6(1+2B\sqrt{d}) + \sqrt{2 \log \frac{2(1+2m)^{2d}}{\delta}} \right) + \frac{14 \log \frac{2(1+2m)^{2d}}{\delta} + 18}{3m}, \tag{39}$$

and B is an absolute positive constant. In addition, when $m \gtrsim n^{1/(2d)}$, $m / \log m \ge d$, and $\delta \asymp 1/n$,

$$C_1(m,d,\delta) \lesssim \sqrt{\frac{d \log m}{m}}. \tag{40}$$

Theorem C.2 (Adapted from (Yang & Li, 2025, Theorem VI.8) for $d \ge 4$)**.** Let $\mathbf{W}(0) = \left\{ \vec{\mathbf{w}}_r(0) \right\}_{r=1}^m$, where each $\vec{\mathbf{w}}_r(0) \sim \mathcal{N}(\mathbf{0}, \kappa^2 \mathbf{I}_d)$ for $r \in [m]$. Suppose $\eta \lesssim 1$, $m \gtrsim 1$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ over $\mathbf{W}(0)$,

$$\sup_{\mathbf{x} \in \mathcal{X}} \left| \widehat{v}_R(\mathbf{W}(0), \mathbf{x}) - \frac{2R}{\sqrt{2\pi}\kappa} \right| \le C_2(m, d, \delta), \tag{41}$$

where

$$C_2(m,d,\delta) := 3\sqrt{\frac{d}{\kappa}} m^{-\frac{3}{16}} T^{\frac{1}{2}} + \sqrt{\frac{2\log \frac{2(1+2\sqrt{m})^d}{\delta}}{m}} + \frac{7\log \frac{2(1+2\sqrt{m})^d}{\delta}}{3m}. \tag{42}$$

In addition, when $m \gtrsim n^{2/d}$, $m / \log^{\frac{8}{5}} m \ge d$, and $\delta \asymp 1/n$,

$$C_2(m,d,\delta) \lesssim \sqrt{d} m^{-\frac{3}{16}} T^{\frac{1}{2}}. \tag{43}$$

Lemma C.3. Suppose

$$m \gtrsim T^8 d^{\frac{8}{3}} / \tau^{\frac{16}{3}}, \tag{44}$$

and the neural network $f(\mathbf{W}(t), \cdot)$ trained by gradient decent with the learning rate $\eta = \Theta(1) \in (0, 1/\widehat{\lambda}_1)$ on the random initialization $\mathbf{W}(0) \in \mathcal{W}_0$. Then with probability at least $1 - \exp(-\Theta(n))$ over the random noise $\mathbf{w}$, $\mathbf{W}(t) \in \mathcal{W}(\mathbf{S}, \mathbf{W}(0), T)$. Moreover, for all $t \in [0, T]$, $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ where $\mathbf{u}(t) = \widehat{\mathbf{y}}(t) - \mathbf{y}$, $\mathbf{v}(t) \in \mathcal{V}_{K,t}$, $\mathbf{e}(t) \in \mathcal{E}_{K,t,\tau}$, and $\|\mathbf{u}(t)\|_2 \le c_{\mathbf{u}} \sqrt{n}$.

Proof. First, when $m \gtrsim T^8 d^{\frac{8}{3}} / \tau^{\frac{16}{3}}$ with a proper constant, it can be verified that $\mathbf{E}_{m,\eta,\tau} \le \tau\sqrt{n}/T$ where $\mathbf{E}_{m,\eta,\tau}$ is defined by (54) of Lemma C.5. Also, Theorem C.1 and Theorem C.2 hold when (44) holds. We then use mathematical induction to prove the lemma. We will first prove that $\mathbf{u}(t) = \mathbf{v}(t) + \mathbf{e}(t)$ where $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, and $\|\mathbf{u}(t)\|_2 \le c_{\mathbf{u}}\sqrt{n}$ for for all $t \in [0, T]$.

When $t = 0$, we have

$$\mathbf{u}(0) = -\mathbf{y} = \mathbf{v}(0) + \mathbf{e}(0), \tag{45}$$

where $\mathbf{v}(0) := -f^*(\mathbf{S}) = -\left(\mathbf{I} - \eta \mathbf{K}_n\right)^0 f^*(\mathbf{S})$, $\mathbf{e}(0) = -\mathbf{w} = \vec{\mathbf{e}}_1(0) + \vec{\mathbf{e}}_2(0)$ with $\vec{\mathbf{e}}_1(0) = -\left(\mathbf{I} - \eta \mathbf{K}_n\right)^0 \mathbf{w}$ 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 C.4 that $\|\mathbf{u}(0)\|_2 \le c_{\mathbf{u}}$ with probability at least $1 - \exp(-\Theta(n))$ over the random noise $\mathbf{w}$.

Suppose that for all $t_1 \in [0, t]$ with $t \in [0, T-1]$, $\mathbf{u}(t_1) = \mathbf{v}(t_1) + \mathbf{e}(t_1)$ where $\mathbf{v}(t_1) \in \mathcal{V}_{t_1}$, and $\mathbf{e}(t_1) = \overrightarrow{\mathbf{e}}_1(t_1) + \overrightarrow{\mathbf{e}}_2(t_1)$ 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 C.5 that the recursion $\mathbf{u}(t'+1) = (\mathbf{I} - \eta\mathbf{K}_n)\,\mathbf{u}(t') + \mathbf{E}(t'+1)$ holds for all $t' \in [0, t]$. As a result, we have

$$\begin{aligned}
\mathbf{u}(t+1) &= (\mathbf{I} - \eta\mathbf{K}_n)\,\mathbf{u}(t) + \mathbf{E}(t+1) \\
&= -\left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1} f^*(\mathbf{S}) - \left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1} \mathbf{w} + \sum_{t'=1}^{t+1} \left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1-t'} \mathbf{E}(t') \\
&= \mathbf{v}(t+1) + \mathbf{e}(t+1),
\end{aligned} \tag{46}$$

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

$$\mathbf{v}(t+1) \coloneqq -\left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1} f^*(\mathbf{S}) \in \mathcal{V}_{t+1}, \tag{47}$$

$$\mathbf{e}(t+1) \coloneqq \underbrace{-\left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1} \mathbf{w}}_{\overrightarrow{\mathbf{e}}_1(t+1)} + \underbrace{\sum_{t'=1}^{t+1} \left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t+1-t'} \mathbf{E}(t')}_{\overrightarrow{\mathbf{e}}_2(t+1)}. \tag{48}$$

We now prove the upper bound for $\overrightarrow{\mathbf{e}}_2(t+1)$. With $\eta \in (0, 1/\widehat{\lambda}_1)$, we have $\|\mathbf{I} - \eta\mathbf{K}_n\|_2 \in (0,1)$. It follows that

$$\left\|\overrightarrow{\mathbf{e}}_2(t+1)\right\|_2 \le \sum_{t'=1}^{t+1} \|\mathbf{I} - \eta\mathbf{K}_n\|_2^{t+1-t'} \|\mathbf{E}(t')\|_2 \le \tau\sqrt{n}, \tag{49}$$

where the last inequality follows from the fact that $\|\mathbf{E}(t)\|_2 \le \mathbf{E}_{m,\eta,\tau} \le \tau\sqrt{n}/T$ for all $t \in [T]$ and the induction hypothesis. It follows that $\mathbf{e}(t+1) \in \mathcal{E}_{t+1,\tau}$. Also, it follows from Lemma C.4 that

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

which completes the induction step and also the proof. □

Lemma C.4. Let $t \in [T]$, $\mathbf{v} = -\left(\mathbf{I} - \eta\mathbf{K}_n\right)^t f^*(\mathbf{S})$, $\mathbf{e} = -\left(\mathbf{I} - \eta\mathbf{K}_n\right)^t \mathbf{w}$, and $\eta \in (0, 1/\widehat{\lambda}_1)$. Then with probability at least $1 - \exp\left(-\Theta(n)\right)$ over the random noise $\mathbf{w}$,

$$\|\mathbf{v}\|_2 + \|\mathbf{e}\|_2 \le \left(\frac{\mu_0}{\sqrt{2e\eta}} + \sigma_0 + 1\right)\sqrt{n}. \tag{50}$$

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

$$\|\mathbf{v}(t)\|_2^2 = \sum_{i=1}^{n} \left(1 - \eta\widehat{\lambda}_i\right)^{2t} \left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \overset{①}{\le} \sum_{i=1}^{n} \frac{1}{2e\eta\widehat{\lambda}_i t} \left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \overset{②}{\le} \frac{n\mu_0^2}{2e\eta t}. \tag{51}$$

Here ① follows Lemma C.13, ② follows by Lemma C.12.

Moreover, it follows from the concentration inequality about quadratic forms of sub-Gaussian random variables in (Wright, 1973) that

$$\Pr\left[\|\mathbf{w}\|_2^2 - \mathbb{E}\left[\|\mathbf{w}\|_2^2\right] > n\right] \le \exp\left(-\Theta(n)\right), \tag{52}$$

and $\mathbb{E}\left[\|\mathbf{w}\|_2\right] \le \sqrt{\mathbb{E}\left[\|\mathbf{w}\|_2^2\right]} = \sqrt{n}\sigma_0$. Therefore, $\Pr\left[\|\mathbf{w}\|_2 - \sqrt{n}\sigma_0 > \sqrt{n}\right] \le \exp\left(-\Theta(n)\right)$.

