Regularization Matters: A Nonparametric Perspective on Overparametrized Neural Network

Wenjia Wang [*]
SAMSI[†]
wenjia.wang234@duke.edu

Tianyang Hu [*]
Purdue University
hu478@purdue.edu

Cong Lin
East China Normal University
conglin1104@gmail.com

Guang Cheng
Purdue University
chengg@purdue.edu

Abstract

Overparametrized neural networks trained by gradient descent (GD) can provably overfit any training data. However, the generalization guarantee may not hold for noisy data. From a nonparametric perspective, this paper studies how well overparametrized neural networks can recover the true target function in the presence of random noises. We establish a lower bound on the L_2 estimation error with respect to the GD iteration, which is away from zero without a delicate choice of early stopping. In turn, through a comprehensive analysis of ℓ_2-regularized GD trajectories, we prove that for overparametrized one-hidden-layer ReLU neural network with the ℓ_2 regularization: (1) the output is close to that of the kernel ridge regression with the corresponding neural tangent kernel; (2) minimax optimal rate of L_2 estimation error is achieved. Numerical experiments confirm our theory and further demonstrate that the ℓ_2 regularization approach improves the training robustness and works for a wider range of neural networks.

1 Introduction

Deep learning has shown outstanding empirical successes and demonstrates superior performance in many standard machine learning tasks, such as image classification [1, 2, 3], generative modeling [4, 5], etc. Despite common accusations of being a black box with no theoretical guarantee, deep neural network (DNN) tends to achieve higher accuracy than other classical methods in various prediction tasks, which attracts plenty of interests from researchers. In contrast to the huge empirical success, little is yet settled from the theoretical side why DNN outperforms other methods. Without enough understanding, practical use of deep learning models could be inefficient or unreliable.

Recently, many efforts have been devoted to provable deep learning methods with algorithmic guarantees, particularly training overparametrized neural networks by gradient descent (GD) or other gradient-based optimization. It has been shown that with enough overparametrization, e.g., neural network width tends to infinity, training DNN resembles a kernel method with a specific kernel called as "neural tangent kernel" (NTK) [6]. In the NTK regime, GD can provably minimize the training error to zero in both regression [7, 8, 9, 10] and classification [11, 12, 13] settings. The corresponding generalization error bounds are developed to ensure prediction performance on unseen data. However, a closer inspection of these generalization results reveals that they only hold under the noiseless assumption, i.e., the response variable is deterministic given the explanatory

[*] These authors contributed equally to this manuscript.

[†] The Statistical and Applied Mathematical Sciences Institute

Preprint. Under review.

variables. For overparametrized neural networks, the training loss can be minimized to zero so that the generalization error equals the population loss, which cannot be zero in the presence of noises. As random noises are ubiquitous in the real world, theoretical guarantees and provable learning algorithms that take into account of random noises are much needed in practice.

In contrast, classic nonparametric statistics literature demonstrate that in the presence of noises, the L_2 estimation error can still go to zero with possibly optimal rates as established in [14]. To further investigate how overparametrized neural networks trained via GD work and how well they can learn the underlying true function with noisy data, we consider the classic nonparametric regression setting. Assume data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are generated from

$$y_i = f^*(\boldsymbol{x}_i) + \epsilon_i, \tag{1.1}$$

where f^* is the ground truth, $\boldsymbol{x}_i \in \mathbb{R}^d$, and ϵ_i's are i.i.d. noises with mean 0 and finite variance σ^2. The goal is to construct a neural network estimator $\widehat{f}$ from data to estimate f^* and investigate how fast the L_2 estimation error $[\mathbb{E}(\widehat{f} - f^*)^2]^{1/2}$ converges to zero as sample size grows. Note that the L_2 convergence rate critically depends on the assumptions of the true function, such as smoothness, based on which minimax lower bounds are established [15]. An estimation method is said to be *minimax-optimal* if it achieves the lower bound, indicating that it performs the best in the worst possible scenario. The above nonparametric perspective provides a sharp characterization of the employed estimation method and complements the existing optimization/generalization framework.

The main contributions of this paper are twofold:

- We prove that overparametrized one-hidden-layer ReLU neural networks trained using GD do not recover the true function in the classic nonparametric regression setting (1.1), i.e., the L_2 estimation error is bounded away from zero as sample size diverges. To predict well on unseen data, a delicate early stopping rule has to be deployed.
- We analyze the ℓ_2-regularized GD trajectory and show that the ℓ_2 penalty on the network weights amounts to penalizing the RKHS (induced by NTK) norm of the associated neural network. We further prove that by adding the ℓ_2 regularization, neural network with sufficient overparametrization achieves the *minimax-optimal* L_2 convergence rate.

In general, this work connects the recent advances in deep learning theory, e.g., analyzing the trajectory of GD updates, implicit bias of overparametrization, etc., to the classical nonparametric statistics literature. More specifically, our findings not only contribute to the theoretical (in particular, nonparametric) understanding of training overparametrized DNN on noisy data but also promotes the use of ℓ_2 penalty or weight decay in practice for better theoretical guarantee.

2 Related works

Neural tangent kernel The seminal paper [6] proves that the evolution of DNNs during training can be described by the so-called neural tangent kernel (NTK), which is central to characterize the convergence and generalization behaviors. [7, 9, 8] investigate specifically for one-hidden-layer ReLU neural network and show explicitly that with enough overparametrization, the weight vectors and the corresponding NTK do not change much during GD training. Similar investigations have been done for other neural networks and other settings [10, 12]. Among others, [9, 16] provide generalization error bounds and provable learning scenarios, but only hold for *noiseless* data. For noisy data, explicit regularizations have recently been considered in the NTK literature. [17] promote the ℓ_2 penalty in the NTK setting by showing that in a constructed classification example, sample efficiency can benefit from regularization. [18] consider classification with noisy labels and propose to add ℓ_2 regularization to ensure robustness. However, their analyses only apply to the kernel estimator directly using NTK, but not overparametrized neural networks, which greatly restricts the model class capacity. In comparison, we directly analyze GD trajectories of training neural networks and prove that the NTK solutions can be well-approximated after a polynomial number of GD iterations. To the best of our knowledge, there are no existing regression results that establish L_2 convergence rate for trained neural networks under noisy data.