As a result, we have

$$\left\|\mathbf{v}\right\|_2 + \left\|\mathbf{e}\right\|_2 \le \sqrt{\frac{n\mu_0^2}{2e\eta}} + \left\|\mathbf{w}\right\|_2 \le \left(\frac{\mu_0}{\sqrt{2e\eta}} + \sigma_0 + 1\right)\sqrt{n}.$$

□

Lemma C.5. Let $0 < \eta < 1$, $0 \le t \le T-1$ for $T \ge 1$, and suppose that $\left\|\widehat{\mathbf{y}}(t') - \mathbf{y}\right\|_2 \le c_{\mathbf{u}}\sqrt{n}$ holds for all $0 \le t' \le t$ and the random initialization $\mathbf{W}(0) \in \mathcal{W}_0$. Then

$$\widehat{\mathbf{y}}(t+1) - \mathbf{y} = \left(\mathbf{I} - \eta\mathbf{K}_n\right)\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) + \mathbf{E}(t+1), \tag{53}$$

where $\left\|\mathbf{E}(t+1)\right\|_2 \le \mathbf{E}_{m,\eta,\tau}$, and $\mathbf{E}_{m,\eta,\tau}$ is defined by

$$\mathbf{E}_{m,\eta,\tau} \coloneqq \eta c_{\mathbf{u}}\sqrt{n}\left(4\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2,d,1/n)\right) + C_1(m/2,d,1/n)\right) \lesssim \eta c_{\mathbf{u}}\sqrt{dn}\, m^{-\frac{3}{16}} T^{\frac{1}{2}}. \tag{54}$$

Proof. Because $\left\|\widehat{\mathbf{y}}(t') - \mathbf{y}\right\|_2 \le \sqrt{n} c_{\mathbf{u}}$ holds for all $t' \in [0,t]$, by Lemma C.6, we have

$$\left\|\overrightarrow{\mathbf{w}}_r(t') - \overrightarrow{\mathbf{w}}_r(0)\right\|_2 \le R, \quad \forall\, 0 \le t' \le t+1. \tag{55}$$

Define two sets of indices

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

We have

$$\begin{aligned}
\widehat{\mathbf{y}}_i(t+1) - \widehat{\mathbf{y}}_i(t) &= \frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r\left(\sigma\left(\overrightarrow{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\overrightarrow{\mathbf{x}}_i\right) - \sigma\left(\overrightarrow{\mathbf{w}}_{\mathbf{S},r}^\top(t)\overrightarrow{\mathbf{x}}_i\right)\right) \\
&= \underbrace{\frac{1}{\sqrt{m}}\sum_{r \in E_{i,R}} a_r\left(\sigma\left(\overrightarrow{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\overrightarrow{\mathbf{x}}_i\right) - \sigma\left(\overrightarrow{\mathbf{w}}_{\mathbf{S},r}^\top(t)\overrightarrow{\mathbf{x}}_i\right)\right)}_{\coloneqq \mathbf{D}_i^{(1)}} \\
&\quad + \underbrace{\frac{1}{\sqrt{m}}\sum_{r \in \bar{E}_{i,R}} a_r\left(\sigma\left(\overrightarrow{\mathbf{w}}_{\mathbf{S},r}^\top(t+1)\overrightarrow{\mathbf{x}}_i\right) - \sigma\left(\overrightarrow{\mathbf{w}}_{\mathbf{S},r}^\top(t)\overrightarrow{\mathbf{x}}_i\right)\right)}_{\coloneqq \mathbf{E}_i^{(1)}} \\
&= \mathbf{D}_i^{(1)} + \mathbf{E}_i^{(1)},
\end{aligned} \tag{56}$$

and $\mathbf{D}^{(1)}, \mathbf{E}^{(1)} \in \mathbb{R}^n$ is a vector with their i-th element being $\mathbf{D}_i^{(1)}$ and $\mathbf{E}_i^{(1)}$ defined on the RHS of (56). Now we derive the upper bound for $\mathbf{E}_i^{(1)}$. For all $i \in [n]$ we have

$$\begin{aligned}
\left|\mathbf{E}_i^{(1)}\right| &= \left|\frac{1}{\sqrt{m}}\sum_{r \in \bar{E}_{i,R}} a_r\left(\sigma\left(\overrightarrow{\mathbf{w}}_{\mathbf{S},r}(t+1)^\top\overrightarrow{\mathbf{x}}_i\right) - \sigma\left(\overrightarrow{\mathbf{w}}_{\mathbf{S},r}(t)^\top\overrightarrow{\mathbf{x}}_i\right)\right)\right| \\
&\le \frac{1}{\sqrt{m}}\sum_{r \in \bar{E}_{i,R}} \left|\overrightarrow{\mathbf{w}}_{\mathbf{S},r}(t+1)^\top\overrightarrow{\mathbf{x}}_i - \overrightarrow{\mathbf{w}}_{\mathbf{S},r}(t)^\top\overrightarrow{\mathbf{x}}_i\right| \\
&\le \frac{1}{\sqrt{m}}\sum_{r \in \bar{E}_{i,R}} \left\|\overrightarrow{\mathbf{w}}_{\mathbf{S},r}(t+1) - \overrightarrow{\mathbf{w}}_{\mathbf{S},r}(t)\right\|_2 \\
&\overset{①}{=} \frac{1}{\sqrt{m}}\sum_{r \in \bar{E}_{i,R}} \left\|\frac{\eta}{n}\left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d+1:rd]}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)\right\|_2
\end{aligned}$$

$$\overset{\textcircled{2}}{\leq} \frac{c_{\mathbf{u}}}{\sqrt{m}} \sum_{r \in \bar{E}_{i,R}} \frac{\eta}{\sqrt{m}} \leq \eta c_{\mathbf{u}} \cdot \frac{\left|\bar{E}_{i,R}\right|}{m}. \tag{57}$$

Here ①, ② follow from (74) and (75) in the proof of Lemma C.6.

Let m be sufficiently large such that $R \leq R_0$ for the absolute positive constant $R_0 < \kappa$ specified in Theorem 6.1. Since $\mathbf{W}(0) \in \mathcal{W}_0$, we have

$$\sup_{\mathbf{x} \in \mathcal{X}} \left| \widehat{v}_R(\mathbf{W}(0), \mathbf{x}) - \frac{2R}{\sqrt{2\pi}\kappa} \right| \leq C_2(m/2, d, 1/n), \tag{58}$$

where $\widehat{v}_R(\mathbf{W}(0), \mathbf{x}) = \frac{1}{m} \sum\limits_{r=1}^{m} \mathbb{1}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \mathbf{x}\right| \leq R\right\}}$, so that $\widehat{v}_R(\mathbf{W}(0), \vec{\mathbf{x}}_i) = \left|\bar{E}_{i,R}\right|/m$. It follows from (57), (58) and the induction hypothesis that

$$\left|\mathbf{E}_i^{(1)}\right| \leq \eta c_{\mathbf{u}} \left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n) \right). \tag{59}$$

It follows from (59) that $\left\|\mathbf{E}^{(1)}\right\|_2$ can be bounded by

$$\left\|\mathbf{E}^{(1)}\right\|_2 \leq \eta c_{\mathbf{u}} \sqrt{n} \left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n) \right). \tag{60}$$

$\mathbf{D}_i^{(1)}$ on the RHS of (56) is expressed by

$$\begin{aligned}
\mathbf{D}_i^{(1)} &= \frac{1}{\sqrt{m}} \sum_{r \in E_{i,R}} a_r \left(\sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t+1) \vec{\mathbf{x}}_i \right) - \sigma\left(\vec{\mathbf{w}}_{\mathbf{S},r}^\top(t) \vec{\mathbf{x}}_i \right) \right) \\
&= \frac{1}{\sqrt{m}} \sum_{r \in E_{i,R}} a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(\vec{\mathbf{w}}_{\mathbf{S},r}(t+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t) \right)^\top \vec{\mathbf{x}}_i \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(-\frac{\eta}{n} \left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d:rd]} \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) \right)^\top \vec{\mathbf{x}}_i \\
&\quad + \frac{1}{\sqrt{m}} \sum_{r \in \bar{E}_{i,R}} a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(\frac{\eta}{n} \left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d:rd]} \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) \right)^\top \vec{\mathbf{x}}_i \\
&= \underbrace{-\frac{\eta}{n} \left[\mathbf{H}(t)\right]_i \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)}_{:=\mathbf{D}_i^{(2)}} \\
&\quad \underbrace{+ \frac{1}{\sqrt{m}} \sum_{r \in \bar{E}_{i,R}} a_r \mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \geq 0\right\}} \left(\frac{\eta}{n} \left[\mathbf{Z}_{\mathbf{S}}(t)\right]_{[(r-1)d:rd]} \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) \right)^\top \vec{\mathbf{x}}_i}_{:=\mathbf{E}_i^{(2)}} \\
&= \mathbf{D}_i^{(2)} + \mathbf{E}_i^{(2)},
\end{aligned} \tag{61}$$

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

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

Let $\mathbf{D}^{(2)}, \mathbf{E}^{(2)} \in \mathbb{R}^n$ be a vector with their i-the element being $\mathbf{D}_i^{(2)}$ and $\mathbf{E}_i^{(2)}$ defined on the RHS of (61). $\mathbf{E}^{(2)}$ can be expressed by $\mathbf{E}^{(2)} = \frac{\eta}{n} \tilde{\mathbf{E}}^{(2)} \left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right)$ with $\tilde{\mathbf{E}}^{(2)} \in \mathbb{R}^{n \times n}$ and

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

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

$$\left\|\tilde{\mathbf{E}}^{(2)}\right\|_2 \le \left\|\tilde{\mathbf{E}}^{(2)}\right\|_{\mathrm{F}} \le n\frac{\left|\bar{E}_{i,R}\right|}{m} \overset{①}{\le} n\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right), \tag{62}$$

where ① follows from (58). Also, $\|\mathbf{H}(t)\|_2 \le \|\mathbf{H}(t)\|_{\mathrm{F}} \le \sqrt{nN}$ for all $t \ge 0$. It follows from (62) that $\left\|\mathbf{E}^{(2)}\right\|_2$ can be bounded by

$$\left\|\mathbf{E}^{(2)}\right\|_2 \le \frac{\eta}{n}\left\|\tilde{\mathbf{E}}^{(2)}\right\|_2 \|\mathbf{y}(t) - \mathbf{y}\|_2 \le \eta c_{\mathbf{u}}\sqrt{n}\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right). \tag{63}$$

$\mathbf{D}_i^{(2)}$ on the RHS of (61) is expressed by

$$\begin{aligned}\mathbf{D}^{(2)} &= -\frac{\eta}{n}\mathbf{H}(t)\left(\hat{\mathbf{y}}(t) - \mathbf{y}\right) = \underbrace{-\frac{\eta}{n}\mathbf{K}\left(\hat{\mathbf{y}}(t) - \mathbf{y}\right)}_{:=\mathbf{D}^{(3)}} + \underbrace{\frac{\eta}{n}\left(\mathbf{K} - \mathbf{H}(0)\right)\left(\hat{\mathbf{y}}(t) - \mathbf{y}\right)}_{:=\mathbf{E}^{(3)}} + \underbrace{\frac{\eta}{n}\left(\mathbf{H}(0) - \mathbf{H}(t)\right)\left(\hat{\mathbf{y}}(t) - \mathbf{y}\right)}_{:=\mathbf{E}^{(4)}} \\ &= \mathbf{D}^{(3)} + \mathbf{E}^{(3)} + \mathbf{E}^{(4)}.\end{aligned} \tag{64}$$

On the RHS of (64), $\mathbf{D}^{(3)}, \mathbf{E}^{(3)}, \mathbf{E}^{(4)} \in \mathbb{R}^n$ are vectors which are analyzed as follows. $\left\|\tilde{\mathbf{E}}^{(3)}\right\|_2$ is bounded by

$$\|\mathbf{K} - \mathbf{H}(0)\|_2 \le \|\mathbf{K} - \mathbf{H}(0)\|_F \le nC_1(m/2, d, 1/n), \tag{65}$$

where the last inequality holds due to $\mathbf{W}(0) \in \mathcal{W}_0$.