Nonparametric regression In nonparametric statistics, [14] show that for the L_2 estimation error, the optimal rate of convergence is $n^{-\beta/(2\beta+d)}$ when f^* is d-variate and β-times differentiable. Many popular methods such as kernel methods, Gaussian process, splines, etc. achieve this rate. It has been recently shown that DNN (with certain compositional structure) can also achieve optimal convergence rates [19, 20, 21, 22] and even for non-smooth functions [23]. However, this type of results only applies to the empirical risk minimizer or some specially constructed DNNs without any algorithmic guarantee. In this sense, the aforementioned results are less helpful in understanding deep neural network models whose optimization is nontrivial, say highly non-convex.

Our algorithm-dependent statistical analysis bridges the gap between these two types of research. Based on the GD trajectories and the corresponding NTK, we are able to analyze the trained overparametrized neural networks within the nonparametric framework and show they can also achieve the optimal convergence rate with proper regularizations.

3 Preliminaries

Notation For any function $f(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$, denote $\|f\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x})|$ and $\|f\|_p = (\int_{\mathcal{X}} |f(\boldsymbol{x})|^p d\boldsymbol{x})^{1/p}$. For any vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|_p$ denotes its p-norm, for $1 \le p \le \infty$. For two given sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ of real numbers, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \le Cb_n$ for all sufficiently large n. Let $\Omega(\cdot)$ be the counterpart of $O(\cdot)$ that $a_n = \Omega(b_n)$ means $a_n \gtrsim b_n$. Further, $a_n = \widetilde{O}(b_n)$ and $a_n = \widetilde{\Omega}(b_n)$ are used to indicate there are specific requirements for the multiplicative constants. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. Let $[N] = \{1, \ldots, N\}$ for $N \in \mathbb{N}$ and let $\lambda_{\min}(\boldsymbol{A})$ be the minimum eigenvalue of a symmetric matrix $\boldsymbol{A}$. We use $\mathbb{I}$ to denote the indicator function and $\boldsymbol{I}_d$ to denote the $d \times d$ identity matrix. $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ and $\mathrm{poly}(t_1, t_2, \ldots)$ denotes some polynomial function with arguments $t_1, t_2, \ldots$.

Neural network setup Consider the one-hidden-layer ReLU neural network family $\mathcal{F}$ with m nodes in the hidden layer, expressed as

$$f_{\boldsymbol{W},\boldsymbol{a}}(\boldsymbol{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(\boldsymbol{w}_r^\top \boldsymbol{x}),$$

where $\boldsymbol{x} \in \mathbb{R}^d$ denotes the input, $\boldsymbol{W} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m) \in \mathbb{R}^{d\times m}$ is the weight matrix in the hidden layer, $\boldsymbol{a} = (a_1, \cdots, a_m)^\top \in \mathbb{R}^m$ is the weight vector in the output layer, $\sigma(z) = \max\{0, z\}$ is the rectified linear unit (ReLU). The initial values of the weights are independently generated from

$$\boldsymbol{w}_r(0) \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I}_m), \;\; a_r \sim \mathrm{unif}\{-1, 1\}, \;\; \forall r \in [m].$$

When $m \gg n$, the neural network is overparametrized. As is usually assumed in the NTK literature [9, 18, 24], we consider data on the unit sphere $\mathbb{S}^{d-1}$, i.e., $\|\boldsymbol{x}_i\|_2 = 1$ for any $i \in [n]$. Throughout this work, we further assume that $\boldsymbol{x}_i$'s are *uniformly* distributed on $\mathbb{S}^{d-1}$ so that $\mathbb{E}(f - f^*)^2$ and $\|f - f^*\|_2^2$ are equal up to a constant multiplier and thus will be used interchangeably.

Gradient descent Let $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$ and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^\top$. Denote $u_i = f_{\boldsymbol{W},\boldsymbol{a}}(\boldsymbol{x}_i)$ to be the network's prediction on $\boldsymbol{x}_i$ and let $\boldsymbol{u} = (u_1, ..., u_n)^\top$. Without loss of generality, we consider fixing the second layer $\boldsymbol{a}$ after initialization and only training the first layer $\boldsymbol{W}$ by GD. Fixing the last layer is not a strong restriction since $a \cdot \sigma(z) = \mathrm{sign}(a) \cdot \sigma(|a|z)$ and we can always reparametrize the network to have all a_i's to be either 1 or -1. Denote the empirical squared loss as $\Phi(\boldsymbol{W}) = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{u}\|_2^2$. The gradient of Φ w.r.t. $\boldsymbol{w}_r$ is

$$\frac{\partial \Phi(\boldsymbol{W})}{\partial \boldsymbol{w}_r} = \frac{1}{\sqrt{m}} a_r \sum_{i=1}^{n} (u_i - y_i) \mathbb{I}_{r,i} \boldsymbol{x}_i, \quad r \in [m],$$

where $\mathbb{I}_{r,i} = \mathbb{I}\{\boldsymbol{w}_r^\top \boldsymbol{x}_i \ge 0\}$. Then the GD update rule at the k-th iteration is given by

$$\boldsymbol{w}_r(k+1) = \boldsymbol{w}_r(k) - \eta \frac{\partial \Phi(\boldsymbol{W})}{\partial \boldsymbol{w}_r}\bigg|_{\boldsymbol{W}=\boldsymbol{W}(k)},$$

where $\eta > 0$ is the step size (a.k.a. learning rate). In the rest of this work, we use k to index variables at the k-th iteration, e.g., $u_i(k) = f_{\boldsymbol{W}(k),\boldsymbol{a}}(\boldsymbol{x}_i)$, etc. Define $\mathbb{I}_{r,i}(k) = \mathbb{I}\{\boldsymbol{w}_r(k)^\top \boldsymbol{x}_i \geq 0\}$,

$$\boldsymbol{Z}(k) = \frac{1}{\sqrt{m}} \begin{pmatrix} a_1 \mathbb{I}_{1,1}(k)\boldsymbol{x}_1 & \dots & a_1 \mathbb{I}_{1,n}(k)\boldsymbol{x}_n \\ \vdots & \ddots & \vdots \\ a_m \mathbb{I}_{m,1}(k)\boldsymbol{x}_1 & \dots & a_m \mathbb{I}_{m,n}(k)\boldsymbol{x}_n \end{pmatrix} \in \mathbb{R}^{md\times n}$$

and $\boldsymbol{H}(k) = \boldsymbol{Z}(k)^\top \boldsymbol{Z}(k)$. It is shown that matrices $\boldsymbol{Z}(k)$ and $\boldsymbol{H}(k)$ are close to $\boldsymbol{Z}(0)$ and $\boldsymbol{H}(0)$, respectively for any k, when m is sufficiently large [9]. We can rewrite the GD update rule as

$$\operatorname{vec}(\boldsymbol{W}(k+1)) = \operatorname{vec}(\boldsymbol{W}(k)) - \eta \boldsymbol{Z}(k)(\boldsymbol{u}(k) - \boldsymbol{y}), \tag{3.1}$$

where $\operatorname{vec}(\boldsymbol{W}) = (\boldsymbol{w}_1^\top, \cdots, \boldsymbol{w}_m^\top)^\top \in \mathbb{R}^{md\times 1}$ is the vectorized weight matrix.

Kernel ridge regression with NTK The study of one-hidden-layer ReLU neural networks is closely related to the NTK defined as

$$h(\boldsymbol{s},\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{w}\sim N(0,\boldsymbol{I}_d)} \left(\boldsymbol{s}^\top \boldsymbol{t}\, \mathbb{I}\{\boldsymbol{w}^\top \boldsymbol{s} \geq 0, \boldsymbol{w}^\top \boldsymbol{t} \geq 0\}\right) = \frac{\boldsymbol{s}^\top \boldsymbol{t}(\pi - \arccos(\boldsymbol{s}^\top \boldsymbol{t}))}{2\pi}, \tag{3.2}$$

where $\boldsymbol{s}, \boldsymbol{t}$ are d-dimensional vectors. It can be shown that h is positive definite on the unit sphere $\mathbb{S}^{d-1}$ [24]. Let the Mercer decomposition of h be $h(\boldsymbol{s},\boldsymbol{t}) = \sum_{j=0}^\infty \lambda_j \varphi_j(\boldsymbol{s})\varphi_j(\boldsymbol{t})$, where $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ are the eigenvalues, and $\{\varphi_j\}_{j=1}^\infty$ is an orthonormal basis.

The following lemma states the decay rate of eigenvalues of the NTK associated with one-hidden-layer ReLU neural networks, as a key technical contribution of this work.

Lemma 3.1. Let λ_j be the eigenvalues of NTK h defined above. Then we have $\lambda_j \asymp j^{-\frac{d}{d-1}}$.

Let $\mathcal{N}$ denote the reproducing kernel Hilbert space (RKHS) generated by h on $\mathbb{S}^{d-1}$, equipped with norm $\|\cdot\|_{\mathcal{N}}$. For an unknown function $f^* \in \mathcal{N}$, the kernel ridge regression minimizes

$$\min_{f\in\mathcal{N}} \frac{1}{n}\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2 + \frac{\mu}{n}\|f\|_{\mathcal{N}}^2, \tag{3.3}$$

where $\mu > 0$ is a tuning parameter controlling the regularization strength. The representer theorem says that the solution to (3.3) can be written as

$$\widehat{f}(\boldsymbol{x}) = h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \mu \boldsymbol{I}_n)^{-1}\boldsymbol{y} \tag{3.4}$$

for any point $\boldsymbol{x}$, where $h(\boldsymbol{x},\boldsymbol{X}) = (h(\boldsymbol{x},\boldsymbol{x}_1), ..., h(\boldsymbol{x},\boldsymbol{x}_n)) \in \mathbb{R}^{1\times n}$ and $\boldsymbol{H}^\infty = (h(\boldsymbol{x}_i,\boldsymbol{x}_j))_{n\times n}$ ($\boldsymbol{H}^\infty$ is usually called the NTK matrix). In the following theorem, we show that the function $\widehat{f}$ is close to the true function f^* under L_2 metric.

Theorem 3.2. Let $\widehat{f}$ be as in (3.4). By choosing $\mu \asymp n^{(d-1)/(2d-1)}$, we have

$$\left\|\widehat{f} - f^*\right\|_2^2 = O_{\mathbb{P}}\left(n^{-\frac{d}{2d-1}}\right), \quad \left\|\widehat{f}\right\|_{\mathcal{N}}^2 = O_{\mathbb{P}}(1).$$

The proof of the convergence rate requires an accurate characterization of the complexity of $\mathcal{N}$, which is determined by the eigenvalues and eigenfunction expansion of the NTK h. If the eigenvalues decay at rate $\lambda_j \asymp j^{-2\nu}$, the corresponding minimax optimal rate is $n^{-2\nu/(2\nu+1)}$ [25, 26]. Building on the the eigenvalue decay rate established in Lemma 3.1, it can be shown that the L_2 estimation rate in Theorem 3.2 is minimax-optimal.

In the rest of this work, we assume that $f^* \in \mathcal{N}$.

4 Problems of gradient descent from the nonparametric perspective

In this section, we consider training overparametrized neural networks with the GD update rule (3.1). Among others, [9, 7] prove that as iteration $k \to \infty$, the training data are interpolated, achieving zero training loss. However, in the presence of noises, i.e., ϵ_i in (1.1), such an overfitting to the training data can be harmful for recovering the true function.

The following theorem shows that if k is too small or too large, the L_2 estimation error of the trained neural network is bounded away from zero.

Theorem 4.1. Fix a failure probability $\delta \in (0,1)$. Let λ_0 be the largest number that with probability at least $1-\delta$, $\lambda_{\min}(\boldsymbol{H}^{\infty}) \geq \lambda_0$. Suppose $m \geq \tau^{-2}\mathrm{poly}\left(n, \frac{1}{\lambda_0}, \frac{1}{\delta}\right)$, $\eta = \widetilde{O}\left(\frac{\lambda_0}{n^2}\right)$, and $\tau = \widetilde{O}\left(\frac{\lambda_0 \delta}{n}\right)$. For sufficiently large n, if the iteration $k = \widetilde{\Omega}\left(\frac{\log n}{\eta \lambda_0}\right)$ or $k = \widetilde{O}\left(\frac{1}{n\eta}\right)$, then with probability at least $1-2\delta$, we have

$$\mathbb{E}_{\boldsymbol{\epsilon}} \left\| f_{\boldsymbol{W}(k),\boldsymbol{a}} - f^* \right\|_2^2 = \Omega(1).$$

The conditions on m, η, and τ are similar to those in Theorem 5.1 of [9]. The probability $1-2\delta$ in Theorem 4.1 comes from the randomness of $\lambda_{\min}(\boldsymbol{H}^{\infty})$ and $(\boldsymbol{W}(0), \boldsymbol{a})$.