In order to bound $\mathbf{E}^{(4)}$, we first estimate the upper bound for $|\mathbf{H}_{ij}(t) - \mathbf{H}_{ij}(0)|$ for all $i, j \in [n]$. We note that

$$\mathbb{1}_{\left\{\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i\right\}} \neq \mathbb{1}_{\left\{\mathbf{w}_r(0)^\top \vec{\mathbf{x}}_i\right\}}\right\}} \le \mathbb{1}_{\left\{\left|\mathbf{w}_r(0)^\top \vec{\mathbf{x}}_i\right| \le R\right\}} + \mathbb{1}_{\left\{\left\|\mathbf{w}_{\mathbf{S},r}(t) - \vec{\mathbf{w}}_r(0)\right\|_2 > R\right\}}. \tag{66}$$

It follows from (66) that

$$\begin{aligned}&|\mathbf{H}_{ij}(t) - \mathbf{H}_{ij}(0)| \\ &= \left|\frac{\vec{\mathbf{x}}_i^\top \vec{\mathbf{x}}_j}{m}\sum_{r=1}^m\left(\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \ge 0\right\}}\mathbb{1}_{\left\{\vec{\mathbf{w}}_r(t)^\top \vec{\mathbf{x}}_j \ge 0\right\}} - \mathbb{1}_{\left\{\mathbf{w}_r(0)^\top \vec{\mathbf{x}}_i \ge 0\right\}}\mathbb{1}_{\left\{\mathbf{w}_r(0)^\top \vec{\mathbf{x}}_j \ge 0\right\}}\right)\right| \\ &\le \frac{1}{m}\sum_{r=1}^m\left(\mathbb{1}_{\left\{\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_i \ge 0\right\}} \neq \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_i \ge 0\right\}}\right\}} + \mathbb{1}_{\left\{\mathbb{1}_{\left\{\vec{\mathbf{w}}_{\mathbf{S},r}(t)^\top \vec{\mathbf{x}}_j \ge 0\right\}} \neq \mathbb{1}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j \ge 0\right\}}\right\}}\right) \\ &\le \frac{1}{m}\sum_{r=1}^m\left(\mathbb{1}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_i\right| \le R\right\}} + \mathbb{1}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j\right| \le R\right\}} + 2\mathbb{1}_{\left\{\left\|\mathbf{w}_{\mathbf{S},r}(t) - \vec{\mathbf{w}}_r(0)\right\|_2 > R\right\}}\right) \\ &\le v_R(\mathbf{W}(0), \vec{\mathbf{x}}_i) + v_R(\mathbf{W}(0), \vec{\mathbf{x}}_j) \overset{①}{\le} \frac{4R}{\sqrt{2\pi}\kappa} + 2C_2(m/2, d, 1/n),\end{aligned} \tag{67}$$

where ① follows from (58).

It follows from (65) and (67) that $\left\|\mathbf{E}^{(3)}\right\|_2, \left\|\mathbf{E}^{(4)}\right\|_2$ are bounded by

$$\left\|\mathbf{E}^{(3)}\right\|_2 \le \frac{\eta}{n}\|\mathbf{K} - \mathbf{H}(0)\|_2\|\hat{\mathbf{y}}(t) - \mathbf{y}\|_2 \le \frac{\eta}{n} \cdot nC_1(m/2, d, 1/n) \cdot \|\mathbf{y}(t) - \mathbf{y}\|_2 \le \eta c_{\mathbf{u}}\sqrt{n}C_1(m/2, d, 1/n), \tag{68}$$

$$\begin{aligned}\left\|\mathbf{E}^{(4)}\right\|_2 &\le \frac{\eta}{n}\|\mathbf{H}(0) - \mathbf{H}(t)\|_2\|\hat{\mathbf{y}}(t) - \mathbf{y}\|_2 \le \frac{\eta}{n} \cdot n\left(\frac{4R}{\sqrt{2\pi}\kappa} + 2C_2(m/2, d, 1/n)\right) \cdot \|\mathbf{y}(t) - \mathbf{y}\|_2 \\ &\le \eta c_{\mathbf{u}}\sqrt{n}\left(\frac{4R}{\sqrt{2\pi}\kappa} + 2C_2(m/2, d, 1/n)\right).\end{aligned} \tag{69}$$

It follows from (61) and (64) that

$$\mathbf{D}_i^{(1)} = \mathbf{D}_i^{(3)} + \mathbf{E}_i^{(2)} + \mathbf{E}_i^{(3)} + \mathbf{E}_i^{(4)}. \tag{70}$$

It then follows from (56) that

$$\begin{aligned}\widehat{\mathbf{y}}_i(t+1) - \widehat{\mathbf{y}}_i(t) &= \mathbf{D}_i^{(1)} + \mathbf{E}_i^{(1)} = \mathbf{D}_i^{(3)} + \underbrace{\mathbf{E}_i^{(1)} + \mathbf{E}_i^{(2)} + \mathbf{E}_i^{(3)} + \mathbf{E}_i^{(4)}}_{:=\mathbf{E}_i} \\ &= -\frac{\eta}{n}\mathbf{K}\left(\widehat{\mathbf{y}}(t) - \mathbf{y}\right) + \mathbf{E}_i, \end{aligned} \tag{71}$$

where $\mathbf{E} \in \mathbb{R}^n$ with its i-th element being $\mathbf{E}_i$, and $\mathbf{E} = \mathbf{E}^{(1)} + \mathbf{E}^{(2)} + \mathbf{E}^{(3)} + \mathbf{E}^{(4)}$. It then follows from (60), (63), (68), and (69) that

$$\|\mathbf{E}\|_2 \le \eta c_{\mathbf{u}}\sqrt{n}\left(4\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2,d,1/n)\right) + C_1(m/2,d,1/n)\right). \tag{72}$$

Finally, (71) can be rewritten as

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

which proves (53). The upper bound for $\|\mathbf{E}\|_2$ in (54) follows from (72), Theorem 6.1, and noting that $\eta c_{\mathbf{u}} \le \Theta(1)$.

□

Lemma C.6. Suppose that $t \in [0:T-1]$ for $T \ge 1$, and $\|\widehat{\mathbf{y}}(t') - \mathbf{y}\|_2 \le \sqrt{n}c_{\mathbf{u}}$ holds for all $0 \le t' \le t$. Then

$$\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t') - \vec{\mathbf{w}}_r(0)\right\|_2 \le R, \quad \forall\, 0 \le t' \le t+1. \tag{73}$$

Proof. Let $[\mathbf{Z}_{\mathbf{S}}(t)]_{[(r-1)d:rd]}$ denotes the submatrix of $\mathbf{Z}_{\mathbf{S}}(t)$ formed by the the rows of $\mathbf{Z}_{\mathbf{Q}}(t)$ with row indices in $[(r-1)d : rd]$. By the GD update rule we have for every $t'' \in [0, T-1]$ that

$$\vec{\mathbf{w}}_{\mathbf{S},r}(t''+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t'') = -\frac{\eta}{n}\left[\mathbf{Z}_{\mathbf{S}}(t'')\right]_{[(r-1)d:rd]}\left(\widehat{\mathbf{y}}(t'') - \mathbf{y}\right), \tag{74}$$

We have $\left\|\left[\mathbf{Z}_{\mathbf{S}}(t'')\right]_{[(r-1)d:rd]}\right\|_2 \le \sqrt{n/m}$. It then follows from (74) that

$$\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t''+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t'')\right\|_2 \le \frac{\eta}{n}\left\|\left[\mathbf{Z}_{\mathbf{S}}(t'')\right]_{[(r-1)d:rd]}\right\|_2 \|\widehat{\mathbf{y}}(t'') - \mathbf{y}\|_2 \le \frac{\eta c_{\mathbf{u}}}{\sqrt{m}}, \forall t'' \in [0:t]. \tag{75}$$

Note that (73) trivially holds for $t' = 0$. For $t' \in [1, t+1]$, it follows from (75) that

$$\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t') - \vec{\mathbf{w}}_r(0)\right\|_2 \le \sum_{t''=0}^{t'-1}\left\|\vec{\mathbf{w}}_{\mathbf{S},r}(t''+1) - \vec{\mathbf{w}}_{\mathbf{S},r}(t'')\right\|_2 \le \frac{\eta}{\sqrt{m}}\sum_{t''=0}^{t'-1} c_{\mathbf{u}} \le \frac{\eta c_{\mathbf{u}} T}{\sqrt{m}} = R, \tag{76}$$

which completes the proof. □

Lemma C.7. Let $h(\cdot) = \sum_{t'=0}^{t-1} h(\cdot, t')$ for $t \in [T]$, $T \le \widehat{T}$ where

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

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

$$\|h\|_{\mathcal{H}_K} \le B_h = \mu_0 + 1 + \sqrt{2}, \tag{77}$$

and B_h is also defined in (33).

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

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

Let $\mathbf{\Sigma}$ be the diagonal matrix containing eigenvalues of $\mathbf{K}_n$, we then have

$$\sum_{t'=0}^{t-1} v(\mathbf{x},t') = \frac{\eta}{n}\sum_{j=1}^{n}\sum_{t'=0}^{t-1}\left[\left(\mathbf{I} - \eta\mathbf{K}_n\right)^{t'} f^*(\mathbf{S})\right]_j K(\overrightarrow{\mathbf{x}}_j, \mathbf{x}) = \frac{\eta}{n}\sum_{j=1}^{n}\sum_{t'=0}^{t-1}\left[\mathbf{U}\left(\mathbf{I} - \eta\mathbf{\Sigma}\right)^{t'}\mathbf{U}^\top f^*(\mathbf{S})\right]_j K(\overrightarrow{\mathbf{x}}_j, \mathbf{x}). \tag{78}$$

It follows from (78) that

$$\begin{aligned}\left\|\sum_{t'=0}^{t-1} v(\cdot,t')\right\|_{\mathcal{H}_K}^2 &= \frac{\eta^2}{n^2} f^*(\mathbf{S})^\top\mathbf{U}\sum_{t'=0}^{t-1}\left(\mathbf{I} - \eta\mathbf{\Sigma}\right)^{t'}\mathbf{U}^\top\mathbf{K}\mathbf{U}\sum_{t'=0}^{t-1}\left(\mathbf{I} - \eta\mathbf{\Sigma}\right)^{t'}\mathbf{U}^\top f^*(\mathbf{S}) \\ = \frac{1}{n}\left\|\eta\left(\mathbf{K}_n\right)^{1/2}\mathbf{U}\sum_{t'=0}^{t-1}\left(\mathbf{I} - \eta\mathbf{\Sigma}\right)^{t'}\mathbf{U}^\top f^*(\mathbf{S})\right\|_2^2 &\le \frac{1}{n}\sum_{i=1}^{n}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 \le \mu_0^2,\end{aligned} \tag{79}$$

where the last inequality follows from Lemma C.12.

Similarly, we have

$$\left\|\sum_{t'=0}^{t-1}\widehat{e}_1(\cdot,t')\right\|_{\mathcal{H}_K}^2 \le \frac{1}{n}\sum_{i=1}^{n}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i}\left[\mathbf{U}^\top\mathbf{w}\right]_i^2. \tag{80}$$

It then follows from the argument in the proof of (Raskutti et al., 2014, Lemma 9) that the RHS of (80) is bounded with high probability. We define a diagonal matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ with $\mathbf{R}_{ii} = \left(1-(1-\eta\widehat{\lambda}_i)^t\right)^2/\widehat{\lambda}_i$ for $i \in [n]$. Then the RHS of (80) is $1/n \cdot \mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)$. It follows from (Wright, 1973) that

$$\Pr\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right) - \mathbb{E}\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)\right] \ge u\right] \le \exp\left(-c\min\left\{nu/\|\mathbf{R}\|_2, n^2u^2/\|\mathbf{R}\|_{\mathrm{F}}^2\right\}\right) \tag{81}$$

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

$$\begin{aligned}\mathbb{E}\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)\right] &\le \frac{\sigma_0^2}{n}\sum_{i=1}^{n}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i} \overset{①}{\le} \frac{\sigma_0^2}{n}\sum_{i=1}^{n}\min\left\{\frac{1}{\widehat{\lambda}_i}, \eta_t^2\widehat{\lambda}_i\right\} \le \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^{n}\min\left\{\frac{1}{\eta_t\widehat{\lambda}_i}, \eta_t\widehat{\lambda}_i\right\} \\ &\overset{②}{\le} \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^{n}\min\left\{1, \eta_t\widehat{\lambda}_i\right\} = \frac{\sigma_0^2\eta_t^2}{n}\sum_{i=1}^{n}\min\left\{\eta_t^{-1}, \widehat{\lambda}_i\right\} = \sigma_0^2\eta_t^2\widehat{R}_K^2(\sqrt{1/\eta_t}) \le 1.\end{aligned} \tag{82}$$

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

$$\|\mathbf{R}\|_2 \le \max_{i\in[n]}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^2}{\widehat{\lambda}_i} \le \max_{i\in[n]}\min\left\{\frac{1}{\widehat{\lambda}_i}, \eta_t^2\widehat{\lambda}_i\right\} \le \eta_t. \tag{83}$$

We also have

$$\begin{aligned}\frac{1}{n}\|\mathbf{R}\|_{\mathrm{F}}^2 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^4}{(\widehat{\lambda}_i)^2} \le \frac{\eta_t^3}{n}\sum_{i=1}^{n}\min\left\{\frac{1}{\eta_t^3\widehat{\lambda}_i^2},\eta_t\widehat{\lambda}_i^2\right\}\\ &\overset{\textcircled{3}}{\le} \frac{\eta_t^3}{n}\sum_{i=1}^{n}\min\left\{\widehat{\lambda}_i,\frac{1}{\eta_t}\right\} = \eta_t^3\widehat{R}_K^2(\sqrt{1/\eta_t}) \le \frac{\eta_t}{\sigma_0^2},\end{aligned} \tag{84}$$

where ③ follows from

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

Combining (80)-(84) with $u=1$ in (81), we have

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

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

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

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

so that

$$\left\|\sum_{t'=0}^{t-1}\widehat{e}_2(\cdot,t')\right\|_{\mathcal{H}_K}\le\sum_{t'=0}^{t-1}\|\widehat{e}_2(\cdot,t')\|_{\mathcal{H}_K}\le T\eta\sqrt{\widehat{\lambda}_1}\tau\le 1, \tag{85}$$

if $\tau\lesssim 1/(\eta T)$.

Finally, we have

$$\|h\|_{\mathcal{H}_K}\le\left\|\sum_{t'=0}^{t-1}\widehat{v}(\cdot,t')\right\|_{\mathcal{H}_K}+\left\|\sum_{t'=0}^{t-1}\widehat{e}_1(\cdot,t')\right\|_{\mathcal{H}_K}+\left\|\sum_{t'=0}^{t-1}\widehat{e}_2(\cdot,t')\right\|_{\mathcal{H}_K}\le\mu_0+1+\sqrt{2}=B_h.$$

□

Theorem C.8. For every $t\in[T]$, let the neural network $f(\cdot)=f(\mathbf{W}(t),\cdot)$ be trained by gradient descent with the learning rate $\eta=\Theta(1)\in(0,1/\widehat{\lambda}_1)$ on the random initialization $\mathbf{W}(0)\in\mathcal{W}_0$ with $T\le\widehat{T}$. Then with probability at least $1-\exp\left(-\Theta(n)\right)-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$, $f\in\mathcal{F}_{\mathrm{NN}}(\mathbf{S},\mathbf{W}(0),T)$, and f can be decomposed by

$$f=h+e\in\mathcal{F}(B_h,w), \tag{86}$$

where $h\in\mathcal{H}_K(B_h)$ with B_h defined in (33), $e\in L^\infty$. When

$$m\gtrsim\max\left\{T^8d^{\frac{8}{3}}/w^{\frac{16}{3}},T^{\frac{40}{3}}d^{\frac{8}{3}}\right\}, \tag{87}$$

then

$$\|e\|_\infty\le w. \tag{88}$$

In addition,

$$\|f\|_\infty\le\frac{B_h}{\sqrt{2}}+w. \tag{89}$$

Remark. We consider the kernel regression problem with the training loss $L(\boldsymbol{\alpha}) = 1/2 \cdot \|\mathbf{K}_n \boldsymbol{\alpha} - \mathbf{y}\|_2^2$. Letting $\boldsymbol{\beta} = \mathbf{K}_n^{1/2} \boldsymbol{\alpha}$ and then performing GD on $\boldsymbol{\beta}$ with this training loss and the learning rate η, it can be verified that the kernel regressor right after the t-th step of GD is

$$\widehat{f}_t^{(\mathrm{NTK})} = \frac{\eta}{n} \sum_{t'=0}^{t-1} \sum_{i=1}^{n} K(\cdot, \overset{\rightharpoonup}{\mathbf{x}}_i) \boldsymbol{\alpha}_i^{(t')}, \tag{90}$$

where $\boldsymbol{\alpha}^{(t')} = (\mathbf{I}_n - \eta \mathbf{K}_n)^{t'} \mathbf{y}$. Following from the proof of Lemma C.6 and Theorem C.8, under the conditions of Theorem C.8 we have

$$h_t = \widehat{f}_t^{(\mathrm{NTK})} + \widehat{e}_2(\cdot, t),$$

where $\widehat{e}_2(\cdot, t) = \frac{\eta}{n} \sum_{t'=0}^{t-1} \sum_{j=1}^{n} K(\cdot, \overset{\rightharpoonup}{\mathbf{x}}_j) \left[\overset{\rightharpoonup}{\mathbf{e}}_2(t') \right]_j$ and $\overset{\rightharpoonup}{\mathbf{e}}_2(t')$ appears in the definition of $\mathcal{E}_{t,\tau}$ in (29). It is remarked that in our analysis, we approximate f_t by $h_t \in \mathcal{H}_K(B_h)$ with a small approximation error w, and we do not need to approximate f_t by the kernel regressor $\widehat{f}_t^{(\mathrm{NTK})}$ with a sufficiently small approximation error which is the common strategy used in existing works (Hu et al., 2021; Suh et al., 2022; Li et al., 2024). In fact, our analysis only requires m is suitably large so that the $\mathcal{H}_K$-norm of $\widehat{e}_2(\cdot, t) = h_t - \widehat{f}_t^{(\mathrm{NTK})}$ is bounded by a positive constant rather than an infinitesimal number as $m \to \infty$, that is, $\|\widehat{e}_2(\cdot, t)\|_{\mathcal{H}_K} \le 1$, which is revealed by the proof of Lemma C.7.

Proof. It follows from Lemma C.3 and its proof that conditioned on an event with probability at least$1 - \exp(-\Theta(n))$, $f \in \mathcal{F}_{\mathrm{NN}}(\mathbf{S}, \mathbf{W}(0), T)$ with $\mathbf{W}(0) \in \mathcal{W}_0$. Moreover, $f(\cdot) = f(\mathbf{W}, \cdot)$ with $\mathbf{W} = \left\{ \overset{\rightharpoonup}{\mathbf{w}}_r \right\}_{r=1}^{m} \in \mathcal{W}(\mathbf{S}, \mathbf{W}(0), T)$, and $\mathrm{vec}(\mathbf{W}) = \mathrm{vec}(\mathbf{W}_{\mathbf{S}}) = \mathrm{vec}(\mathbf{W}(0)) - \sum_{t'=0}^{t-1} \eta/n \cdot \mathbf{Z}_{\mathbf{S}}(t') \mathbf{u}(t')$ 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 $t' \in [0, t-1]$.

$\overset{\rightharpoonup}{\mathbf{w}}_r$ is expressed as

$$\overset{\rightharpoonup}{\mathbf{w}}_r = \overset{\rightharpoonup}{\mathbf{w}}_{\mathbf{S},r}(t) = \overset{\rightharpoonup}{\mathbf{w}}_r(0) - \sum_{t'=0}^{t-1} \frac{\eta}{n} \left[\mathbf{Z}_{\mathbf{S}}(t') \right]_{[(r-1)d:rd]} \mathbf{u}(t'), \tag{91}$$

where the notation $\overset{\rightharpoonup}{\mathbf{w}}_{\mathbf{S},r}$ emphasizes that $\overset{\rightharpoonup}{\mathbf{w}}_r$ depends on the training data $\mathbf{S}$.