Theorem 4.1 states that the estimation error for non-regularized one-hidden-layer neural networks is bounded away from zero by some constant if trained for too short or too long. The latter scenario indicates overfitting is harmful in terms of L_2 estimation error. Similar results have been shown in [27] for specifically designed overparametrized DNNs that is a linear combination of $\Omega(n^{10d^2})$ smaller neural networks, which is very restrictive.

To have low L_2 estimation errors, Theorem 4.1 requires $(\eta\lambda_0)^{-1}\log n \lesssim k \lesssim (n\eta)^{-1}$. However, deriving a precise order of k, which leads to the best rate of convergence, could be extremely challenging. Alternatively, we consider the infinite-width limit of one-hidden-layer ReLU networks, i.e., the NTK (3.2) in kernel regression. This may shed some light on the optimal stopping time for practical overparametrized neural networks.

In kernel regression, the objective becomes

$$\min_{f \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2, \tag{4.1}$$

whose solution can be explicitly expressed as $h(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}$, by setting $\mu = 0$ in (3.4). However, inverting the kernel matrix can be computationally intensive. In practice, gradient-based methods are often applied to solve (4.1) [26]. The following theorem establishes estimation error results for the NTK estimators trained by GD, complementary to Theorem 4.1.

Theorem 4.2. Consider using GD to optimize (4.1) with a sufficiently small step size η depending on n (but not on k). There exists a stopping time k^* depending on data, such that

$$\mathbb{E} \left\| \widehat{f}_{k^*} - f^* \right\|_2^2 = O\left(n^{-\frac{d}{2d-1}}\right),$$

where $\widehat{f}_k$ is the predictor obtained at the k-th iteration. Moreover, if $k \to \infty$, the interpolated estimator $\widehat{f}_{\infty}$ satisfies

$$\mathbb{E} \left\| \widehat{f}_{\infty} - f^* \right\|_2^2 = \Omega(1).$$

To specify the optimal stopping time k^* in Theorem 4.2, we first introduce the local empirical Rademacher complexity defined as $\widehat{\mathcal{R}}_{\boldsymbol{H}^{\infty}}(\varepsilon) := \left(\frac{1}{n}\sum_{i=1}^{n} \min\left\{\widehat{\lambda}_i/n, \varepsilon^2\right\}\right)^{1/2}$, which relies on the eigenvalues $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n > 0$ of $\boldsymbol{H}^{\infty}$. Then, the stopping time k^* is defined to be

$$k^* := \mathrm{argmin}\left\{ k \in \mathbb{N} \mid \widehat{\mathcal{R}}_{\boldsymbol{H}^{\infty}}\left(1/\sqrt{\eta k}\right) > (2e\sigma\eta k)^{-1} \right\} - 1. \tag{4.2}$$

In essence, the optimal stopping time decreases with noise level σ and increases with the model complexity, measured by the eigenvalues of $\boldsymbol{H}^{\infty}$.

To derive the order of k^* for NTK, a sharp characterization of the eigen-distribution of $\boldsymbol{H}^{\infty}$ is needed. To the best of our knowledge, no such results are available yet. Even though as $m \to \infty$, neural network resembles its linearization (NTK), it doesn't necessarily mean such a stopping rule can be easily derived for finite-width neural networks. In general, theoretical guarantees of an early stopping rule for training overparametrized neural networks is challenging and left for future work.

Instead, explicit regularizations are usually employed in deep learning models, for example, weight decay [28], batch normalization [29], dropout [30], etc., to prevent overfitting. In the next section, we consider the ℓ_2 regularization [31, 32, 33] and demonstrate its effectiveness in the nonparametric regression setting.

5 ℓ_2-regularized gradient descent for noisy data

Without any regularization, GD overfits the training data and the estimation error is bounded away from zero. Instead, we propose using the ℓ_2-regularized gradient descent defined as

$$\mathrm{vec}(\boldsymbol{W}_D(k+1)) = \mathrm{vec}(\boldsymbol{W}_D(k)) - \eta_1 \boldsymbol{Z}_D(k)(\boldsymbol{u}_D(k) - \boldsymbol{y}) - \eta_2 \mu \mathrm{vec}(\boldsymbol{W}_D(k)), \tag{5.1}$$

where $\eta_1, \eta_2 > 0$ are step sizes, and $\mu > 0$ is a tuning parameter. It can be easily seen that (5.1) is the GD update rule on the following loss function

$$\Phi_1(\boldsymbol{W}) = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{u}\|_2^2 + \frac{\mu}{2} \|\mathrm{vec}(\boldsymbol{W})\|_2^2 . \tag{5.2}$$

ℓ_2 regularization has long been used in training neural networks and is equivalent to "weight decay" [28] when using GD [34]. The ℓ_2 regularization is also considered in theoretical analysis of training overparametrized neural networks as a way to improve generalization [17, 18]. In the rest of this work, we use subscript D to denote the variables under the regularized GD (5.1), e.g., $\boldsymbol{u}_D(k)$ for the predictions at the k-th iteration.