We define the event

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

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

$$\begin{aligned}
&|f(\mathbf{W}, \mathbf{x}) - g(\mathbf{x})| \\
&= \frac{1}{\sqrt{m}} \left| \sum_{r=1}^{m} a_r \sigma\left(\overset{\rightharpoonup}{\mathbf{w}}_r^\top \mathbf{x} \right) - \sum_{r=1}^{m} a_r \mathbb{I}_{\left\{ \overset{\rightharpoonup}{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0 \right\}} \overset{\rightharpoonup}{\mathbf{w}}_r^\top \mathbf{x} \right| \\
&\le \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \left| a_r \left(\mathbb{I}_{\{E_r(R)\}} + \mathbb{I}_{\{\bar{E}_r(R)\}} \right) \left(\sigma\left(\overset{\rightharpoonup}{\mathbf{w}}_r^\top \mathbf{x} \right) - \mathbb{I}_{\left\{ \overset{\rightharpoonup}{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0 \right\}} \overset{\rightharpoonup}{\mathbf{w}}_r^\top \mathbf{x} \right) \right| \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}_{\{E_r(R)\}} \left| \sigma\left(\overset{\rightharpoonup}{\mathbf{w}}_r^\top \mathbf{x} \right) - \mathbb{I}_{\left\{ \overset{\rightharpoonup}{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0 \right\}} \overset{\rightharpoonup}{\mathbf{w}}_r^\top \mathbf{x} \right| \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}_{\{E_r(R)\}} \left| \sigma\left(\overset{\rightharpoonup}{\mathbf{w}}_r^\top \mathbf{x} \right) - \sigma\left(\overset{\rightharpoonup}{\mathbf{w}}_r(0)^\top \mathbf{x} \right) - \mathbb{I}_{\left\{ \overset{\rightharpoonup}{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0 \right\}} \left(\overset{\rightharpoonup}{\mathbf{w}}_r - \overset{\rightharpoonup}{\mathbf{w}}_r(0) \right)^\top \mathbf{x} \right| \\
&\le \frac{2R}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{I}_{\{E_r(R)\}}.
\end{aligned} \tag{92}$$

Plugging $R = \frac{\eta c_{\mathbf{u}} T}{\sqrt{m}}$ in (92), we have

$$\begin{aligned} |f(\mathbf{W}, \mathbf{x}) - g(\mathbf{x})| &\le \frac{2R}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{1}_{\{E_r(R)\}} = 2\eta c_{\mathbf{u}} T \cdot \frac{1}{m} \sum_{r=1}^{m} \mathbb{1}_{\{E_r(R)\}} \\ &= 2\eta c_{\mathbf{u}} T \cdot \widehat{v}_R(\mathbf{W}(0), \mathbf{x}) \le 2\eta c_{\mathbf{u}} T \left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n) \right). \end{aligned} \tag{93}$$

Using (91), we can express $g(\mathbf{x})$ as

$$\begin{aligned} g(\mathbf{x}) &= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \overrightarrow{\mathbf{w}}_r(0)^\top \mathbf{x} - \sum_{t'=0}^{t-1} \frac{1}{\sqrt{m}} \sum_{r=1}^{m} \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \left(\frac{\eta}{n} \left[\mathbf{Z}_{\mathbf{S}}(t')\right]_{[(r-1)d:rd]} \mathbf{u}(t') \right)^\top \mathbf{x} \\ &\overset{①}{=} - \sum_{t'=0}^{t-1} \underbrace{\frac{\eta}{nm} \sum_{r=1}^{m} \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \sum_{j=1}^{n} \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(t')^\top \overrightarrow{\mathbf{x}}_j \ge 0\right\}} \mathbf{u}_j(t') \overrightarrow{\mathbf{x}}_j^\top \mathbf{x},}_{:= G_{t'}(\mathbf{x})} \end{aligned} \tag{94}$$

where ① follows from the fact that $\frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \overrightarrow{\mathbf{w}}_r(0)^\top \mathbf{x} = f(\mathbf{W}(0), \mathbf{x}) = 0$ due to the particular initialization of the two-layer NN. For each $G_{t'}$ on the RHS of (94), we have

$$\begin{aligned} G_{t'}(\mathbf{x}) &\overset{②}{=} \frac{\eta}{nm} \sum_{r=1}^{m} \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \sum_{j=1}^{n} d_{t',r,j} \mathbf{u}_j(t') \overrightarrow{\mathbf{x}}_j^\top \mathbf{x} + \frac{\eta}{nm} \sum_{r=1}^{m} \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \sum_{j=1}^{n} \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(0)^\top \overrightarrow{\mathbf{x}}_j \ge 0\right\}} \mathbf{u}_j(t') \overrightarrow{\mathbf{x}}_j^\top \mathbf{x} \\ &\overset{③}{=} \frac{\eta}{n} \sum_{j=1}^{n} K(\mathbf{x}, \overrightarrow{\mathbf{x}}_j) \mathbf{u}_j(t') + \underbrace{\frac{\eta}{n} \sum_{j=1}^{n} q_j \mathbf{u}_j(t')}_{:= E_1(\mathbf{x})} + \underbrace{\frac{\eta}{nm} \sum_{r=1}^{m} \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} \sum_{j=1}^{n} d_{t',r,j} \mathbf{u}_j(t') \overrightarrow{\mathbf{x}}_j^\top \mathbf{x}}_{:= E_2(\mathbf{x})}. \end{aligned} \tag{95}$$

where

$$d_{t',r,j} \coloneqq \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(t')^\top \overrightarrow{\mathbf{x}}_j \ge 0\right\}} - \mathbb{1}_{\left\{\overrightarrow{\mathbf{w}}_r(0)^\top \overrightarrow{\mathbf{x}}_j \ge 0\right\}}$$

in ②, and and

$$q_j \coloneqq \widehat{h}(\mathbf{W}(0), \overrightarrow{\mathbf{x}}_j, \mathbf{x}) - K(\overrightarrow{\mathbf{x}}_j, \mathbf{x})$$

for all $j \in [n]$ in ③.

We now analyze each term on the RHS of (95). Let $h(\cdot, t') \colon \mathcal{X} \to \mathbb{R}$ be defined by

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

then $h(\cdot, t')$ is an element in the RKHS $\mathcal{H}_K$ for each $t' \in [0, t-1]$. We further define

$$h(\cdot) \coloneqq \sum_{t'=0}^{t-1} h(\cdot, t'), \tag{96}$$

Since $\mathbf{W}(0) \in \mathcal{W}_0$, $q_j \le C_1(m/2, d, 1/n)$ for all $j \in [n]$ with $C_1(m/2, d, 1/n)$ defined in (39). Moreover, $\mathbf{u}(t') \le c_{\mathbf{u}} \sqrt{n}$ with high probability, so that we have

$$\|E_1\|_\infty = \left\| \frac{\eta}{n} \sum_{j=1}^{n} q_j \mathbf{u}_j(t') \right\|_\infty \le \frac{\eta}{n} \|\mathbf{u}(t')\|_2 \sqrt{n} C_1(m/2, d, 1/n) \le \eta c_{\mathbf{u}} C_1(m/2, d, 1/n). \tag{97}$$

We now bound the last term on the RHS of (95). Define $\mathbf{X}' \in \mathbb{R}^{d \times n}$ with its j-column being $\mathbf{X}'^{(j)} = \frac{1}{m}\sum_{r=1}^{m} \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \mathbf{x} \ge 0\right\}} d_{t',r,j}\vec{\mathbf{x}}_j$ for all $j \in [n]$, then $E_2(\mathbf{x}) = \frac{\eta}{n}\left(\mathbf{X}'\mathbf{u}(t')\right)^\top \mathbf{x}$.

We need to derive the upper bound for $\|\mathbf{X}'\|_2$. Because $\left\|\vec{\mathbf{w}}_r - \vec{\mathbf{w}}_r(0)\right\|_2 \le R$, it follows that $\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(t')^\top \vec{\mathbf{x}}_j \ge 0\right\}} = \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j \ge 0\right\}}$ when $\left|\vec{\mathbf{w}}_r(0)^\top \mathbf{x}'_{j'}\right| > R$ for all $j' \in [n]$. Therefore,

$$|d_{t',r,j'}| = \left|\mathbb{I}_{\left\{\vec{\mathbf{w}}_r(t')^\top \vec{\mathbf{x}}_j \ge 0\right\}} - \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j \ge 0\right\}}\right| \le \mathbb{I}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j\right| \le R\right\}},$$

and it follows that

$$\begin{aligned}\frac{\left|\sum\limits_{r=1}^{m} \mathbb{I}_{\left\{\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_i \ge 0\right\}} d_{t',r,j}\right|}{m} &\le \frac{\sum\limits_{r=1}^{m} |d_{t',r,j}|}{m} \le \frac{\sum\limits_{r=1}^{m} \mathbb{I}_{\left\{\left|\vec{\mathbf{w}}_r(0)^\top \vec{\mathbf{x}}_j\right| \le R\right\}}}{m} = \widehat{v}_R(\mathbf{W}(0), \vec{\mathbf{x}}_j) \\ &\le \frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n),\end{aligned} \tag{98}$$

where $\widehat{v}_R$ is defined by (14), and the last inequality follows from Theorem C.2.

It follows from (98) that $\|\mathbf{X}'\|_2 \le \sqrt{n}\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right)$, and we have

$$\|E_2(\mathbf{x})\|_\infty \le \frac{\eta}{n}\|\mathbf{X}'\|_2\|\mathbf{u}(t')\|_2\|\mathbf{x}\|_2 \le \eta c_{\mathbf{u}}\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right). \tag{99}$$

Combining (95), (97), and (99), for any $t' \in [0, t-1]$,

$$\begin{aligned}&\|G_{t'}(\mathbf{x}) - h(\mathbf{x}, t')\|_\infty \le \|E_1\|_\infty + \|E_2\|_\infty \\ &\le \eta c_{\mathbf{u}}\left(C_1(m/2, d, 1/n) + \frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right).\end{aligned} \tag{100}$$

Define $e(\cdot) = f(\mathbf{W}, \cdot) - h(\cdot)$, it then follows from (93), (94), and (100) that

$$\begin{aligned}&\|e(\mathbf{x})\|_\infty \le \|f(\mathbf{W}, \cdot) - g\|_\infty + \|g - h\|_\infty \\ &\le \|f(\mathbf{W}, \cdot) - g\|_\infty + \sum_{t'=0}^{t-1} \|G_{t'} - h(\cdot, t')\|_\infty \\ &\overset{②}{\le} 2\eta c_{\mathbf{u}} T\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right) \\ &+ \eta c_{\mathbf{u}} T\left(C_1(m/2, d, 1/n) + \frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right) \\ &\le \eta c_{\mathbf{u}} T\left(C_1(m/2, d, 1/n) + 3\left(\frac{2R}{\sqrt{2\pi}\kappa} + C_2(m/2, d, 1/n)\right)\right) \\ &:= \Delta_{m,n,N,c_x,\eta,\tau}.\end{aligned} \tag{101}$$

We now give estimates for $\Delta_{m,n,N,c_x,\eta,\tau}$. Since $\mathbf{W}(0) \in \mathcal{W}_0$, it follows from Theorem 6.1 that $C_1(m/2, d, 1/n) \lesssim \sqrt{\frac{d\log m}{m}} \lesssim \sqrt{d}m^{-\frac{3}{16}}T^{\frac{1}{2}}$, and $2R/(\sqrt{2\pi}\kappa) + C_2(m/2, d, 1/n) \lesssim \sqrt{d}m^{-\frac{3}{16}}T^{\frac{1}{2}}$. As a result,

$$\Delta_{m,n,N,c_x,\eta,\tau} \lesssim \sqrt{d}m^{-\frac{3}{16}}T^{3/2}.$$

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

$$m \gtrsim T^8 d^{\frac{8}{3}}/w^{\frac{16}{3}},$$

we have $\Delta_{m,n,N,c_x,\eta,\tau} \le w$.

It follows from Lemma C.7 that with probability at least $1-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$,

$$\|h\|_{\mathcal{H}_K} \le B_h, \tag{102}$$

where B_h is defined in (33), and τ are required to satisfy

$$\tau \lesssim 1/(\eta T).$$

Lemma C.3 requires that $m \gtrsim T^8 d^{\frac{8}{3}}/\tau^{\frac{16}{3}}$. As a result, we have

$$m \gtrsim T^{\frac{40}{3}} d^{\frac{8}{3}}.$$

It also follows from the Cauchy-Schwarz inequality that $\|h\|_\infty \le B_h/\sqrt{2}$. This together with (101) proves (89). □

The following lemma gives the upper bound for the Rademacher complexity of a localized function class in $\mathcal{F}(B,w)$ comprising functions with their L^2-norm bounded by every $r>0$.

Lemma C.9. For every $B, w > 0$ every $r > 0$,

$$\mathfrak{R}\left(\left\{f \in \mathcal{F}(B,w)\colon \mathbb{E}_P\left[f^2\right] \le r\right\}\right) \le \varphi_{B,w}(r), \tag{103}$$

where

$$\varphi_{B,w}(r) \coloneqq \min_{Q\colon Q\ge 0}\left((\sqrt{r}+w)\sqrt{\frac{Q}{n}} + B\left(\frac{\sum\limits_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right) + w. \tag{104}$$

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

$$\begin{aligned}&\mathfrak{R}\left(\left\{f\colon f \in \mathcal{F}(B,w), \mathbb{E}_P\left[f^2\right] \le r\right\}\right)\\ &\le \underbrace{\frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}(B,w)\colon \mathbb{E}_P[f^2]\le r}\sum_{i=1}^n \sigma_i h(\overset{\rightharpoonup}{\mathbf{x}}_i)\right]}_{:=\mathcal{R}_1} + \underbrace{\frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}(B,w)\colon \mathbb{E}_P[f^2]\le r}\sum_{i=1}^n \sigma_i e(\overset{\rightharpoonup}{\mathbf{x}}_i)\right]}_{:=\mathcal{R}_2}.\end{aligned} \tag{105}$$

We now analyze the upper bounds for $\mathcal{R}_1, \mathcal{R}_2$ on the RHS of (105).

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

For any $f \in \mathcal{F}(B,w)$, we have $f = h + e$ with $h \in \mathcal{H}_K(B)$, $e \in L^\infty$, $\|e\|_\infty \le w$. When $\mathbb{E}_P\left[f^2\right] \le r$, it follows from the triangle inequality that $\|h\|_{L^2} \le \|f\|_{L^2} + \|e\|_{L^2} \le \sqrt{r} + w \coloneqq r_h$.

We now consider $h \in \mathcal{H}_K(B)$ with $\|h\|_{L^2} \le r_h$ in the remaining of this proof. We have

$$\sum_{i=1}^n \sigma_i f(\overset{\rightharpoonup}{\mathbf{x}}_i) = \sum_{i=1}^n \sigma_i\left(h(\overset{\rightharpoonup}{\mathbf{x}}_i) + e(\overset{\rightharpoonup}{\mathbf{x}}_i)\right) = \left\langle h, \sum_{i=1}^n \sigma_i K(\cdot, \overset{\rightharpoonup}{\mathbf{x}}_i)\right\rangle_{\mathcal{H}_K} + \sum_{i=1}^n \sigma_i e(\overset{\rightharpoonup}{\mathbf{x}}_i). \tag{106}$$

Because $\{v_q\}_{q\ge 1}$ is an orthonormal basis of $\mathcal{H}_K$, for any $0 \le Q \le n$, we further express the first term on the RHS of (106) as

$$\left\langle h, \sum_{i=1}^n \sigma_i K(\cdot, \overset{\rightharpoonup}{\mathbf{x}}_i)\right\rangle_{\mathcal{H}_K}$$

$$= \left\langle \sum_{q=1}^{Q} \sqrt{\lambda_q} \left\langle h, v_q \right\rangle_{\mathcal{H}_K} v_q, \sum_{q=1}^{Q} \frac{1}{\sqrt{\lambda_q}} \left\langle \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} v_q \right\rangle_{\mathcal{H}_K} + \left\langle h, \sum_{q>Q} \left\langle \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} v_q \right\rangle_{\mathcal{H}_K}. \tag{107}$$

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

$$\left\| \sum_{q=1}^{Q} \sqrt{\lambda_q} \left\langle h, v_q \right\rangle_{\mathcal{H}_K} v_q \right\|_{\mathcal{H}_K} = \left\| \sum_{q=1}^{Q} \sqrt{\lambda_q} \boldsymbol{\beta}_q^{(h)} v_q \right\|_{\mathcal{H}_K} = \sqrt{\sum_{q=1}^{Q} \lambda_q {\boldsymbol{\beta}_q^{(h)}}^2} \le \|h\|_{L^2} \le r_h. \tag{108}$$

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

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

so that we have

$$\begin{aligned} \left\langle \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} &= \left\langle \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{\infty} \lambda_j e_j e_j(\vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} = \left\langle \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{\infty} \sqrt{\lambda_j} e_j(\vec{\mathbf{x}}_i) \cdot v_j, v_q \right\rangle_{\mathcal{H}_K} \\ &= \sum_{i=1}^{n} \sigma_i \sqrt{\lambda_q} e_q(\vec{\mathbf{x}}_i). \end{aligned} \tag{109}$$

Combining (107), (108), and (109), we have

$$\begin{aligned} \left\langle h, \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i) \right\rangle &\overset{①}{\le} \left\| \sum_{q=1}^{Q} \sqrt{\lambda_q} \left\langle h, v_q \right\rangle_{\mathcal{H}_K} v_q \right\|_{\mathcal{H}_K} \cdot \left\| \sum_{q=1}^{Q} \frac{1}{\sqrt{\lambda_q}} \left\langle \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} v_q \right\|_{\mathcal{H}_K} \\ &+ \|h\|_{\mathcal{H}_K} \cdot \left\| \sum_{q=Q+1}^{\infty} \left\langle \sum_{i=1}^{n} \sigma_i K(\cdot, \vec{\mathbf{x}}_i), v_q \right\rangle_{\mathcal{H}_K} v_q \right\|_{\mathcal{H}_K} \\ &\le \|h\|_{L^2} \left\| \sum_{q=1}^{Q} \sum_{i=1}^{n} \sigma_i e_q(\vec{\mathbf{x}}_i) v_q \right\|_{\mathcal{H}_K} + B \left\| \sum_{q=Q+1}^{\infty} \sum_{i=1}^{n} \sigma_i \sqrt{\lambda_q} e_q(\vec{\mathbf{x}}_i) v_q \right\|_{\mathcal{H}_K} \\ &\le r_h \sqrt{\sum_{q=1}^{Q} \left(\sum_{i=1}^{n} \sigma_i e_q(\vec{\mathbf{x}}_i) \right)^2} + B \sqrt{\sum_{q=Q+1}^{\infty} \left(\sum_{i=1}^{n} \sigma_i \sqrt{\lambda_q} e_q(\vec{\mathbf{x}}_i) \right)^2}, \end{aligned} \tag{110}$$

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

$$\mathbb{E}\left[\sqrt{\sum_{q=1}^{Q} \left(\sum_{i=1}^{n} \sigma_i e_q(\vec{\mathbf{x}}_i) \right)^2} \right] \le \sqrt{\mathbb{E}\left[\sum_{q=1}^{Q} \left(\sum_{i=1}^{n} \sigma_i e_q(\vec{\mathbf{x}}_i) \right)^2 \right]} \le \sqrt{\mathbb{E}\left[\sum_{q=1}^{Q} \sum_{i=1}^{n} e_q^2(\vec{\mathbf{x}}_i) \right]} = \sqrt{nQ}. \tag{111}$$

and similarly,

$$\mathbb{E}\left[\sqrt{\sum_{q=Q+1}^{\infty} \left(\sum_{i=1}^{n} \sigma_i \sqrt{\lambda_q} e_q(\vec{\mathbf{x}}_i) \right)^2} \right] \le \sqrt{\mathbb{E}\left[\sum_{q=Q+1}^{\infty} \lambda_q \sum_{i=1}^{n} e_q^2(\vec{\mathbf{x}}_i) \right]} = \sqrt{n \sum_{q=Q+1}^{\infty} \lambda_q}. \tag{112}$$

Since (110)-(112) hold for all $Q \ge 0$, it follows that

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}_K(B), \|h\|_{L^2} \le r_h} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(\vec{\mathbf{x}}_i) \right] \le \min_{Q \colon Q \ge 0} \left(r_h \sqrt{nQ} + B \sqrt{n \sum_{q=Q+1}^{\infty} \lambda_q} \right). \tag{113}$$

It follows from (105), (106), and (113) that

$$\mathcal{R}_1 \le \frac{1}{n}\mathbb{E}\left[\sup_{h \in \mathcal{H}_K(B), \|h\|_{L^2} \le r_h} \sum_{i=1}^{n} \sigma_i h(\overrightarrow{\mathbf{x}}_i)\right] \le \min_{Q \colon Q \ge 0} \left(r_h \sqrt{\frac{Q}{n}} + B \left(\frac{\sum\limits_{q=Q+1}^{\infty} \lambda_q}{n} \right)^{1/2} \right). \tag{114}$$

Derivation for the upper bound for $\mathcal{R}_2$.