Theorem 5.1. Let λ_0 be the largest number such that with probability at least $1 - \delta_n$, $\lambda_{\min}(\boldsymbol{H}^\infty) \geq \lambda_0$, and $\delta_n \to 0$ as n goes to infinity[3]. For sufficiently large n, suppose $\mu \asymp n^{\frac{d-1}{2d-1}}$, $\eta_1 \asymp \eta_2 = o(n^{-\frac{3d-1}{2d-1}})$, $\tau = O(1)$, $m \geq \tau^{-2}\mathrm{ploy}(n, \lambda_0^{-1})$, and the iteration number k satisfies $\log\left(\mathrm{ploy}_1(n, \tau, 1/\lambda_0)\right) \lesssim \eta_2 \mu k \lesssim \log\left(\mathrm{ploy}_2(\tau, 1/n, \sqrt{m})\right)$. Then we have

$$\left\|\boldsymbol{u}_D(k) - \boldsymbol{H}^\infty (C\mu I + \boldsymbol{H}^\infty)^{-1} \boldsymbol{y}\right\|_2 = O_{\mathbb{P}}\left(\sqrt{n}(1 - \eta_2\mu)^k\right), \tag{5.3}$$

$$\left\|\mathrm{vec}(\boldsymbol{W}_D(k)) - (1 - \eta_2 \mu)^k \mathrm{vec}(\boldsymbol{W}_D(0))\right\|_2 = O_{\mathbb{P}}(1), \tag{5.4}$$

for some constant $C > 0$. Moreover, during the training process, the mean squared loss satisfies

$$\Phi(\boldsymbol{W}_D(k))/n \leq (1 - \eta_2\mu)^k \Phi(\boldsymbol{W}_D(0))/n + O_{\mathbb{P}}(1). \tag{5.5}$$

In the above theorem, three upper bounds are provided. In (5.3), we provide an upper bound on the difference between the prediction using one-hidden-layer neural networks and the prediction obtained by (3.4), which converges to zero as the sample size goes to infinity. This indicates that the ℓ_2 penalty on neural network weights has similar effects to penalizing the RKHS norm as in (3.3). Combining (5.3) and Theorem 3.2, we can conclude that the ℓ_2-regularized one-hidden-layer ReLU neural network recovers the true function on the training data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

In (5.4), we provide an upper bound on the distance between the weight matrix at the k-th iteration and the "decayed" initialization $\boldsymbol{W}_D(0)$. Under the conditions in Theorem 5.1, their distance measured in Frobenius norm is bounded by some constant depending on the underlying true function. Unlike the results in [9], the upper bound presented in (5.4) does not depend on data. Therefore, as long as the underlying function is within the RKHS generated by NTK, the total movement of all the weights is not large even if the data observed are corrupted by noises.

In (5.5), we give a characterization of how the training objective decreases over iterations, which is reminiscent of Theorem 4.1 in [7]. Unlike the results without regularization, our ℓ_2-regularized objective is not expected to converge to zero, i.e., no data interpolation, which is essential to ensure the best trade-off between bias and variance.

Remark 1. (More iterations) The required iteration number k in Theorem 5.1 is approximately $(\eta_2\mu)^{-1}$, up to a logarithmic term. We believe the upper bound on k is not necessary and may be relaxed. The stated results are expected to hold if $k \to \infty$ and we conjecture that the output will converge to the optimal solution of kernel ridge regression as in (3.4). Simulation results in Section 6 support our conjecture and we leave the technical proof for future work.

Next, we extend the results in Theorem 5.1 and establish the L_2 convergence rate for neural networks trained with ℓ_2-regularized GD.

Theorem 5.2. Suppose the assumptions of Theorem 5.1 hold. Then we have

$$\left\|f_{\boldsymbol{W}_D(k), \boldsymbol{a}} - f^*\right\|_2^2 = O_{\mathbb{P}}(n^{-\frac{d}{2d-1}}).$$

[3] Potential dependency of λ_0 on n is suppressed for notational simplicity.

The above theorem states that with probability tending to one, the neural network estimator can still recover the true function with the optimal convergence rate of $n^{-\frac{d}{2d-1}}$, demonstrating the effectiveness of the ℓ_2 regularization for noisy data. Unlike other optimality results established for neural networks [20, 21], our convergence rate result applies to overparametrized networks and is obtainable using the ℓ_2-regularized GD.

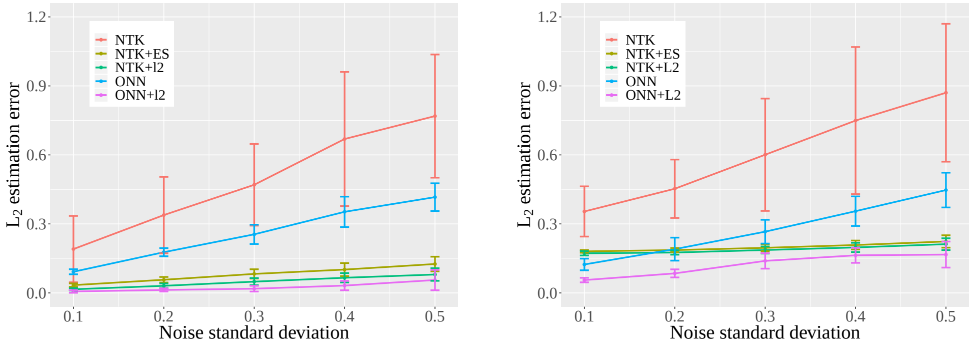
6 Numerical studies

In practice, regularization techniques are widely used in training deep learning models. Among others, [32, 35, 36, 37] have investigated the effectiveness of ℓ_2 regularization and early stopping in training DNNs, and comprehensive comparisons have been made empirically against other regularization techniques. Therefore, one major goal of this section is not to show state-of-the-art performance using ℓ_2 regularization, but to use it as an example to illustrate, from a nonparametric perspective, the necessity of regularization in training overparametrized neural networks with GD. Another goal is to demonstrate the robustness of our theory when some underlying assumptions are violated, e.g., one hidden layer, ReLU activation function and data on a sphere, etc.

Specifically, we consider NTK without regularization (NTK), NTK with early stopping[4] (NTK+ES), NTK with ℓ_2 regularization (NTK+ℓ_2), overparametrized neural network with and without ℓ_2 regularization, denoted as ONN and ONN+ℓ_2, respectively. For ONN, we use two-hidden-layer ReLU neural networks and $m = 500$ for each layer. To train the neural networks, instead of GD, we consider the more popular RMSProp optimizer [38] with the default setting. For ONN+ℓ_2 and NTK+ℓ_2, the tuning parameter μ is selected by cross-validation.