Because $\left|1/n \sum_{i=1}^{n} \sigma_i e(\overrightarrow{\mathbf{x}}_i)\right| \le w$ when $\|e\|_\infty \le w$, we have

$$\mathcal{R}_2 \le \frac{1}{n}\mathbb{E}\left[\sup_{e \in L^\infty \colon \|e\|_\infty \le w} \sum_{i=1}^{n} \sigma_i e(\overrightarrow{\mathbf{x}}_i)\right] \le w. \tag{115}$$

It follows from (114) and (115) that

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

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

Theorem C.10. Suppose $w \in (0,1)$ and m satisfy

$$m \gtrsim \max\left\{T^8 d^{\frac{8}{3}} / w^{\frac{16}{3}}, T^{\frac{40}{3}} d^{\frac{8}{3}}\right\}, \tag{116}$$

and the neural network $f(\mathbf{W}(t), \cdot)$ is trained by GD in Algorithm 1 with the learning rate $\eta = \Theta(1) \in (0, 1/\widehat{\lambda}_1)$ on random initialization $\mathbf{W}(0) \in \mathcal{W}_0$, and $T \le \widehat{T}$. Then for every $t \in [T]$, with probability at least $1 - \exp\left(-\Theta(n)\right) - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right) - \exp\left(-n\varepsilon_n^2\right)$ over the random noise $\mathbf{w}$ and the random training features $\mathbf{S}$,

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \le c_0 \min_{0 \le Q \le n} \left(\frac{B_0 Q}{n} + w\left(\sqrt{\frac{Q}{n}} + 1\right) + B_h \left(\frac{\sum\limits_{q=Q+1}^{\infty} \lambda_q}{n} \right)^{1/2} \right)^2 + c_0 \varepsilon_n^2. \tag{117}$$

Furthermore, with probability at least $1 - \exp\left(-\Theta(n)\right) - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right) - \exp\left(-n\varepsilon_n^2\right)$ over the random noise $\mathbf{w}$, the random training features $\mathbf{S}$,

$$\mathbb{E}_P\left[(f_t - f^*)^2\right] - 2\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \le c_0'(\varepsilon_n^2 + w). \tag{118}$$

Here B_0, c_0, c_0' are absolute positive constants depending on μ_0, and c_0' also depends on σ_0.

Proof. We first remark that the conditions on m, (116), is required by Lemma C.3 and Theorem C.8.

It follows from Lemma C.3 and Theorem C.8 that for every $t \in [T]$, conditioned on an event Ω with probability at least $1 - \exp\left(-\Theta(n)\right) - \exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$, we have $\mathbf{W}(t) \in \mathcal{W}(\mathbf{S}, \mathbf{W}(0), T)$, and

$$f(\mathbf{W}(t), \cdot) = f_t \in \mathcal{F}_{\rm NN}(\mathbf{S}, \mathbf{W}(0), T).$$

Moreover, conditioned on the event Ω,

$$f_t \in \mathcal{F}(B_h, w).$$

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

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

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

We have

$$\begin{aligned}\mathfrak{R}\left(\{F\in\mathcal{F}\colon T(F)\le r\}\right)&=\mathfrak{R}\left(\left\{(f-f^*)^2\colon f\in\mathcal{F}(B_h,w),\mathbb{E}_P\left[(f-f^*)^2\right]\le\frac{r}{B_0^2}\right\}\right)\\&\overset{①}{\le}2B_0\mathfrak{R}\left(\left\{f-f^*\colon f\in\mathcal{F}(B_h,w),\mathbb{E}_P\left[(f-f^*)^2\right]\le\frac{r}{B_0^2}\right\}\right)\\&\overset{②}{\le}4B_0\mathfrak{R}\left(\left\{f\in\mathcal{F}(B_h,w)\colon\mathbb{E}_P\left[f^2\right]\le\frac{r}{4B_0^2}\right\}\right),\end{aligned}\tag{119}$$

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

It follows from (119) and Lemma C.9 that

$$\begin{aligned}B_0^2\mathfrak{R}\left(\{F\in\mathcal{F}\colon T(F)\le r\}\right)&\le4B_0^3\mathfrak{R}\left(\left\{f\colon f\in\mathcal{F}(B_h,w),\mathbb{E}_P\left[f^2\right]\le\frac{r}{4B_0^2}\right\}\right)\\&\le4B_0^3\varphi_{B_h,w}\left(\frac{r}{4B_0^2}\right):=\psi(r).\end{aligned}\tag{120}$$

ψ defined as the RHS of (120) 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 $0\le r\le r^*$. 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 (104), for every $0\le Q\le n$, we have

$$\frac{r}{4B_0^3}\le\left(\frac{\sqrt{r}}{2B_0}+w\right)\sqrt{\frac{Q}{n}}+B_h\left(\frac{\sum\limits_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}+w.\tag{121}$$

Solving the quadratic inequality (121) for r, we have

$$r\le\frac{8B_0^4Q}{n}+8B_0^3\left(w\left(\sqrt{\frac{Q}{n}}+1\right)+B_h\left(\frac{\sum\limits_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right).\tag{122}$$

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

$$r\le8B_0^3\min_{0\le Q\le n}\left(\frac{B_0Q}{n}+w\left(\sqrt{\frac{Q}{n}}+1\right)+B_h\left(\frac{\sum\limits_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right).\tag{123}$$

It then follows from (120) and Theorem A.3 that with probability at least $1-\exp(-x)$ over the random training features $\mathbf{S}$,

$$\mathbb{E}_P\left[(f_t-f^*)^2\right]-\frac{K_0}{K_0-1}\mathbb{E}_{P_n}\left[(f_t-f^*)^2\right]-\frac{x\left(11B_0^2+26B_0^2K_0\right)}{n}\le\frac{704K_0}{B_0^2}r^*,\tag{124}$$

or

$$\mathbb{E}_P\left[(f_t-f^*)^2\right]-2\mathbb{E}_{P_n}\left[(f_t-f^*)^2\right]\lesssim r^*+\frac{x}{n}, \tag{125}$$

with $K_0=2$ in (124).

It follows from (123) and (125) that

$$\begin{aligned}&\mathbb{E}_P\left[(f_t-f^*)^2\right]-2\mathbb{E}_{P_n}\left[(f_t-f^*)^2\right]\\&\lesssim \min_{0\le Q\le n}\left(\frac{B_0Q}{n}+w\left(\sqrt{\frac{Q}{n}}+1\right)+B_h\left(\frac{\sum\limits_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right)+\frac{x}{n}.\end{aligned}$$

Let $x=n\varepsilon_n^2$ in the above inequality, and we note that the above argument requires Theorem C.8 which holds with probability at least $1-\exp\left(-\Theta(n)\right)-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$, then (117) is proved.

We now prove (118). First, it follows from the definition of $\varphi_{B_h,w}$ in (104) that

$$\begin{aligned}\psi(r)&=4B_0^3\varphi_{B_h,w}\left(\frac{r}{4B_0^2}\right)\\&=4B_0^3\min_{Q\colon Q\ge 0}\left(\left(\frac{\sqrt{r}}{2B_0}+w\right)\sqrt{\frac{Q}{n}}+B_h\left(\frac{\sum\limits_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right)+4B_0^3w\\&\le 4B_0^3B_h\min_{Q\colon Q\ge 0}\left(\sqrt{\frac{Qr}{n}}+\left(\frac{\sum\limits_{q=Q+1}^{\infty}\lambda_q}{n}\right)^{1/2}\right)+4B_0^3w\left(\sqrt{\frac{Q}{n}}+1\right)\\&\le\frac{4\sqrt{2}B_0^3B_h}{\sigma_0}\cdot\sigma_0R_K(\sqrt{r})+8B_0^3w:=\psi_1(r),\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. It can be verified that $\psi_1(r)$ is a sub-root function. Let the fixed point of $\psi_1(r)$ be r_1^*. Because the fixed point of $\sigma_0R_K(\sqrt{r})$ as a function of r is ε_n^2, it follows from the properties about the fixed points of sub-root functions in (Yang & Li, 2025, Lemma B.9) that

$$r_1^*\le\max\left\{\frac{32B_0^6B_h^2}{\sigma_0^2},1\right\}\varepsilon_n^2+16B_0^3w. \tag{126}$$

It then follows from Theorem A.3 with $K_0=2$ that with probability at least $1-\exp(-x)$,

$$\mathbb{E}_P\left[(f_t-f^*)^2\right]-2\mathbb{E}_{P_n}\left[(f_t-f^*)^2\right]\lesssim r_1^*+\frac{x}{n}.$$

Letting $x=n\varepsilon_n^2$, then plugging the upper bound for r_1^*, (126), in the above inequality leads to

$$\mathbb{E}_P\left[(f_t-f^*)^2\right]-2\mathbb{E}_{P_n}\left[(f_t-f^*)^2\right]\lesssim\varepsilon_n^2+16B_0^3w, \tag{127}$$

which proves (118). Again, we note that the above argument requires Theorem C.8 which holds with probability at least $1-\exp\left(-\Theta(n)\right)-\exp\left(-\Theta(n\widehat{\varepsilon}_n^2)\right)$ over the random noise $\mathbf{w}$. □

Theorem C.11. Suppose the neural network trained after the t-th step of gradient descent, $f_t=f(\mathbf{W}(t),\cdot)$, 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 $T\le\widehat{T}$. If

$$\eta\in[1,2),\quad\tau\le\frac{1}{\eta T}, \tag{128}$$

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

$$\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] \le \frac{3}{\eta t}\left(\frac{\mu_0^2}{2e} + 3\right). \tag{129}$$