6.1 Simulated Data

Consider the $d = 2$ case where the training data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are i.i.d. sampled from Unif($[-1, 1]^2$). We set $n = 100$ and let noises follow $N(0, \sigma^2)$. Two target functions are considered: $f_1^*(\boldsymbol{x}) = 0$ and $f_2^*(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x}$. The L_2 estimation error is approximated using a noiseless test dataset $\{(\bar{\boldsymbol{x}}_i, f^*(\bar{\boldsymbol{x}}_i))\}_{i=1}^{1000}$ where $\bar{\boldsymbol{x}}_i$'s are new samples i.i.d. from Unif($[-1, 1]^2$). We choose $\sigma = 0.1, 0.2, ..., 0.5$ and for each σ value, 100 replications are run to estimate the mean and standard deviation of the L_2 estimation error. Results are presented in Figure 1. More details and results can be found in Appendix G.



(a) f_1^*

(b) f_2^*

Figure 1: The L_2 estimation errors are shown for all methods vs. σ, with their standard deviations plotted as vertical bars. Similarly for both f_1^* and f_2^*, we observe that NTK and ONN do not recover the true function well. Early stopping and ℓ_2 regularization perform similarly for NTK, especially for f_2^*. ONN+ℓ_2 performs the best in both cases.

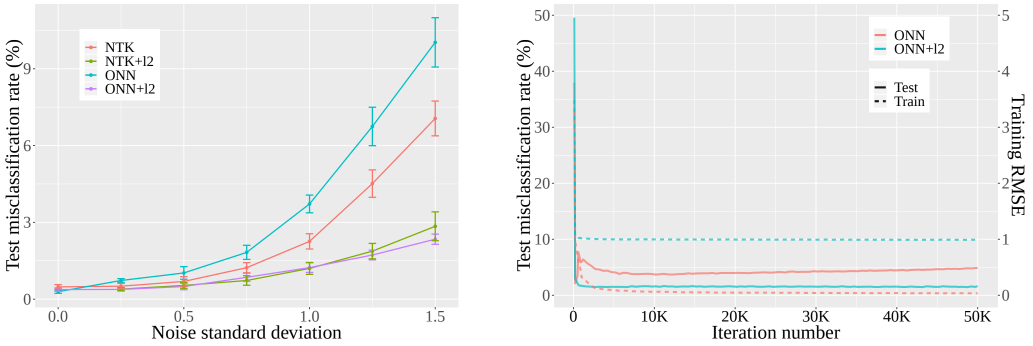
[4] As specified in Theorem 4.2, the optimal stopping time k^* in (4.2) depends on σ, which is to be estimated from data. In our simulation, we directly use the true value. The GD algorithm can found in Appendix G

6.2 Real Data

To showcase our results on the L_2 estimation, an ideal dataset is one that can be well-fitted by neural networks so that we can treat it as noiseless and then manually inject random noises. Inspired by the numerical studies in [18], we consider the MNIST dataset (digits 5 vs. 8 relabeled as -1 and 1), where the test accuracy can reach 99% by shallow fully connected neural networks [39]. Even though the dataset is for classification, we can treat the labels as continuous and learn the true function under the proposed regression setting. We use $\boldsymbol{y}^*$ to denote the true labels and manually add noises $\boldsymbol{\epsilon}$ to the training data, where each element of $\boldsymbol{\epsilon}$ follows $N(0, \sigma^2)$. The perturbed labels are denoted by $\boldsymbol{y} = \boldsymbol{y}^* + \boldsymbol{\epsilon}$. By gradually increase σ, we investigate how ONN and ONN+ℓ_2 perform under the additive label noises.

Remark 2. (Additive label noises) To manually inject noises to classification data, many works consider replacing part of the labels by random labels [37, 9]. However, such noises are not i.i.d. and cannot be applied to the regression setting. Similar additive label noises are also considered in [18].

The training dataset contains $n = 11272$ vectorized images of dimension $d = 784$. The test dataset size is 1866. For ONN+ℓ_2, our training objective function is Φ_1 as in (5.2) and setting $\mu = 0$ corresponds to the objective function of training ONN. On test dataset, which is *not contaminated* by noises, we use the sign of the output for classification and calculate the misclassification rate as a measure of estimation performance. To be more specific, a test image $\bar{\boldsymbol{x}}$ is classified as label 8 if $\widehat{f}(\bar{\boldsymbol{x}}) \geq 0$, and label 5 if $\widehat{f}(\bar{\boldsymbol{x}}) < 0$, where $\widehat{f}$ is the neural network estimator. The misclassification rate is the percentage of incorrect classifications on the test images. We choose $\sigma = 0, 0.25, ..., 1.5$ and for each σ value, 100 replications are run to estimate the mean and standard deviation of the test misclassification rate. How the training root mean square error (RMSE) and test misclassification rate evolve during training when $\sigma = 1$ for ONN and ONN+ℓ_2 is also investigated. The results are reported in Figure 2. More details and results can be found in Appendix G.



(a)

(b)

Figure 2: Figure (a) shows the test misclassification rates for all methods (except NTK+ES, which is deferred to Appendix G) vs. σ with their standard deviations plotted as vertical bars. As σ increases, all misclassification rates increase but NTK+ℓ_2 and ONN+ℓ_2 perform significantly better than NTK and ONN with smaller misclassification rate and better stability, i.e., the standard deviation is smaller. Figure (b) shows how the training RMSE and test misclassification rate evolve across iterations for ONN and ONN+ℓ_2 when $\sigma = 1$. For both methods, the training RMSEs decrease fast in the first 1K iterations. However, as the ONN training RMSE flattens after 10K iterations, its test misclassification rate goes up while that for ONN+ℓ_2 remains flat even after 50K iterations, which supports our conjecture in Remark 1. Figure (b) also reveals the potential early stopping time for ONN around iteration 10K, which has test misclassification rate comparable to that of ONN+ℓ_2.

7 Conclusions and discussion

From a nonparametric perspective, this paper studies overparametrized neural networks trained with GD and establishes optimal L_2 convergence rates for trained neural network estimators under the ℓ_2 regularization. In particular, our results bring algorithmic guarantees into the statistical analysis of deep neural networks. Our simulation results corroborate our theoretical analysis, and imply that the assumptions of our theory may be relaxed. More investigations along this direction will advance our statistical understandings of deep learning. For example, our work can be further improved by relaxing the sphere assumption on the input data and assumptions on the learning rate η_1, η_2 and the iteration number k imposed in Theorems 5.1 and 5.2. Additionally, as empirically shown in numerical experiments, it is possible to extend our theory to multi-layer neural networks with other types of activation functions.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[6] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

[7] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[8] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[9] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[10] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2053–2062, 2019.