Proof. We have

$$f_t(\mathbf{S}) = f^*(\mathbf{S}) + \mathbf{w} + \mathbf{v}(t) + \mathbf{e}(t), \tag{130}$$

where $\mathbf{v}(t) \in \mathcal{V}_t$, $\mathbf{e}(t) \in \mathcal{E}_{t,\tau}$, $\overrightarrow{\mathbf{e}}(t) = \overrightarrow{\mathbf{e}}_1(t) + \overrightarrow{\mathbf{e}}_2(t)$ with $\overrightarrow{\mathbf{e}}_1(t) = -\left(\mathbf{I}_n - \eta\mathbf{K}_n\right)^t \mathbf{w}$ and $\left\|\overrightarrow{\mathbf{e}}_2(t)\right\|_2 \le \sqrt{n}\tau$. We have $\eta\lambda_1 \in (0,1)$ if $\eta \in [1,2)$. It follows from (130) that

$$\begin{aligned}
&\mathbb{E}_{P_n}\left[(f_t - f^*)^2\right] = \frac{1}{n}\|f_t(\mathbf{S}) - f^*(\mathbf{S})\|_2^2 = \frac{1}{n}\|\mathbf{v}(t) + \mathbf{w} + \mathbf{e}(t)\|_2^2 \\
&= \frac{1}{n}\left\|-\left(\mathbf{I} - \eta\mathbf{K}_n\right)^t f^*(\mathbf{S}) + \left(\mathbf{I}_n - \left(\mathbf{I}_n - \eta\mathbf{K}_n\right)^t\right)\mathbf{w} + \overrightarrow{\mathbf{e}}_2(t)\right\|_2^2 \\
&\overset{①}{\le} \frac{3}{n}\sum_{i=1}^n \left(1 - \eta\widehat{\lambda}_i\right)^{2t}\left[\mathbf{U}^\top f^*(\mathbf{S})\right]_i^2 + \frac{3}{n}\sum_{i=1}^n\left(1 - \left(1 - \eta\widehat{\lambda}_i\right)^t\right)^2\left[\mathbf{U}^\top\mathbf{w}\right]_i^2 + \frac{3}{n}\left\|\overrightarrow{\mathbf{e}}_2(t)\right\|_2^2 \\
&\overset{②}{\le} \frac{3\mu_0^2}{2e\eta t} + \frac{3}{n}\sum_{i=1}^n\left(1 - (1 - \eta\lambda_i)^t\right)^2\left[\mathbf{U}^\top\mathbf{w}\right]_i^2 + 3\tau^2 \\
&\le \frac{3}{\eta t}\left(\frac{\mu_0^2}{2e} + \frac{1}{\eta}\right) + 3\cdot\underbrace{\frac{1}{n}\sum_{i=1}^n\left(1 - (1 - \eta\lambda_i)^t\right)^2\left[\mathbf{U}^\top\mathbf{w}\right]_i^2}_{:=E_\varepsilon} \\
&\le \frac{3}{\eta t}\left(\frac{\mu_0^2}{2e} + 1\right) + 3E_\varepsilon.
\end{aligned} \tag{131}$$

Here ① follows from the Cauchy-Schwarz inequality, ② follows from (51) in the proof of Lemma C.4. We then derive the upper bound for E_ε on the RHS of (131). We define the diagonal matrix $\mathbf{R} \in \mathbb{R}^{n\times n}$ with $\mathbf{R}_{ii} = \left(1 - (1 - \eta\lambda_i)^t\right)^2$. Then we have

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

It follows from (Wright, 1973) that

$$\Pr\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right) - \mathbb{E}\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)\right] \ge u\right] \le \exp\left(-c\min\left\{nu/\|\mathbf{R}\|_2, n^2u^2/\|\mathbf{R}\|_{\mathrm{F}}^2\right\}\right). \tag{132}$$

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

$$\begin{aligned}
\mathbb{E}\left[1/n\cdot\mathrm{tr}\left(\mathbf{U}\mathbf{R}\mathbf{U}^\top\mathbf{w}\mathbf{w}^\top\right)\right] &= \frac{\sigma_0^2}{n}\sum_{i=1}^n\left(1 - \left(1 - \eta\widehat{\lambda}_i\right)^t\right)^2 \overset{①}{\le} \frac{\sigma_0^2}{n}\sum_{i=1}^n \min\left\{1, \eta_t^2\widehat{\lambda}_i^2\right\} \le \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^n\min\left\{\frac{1}{\eta_t}, \eta_t\widehat{\lambda}_i^2\right\} \\
&\overset{②}{\le} \frac{\sigma_0^2\eta_t}{n}\sum_{i=1}^n\min\left\{\frac{1}{\eta_t}, \widehat{\lambda}_i\right\} = \sigma_0^2\eta_t\widehat{R}_K^2(\sqrt{1/\eta_t}) \le \frac{1}{\eta_t}.
\end{aligned} \tag{133}$$

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

$$\|\mathbf{R}\|_2 \le \max_{i\in[n]}\left(1 - \left(1 - \eta\widehat{\lambda}_i\right)^t\right)^2 \le \min\left\{1, \eta_t^2\widehat{\lambda}_i^2\right\} \le 1. \tag{134}$$

We also have

$$\frac{1}{n}\|\mathbf{R}\|_{\mathrm{F}}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(1-\left(1-\eta\widehat{\lambda}_i\right)^t\right)^4 \le \frac{\eta_t}{n}\sum_{i=1}^{n}\min\left\{\frac{1}{\eta_t}, \eta_t^3\widehat{\lambda}_i^4\right\} \overset{③}{\le} \frac{\eta_t}{n}\sum_{i=1}^{n}\min\left\{\widehat{\lambda}_i, \frac{1}{\eta_t}\right\} = \eta_t\widehat{R}_K^2(\sqrt{1/\eta_t}) \le \frac{1}{\sigma_0^2\eta_t}. \tag{135}$$

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

Combining (132)-(135), we have

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

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

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

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

$$E_\varepsilon \le u + \frac{1}{\eta_t} = \frac{2}{\eta_t}. \tag{136}$$

It then follows from (131)-(136) that

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

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

□

C.3. Auxiliary Results about Reproducing Kernel Hilbert Spaces

Lemma C.12 (In the proof of (Raskutti et al., 2014, Lemma 8))**.** For any $f \in \mathcal{H}_K(\mu_0)$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\left[\mathbf{U}^\top f(\mathbf{S}')\right]_i^2}{\widehat{\lambda}_i} \le \mu_0^2. \tag{137}$$

Similarly, for $f \in \mathcal{H}_{K^{(\mathrm{int})}}(\mu_0)$, we have $\frac{1}{n}\sum_{i=1}^{n}\frac{\left[\mathbf{U}^\top f(\mathbf{S}')\right]_i^2}{\lambda_i} \le \mu_0^2$.

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

$$(1-a)^t \le e^{-ta} \le \frac{1}{eta}. \tag{138}$$

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 C.14. (Yang & Li, 2025, Lemma B.7)] With probability at least $1-2\exp(-\Theta(n\varepsilon_n^2))$,

$$\varepsilon_n^2 \le c_1\widehat{\varepsilon}_n^2. \tag{139}$$

Furthermore, with probability at least $1-2\exp(-\Theta(n\varepsilon_n^2))$,

$$\widehat{\varepsilon}_n^2 \le c_1\varepsilon_n^2. \tag{140}$$

Here c_1 is an absolute positive constant depending on σ_0.

Remark. Lemma C.14 shows that with probability at least $1-4\exp(-\Theta(n\varepsilon_n^2))$, $\varepsilon_n^2 \asymp \widehat{\varepsilon}_n^2$, which is also a fact used in kernel complexity or local Rademacher based analysis for kernel regression in the statistical learning literature.

D. Simulation Study

We present simulation results for GD in this section. We randomly sample n points $\left\{\overrightarrow{\mathbf{x}}_i\right\}_{i=1}^n$ as a i.i.d. sample of random variables distributed uniformly on the unit sphere in $\mathbb{R}^{50}$. n ranges within $[100, 1000]$ with a step size of 100. We set the target function to $f^*(\mathbf{x}) = \mathbf{s}^\top \mathbf{x}$ where $\mathbf{s} \sim \mathrm{Unif}\left(\mathcal{X}\right)$ is randomly sampled. We also uniformly and independently sample 1000 points on the unit sphere in $\mathbb{R}^{50}$ as the test data. We train the two-layer NN (1) using either GD by Algoirthm 1 or GD by Algoirthm 1 with $m \asymp n^2$ on a NVIDIA A100 GPU card with a learning rate $\eta = 0.1$, and report the test loss in Figure 2. It can be observed that early-stopping is always helpful in training neural networks with better generalization, as the test loss initially decreases and then increases with over-training. Figure 2 illustrates the test loss with respect to the steps (or epochs) of GD for $n = 100, 500, 1000$. For each n in $[100, 1000]$ with a step size of 100, we find the step of GD where minimum test loss is achieved, denoted by $\widehat{t}_n$, which is the empirical early stopping time. We note that the theoretically predicted early stopping time is $1/\widehat{\varepsilon}_n^2 \asymp n^{d/(2d-1)}$, and we compute the ratio of early stopping time for each n by $\widehat{t}_n / n^{d/(2d-1)}$. Such ratios for different values of n are illustrated in the bottom right figure of Figure 2. It is observed that the ratio of early stopping time is roughly stable and distributed between $[8, 10]$, suggesting that predicted early stopping time is empirically proportional to the empirical early stopping time.

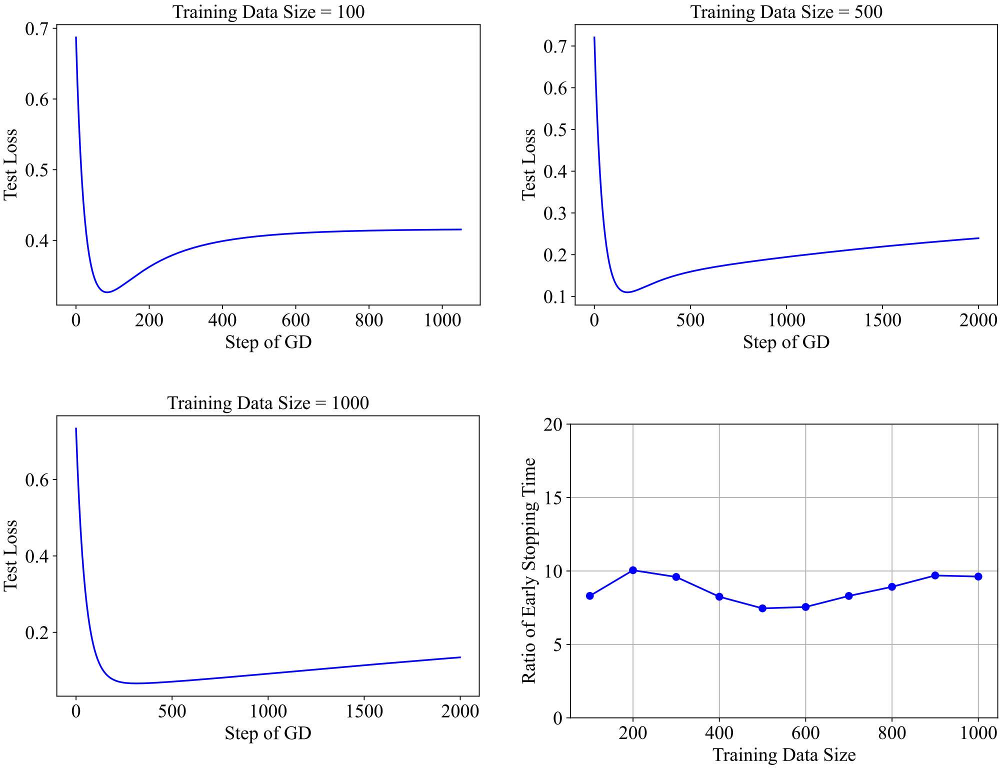

Figure 2: Illustration of the test loss by GD