[11] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019.

[12] Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. *arXiv preprint arXiv:1909.12292*, 2019.

[13] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[14] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

[15] Sidney Siegel. Nonparametric statistics. *The American Statistician*, 11(3):13–19, 1957.

[16] Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning overparameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384*, 2019.

[17] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9709–9721, 2019.

[18] W Hu, Z Li, and D Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *International Conference on Learning Representations*, 2020.

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

[20] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv preprint arXiv:1708.06633*, 2017.

[21] Benedikt Bauer, Michael Kohler, et al. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

[22] Ruiqi Liu, Ben Boukai, and Zuofeng Shang. Optimal nonparametric inference via deep neural network. *arXiv preprint arXiv:1902.01687*, 2019.

[23] Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. *arXiv preprint arXiv:1802.04474*, 2018.

[24] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, pages 12873–12884, 2019.

[25] Ming Yuan, Ding-Xuan Zhou, et al. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–2593, 2016.

[26] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

[27] Michael Kohler and Adam Krzyzak. Over-parametrized deep neural networks do not generalize well. *arXiv preprint arXiv:1912.03925*, 2019.

[28] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, pages 950–957, 1992.

[29] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[31] Berkin Bilgic, Itthi Chatnuntawech, Audrey P Fan, Kawin Setsompop, Stephen F Cauley, Lawrence L Wald, and Elfar Adalsteinsson. Fast image reconstruction with l2-regularization. *Journal of magnetic resonance imaging*, 40(1):181–191, 2014.

[32] Twan Van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

[33] Ekachai Phaisangittisagul. An analysis of the regularization between l2 and dropout in single hidden layer neural network. In *2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*, pages 174–179. IEEE, 2016.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[35] Rich Caruana, Steve Lawrence, and C Lee Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems*, pages 402–408, 2001.

[36] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the Trade*, pages 55–69. Springer, 1998.

[37] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[38] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

[39] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[40] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198*, 2019.

[41] Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

[42] Sara van de Geer. On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electronic Journal of Statistics*, 8(1):543–574, 2014.

[43] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

[44] Richard S Varga. *Gershgorin and His Circles*, volume 36. Springer Science & Business Media, 2010.

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

[46] Kendall Atkinson and Weimin Han. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, volume 2044. Springer Science & Business Media, 2012.

[47] Efthimiou Costas and Frye Christopher. *Spherical Harmonics in p Dimensions*. World Scientific, 2014.

[48] Johann S Brauchart and Josef Dick. A characterization of Sobolev spaces on the sphere and an extension of Stolarsky's invariance principle to arbitrary smoothness. *Constructive Approximation*, 38(3):397–445, 2013.

[49] He Ping Wang, Kai Wang, and Jing Wang. Entropy numbers of Besov classes of generalized smoothness on the sphere. *Acta Mathematica Sinica, English Series*, 30(1):51–60, 2014.

[50] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

A More notations

We introduce some additional notations to be used in the Appendix. Denote $\boldsymbol{y}^* = (f^*(x_1), \cdots, f^*(x_n))^\top$ as the the vector of underlying function's functional values at sample points. Let $\mathbb{I}_r(\boldsymbol{x}) = \mathbb{I}\{\boldsymbol{w}_r^\top \boldsymbol{x} \geq 0\}$ and

$$\boldsymbol{z}(\boldsymbol{x}) = \frac{1}{\sqrt{m}} \begin{pmatrix} a_1 \mathbb{I}_1(\boldsymbol{x})\boldsymbol{x} \\ \vdots \\ a_m \mathbb{I}_m(\boldsymbol{x})\boldsymbol{x} \end{pmatrix} \in \mathbb{R}^{md\times 1}. \tag{A.1}$$

Thus, $\boldsymbol{Z}(k) = (\boldsymbol{z}(\boldsymbol{x}_1), ..., \boldsymbol{z}(\boldsymbol{x}_n))|_{\boldsymbol{W}=\boldsymbol{W}(k)}$. When the context is clear, we omit the dimension and write $\boldsymbol{I}_d$ as $\boldsymbol{I}$.

B Proof of Lemma 3.1

We will use the following lemma, which states the Mercer decomposition of h as in (3.2).

Lemma B.1 (Mercer decomposition of NTK h)**.** For any $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{S}^{d-1}$, we have the following decomposition of the NTK,

$$h(\boldsymbol{s}, \boldsymbol{t}) = \sum_{k=0}^{\infty} \mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}(\boldsymbol{s}) Y_{k,j}(\boldsymbol{t}),$$

where $Y_{k,j}$, $j = 1, ..., N(d,k)$ are spherical harmonic polynomials of degree k, and the non-negative eigenvalues μ_k satisfy $\mu_k \asymp k^{-d}$, and $\mu_k = 0$ if $k = 2j+1$ for $k \geq 2$.

The proof of Lemma B.1 is similar to the proof of Proposition 5 in [24]. The difference is that the Proposition 5 in [24] considers the kernel function

$$h_1(\boldsymbol{s}, \boldsymbol{t}) = 4h(\boldsymbol{s}, \boldsymbol{t}) + \frac{\sqrt{1 - (\boldsymbol{s}^\top \boldsymbol{t})^2}}{\pi},$$

and we only need to consider the kernel function $h(\boldsymbol{s}, \boldsymbol{t})$. A generalization of Proposition 5 in [24] can be found in Theorem 3.5 of [40].

Note that in the proof of Lemma B.1,

$$N(d,j) = \frac{2j+d-2}{j} \begin{pmatrix} j+d-3 \\ d-2 \end{pmatrix} = \frac{\Gamma(j+d-2)}{\Gamma(d-1)\Gamma(j)},$$

where Γ is the Gamma function. By the Stirling approximation, we have $\Gamma(x) \approx \sqrt{2\pi} x^{x-1/2} e^{-x}$. Therefore, we have the number $N(d,j)$ is equivalent to j^{d-2}. Thus, by Lemma B.1, the j-th eigenvalue λ_j can be denoted by

$$\lambda_j = \mu_l, \text{ for } \sum_{i=1}^{l-1} N(d, 2i) \leq j < \sum_{i=1}^{l} N(d, 2i),$$

which can be approximated by $\lambda_j \asymp \mu_l$, for $(2l-2)^{d-1} \leq j < (2l)^{d-1}$. By Lemma B.1, we have $\mu_l \asymp l^{-d}$, which implies $\lambda_j \asymp j^{-\frac{d}{d-1}}$.

C Proof of Theorem 3.2

Let $\mathcal{G}$ be a metric space equipped with a metric d_g. The δ-covering number of the metric space $(\mathcal{G}, d_g)$, denoted by $N(\delta, \mathcal{G}, d_g)$, is the minimum integer N so that there exist N distinct balls in $(\mathcal{G}, d_g)$ with radius δ, and the union of these balls covers $\mathcal{G}$. Let $H(\delta, \mathcal{G}, d_g) = \log N(\delta, \mathcal{G}, d_g)$ be the entropy of the metric space $(\mathcal{G}, d_g)$. We first present an upper bound on the entropy of the metric space $(\mathcal{N}, \|\cdot\|_\infty)$, where the proof can be found in Appendix F.

Lemma C.1. Let $\mathcal{N}$ be the reproducing kernel Hilbert space generated by the NTK h defined in (3.2), equipped with norm $\|\cdot\|_{\mathcal{N}}$. The entropy $H(\delta, \mathcal{N}(1), \|\cdot\|_\infty)$ can be bounded by

$$H(\delta, \mathcal{N}(1), \|\cdot\|_\infty) \leq A_0 \delta^{-\frac{2(d-1)}{d}}, \tag{C.1}$$

where $\mathcal{N}(1) = \{f : f \in \mathcal{N}, \|f\|_{\mathcal{N}} \leq 1\}$, and $A_0 > 0$ is a constant not depending on δ.

For the regression problem, consider a general penalized least-square estimator

$$\widehat{f} := \operatorname*{argmin}_{f \in \mathcal{N}} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda_n^2 I^v(f) \right),$$

where $\lambda_n > 0$ is the smoothing parameter and $I : \mathcal{N} \to [0, \infty)$ is a pseudo-norm measuring the complexity. We use the RKHS norm $\|f\|_{\mathcal{N}(\Omega)}$ in our case. Let $\|\cdot\|_n$ denote the empirical norm. The following lemma establishes the rate of convergence for the estimator $\widehat{f}$.

Lemma C.2 (Lemma 10.2 in [41])**.** Assume Gaussian noises and entropy bound $H(\delta, \mathcal{N}(1), \|\cdot\|_n) \leq A\delta^{-\alpha}$ for some constants $A > 0$ and $0 < \alpha < 2$. If $v \geq \frac{2\alpha}{2+\alpha}$, $I(f^*) > 0$ and

$$\lambda_n^{-1} = O_{\mathbb{P}} \left(n^{1/(2+\alpha)} \right) I^{(2v - 2\alpha + v\alpha)/2(2+\alpha)}(f^*).$$

Then we have

$$\left\| \widehat{f} - f^* \right\|_n = O_{\mathbb{P}}(\lambda_n) I^{v/2}(f^*)$$

and $I(\widehat{f}) = O_{\mathbb{P}}(1) I(f^*)$.

To bound the difference between empirical norm and L_2 norm, we utilize the following lemma. For a class of functions $\mathcal{F}$, define for $z > 0$

$$J_\infty(z, \mathcal{F}) := C_0 \inf_{\delta > 0} \left[z \int_{\delta/4}^{1} \sqrt{\mathcal{H}_\infty(uz/2, \mathcal{F})} du + \sqrt{n} \delta z \right].$$

Lemma C.3 (Theorem 2.2 in [42])**.** Let

$$R := \sup_{f \in \mathcal{F}} \|f\|_2, \; K := \sup_{f \in \mathcal{F}} \|f\|_\infty$$

Then, for all $t > 0$, with probability at least $1 - \exp[-t]$,

$$\sup_{f \in \mathcal{F}} \left| \|f\|_n^2 - \|f\|_2^2 \right| / C_1 \leq \frac{2RJ_\infty(K, \mathcal{F}) + RK\sqrt{t}}{\sqrt{n}} + \frac{4J_\infty^2(K, \mathcal{F}) + K^2 t}{n}$$

where $C_1 > 0$ is some constant not depending on n.

Proof of Theorem 3.2. Consider our estimator $\widehat{f}$ as in (3.4), in which case, $v = 2$ and $I(f)$ is the RKHS norm of f. Since $\|f\|_n \leq \|f\|_\infty$, Lemma C.1 indicates that $\alpha = 2(d-1)/d < 2$. By choosing $\lambda_n \asymp n^{-d/(4d-2)}$, which corresponds to $\mu \asymp n^{(d-1)/(2d-1)}$ in (3.3), Lemma C.2 yields that

$$\left\| \widehat{f} - f^* \right\|_n^2 = O_{\mathbb{P}}(n^{-d/(2d-1)}) \quad \text{and} \quad \left\| \widehat{f} \right\|_{\mathcal{N}}^2 = O_{\mathbb{P}}(1).$$

Now we use Lemma C.3 to obtain a bound on $\left\| \widehat{f} - f^* \right\|_2$. First consider $\{f - f^* : f \in \mathcal{N}(1)\}$. Since $\|f\|_{\mathcal{N}} \leq 1$ for every $f \in \mathcal{N}(1)$, we have $K, R = O(1)$. By the entropy bound in Lemma C.1 we have $J_\infty(z, \mathcal{N}(1)) \leq 2C_0 z^{1/d}$. Therefore, Lemma C.3 yields

$$\sup_{f \in \mathcal{N}(1)} \left| \|f - f^*\|_n^2 - \|f - f^*\|_2^2 \right| = O_{\mathbb{P}} \left(\sqrt{\frac{1}{n}} \right).$$

Combined with $\left\|\widehat{f}-f^*\right\|_n^2 = O_{\mathbb{P}}(n^{-d/(2d-1)})$, we can conclude that for any $t > 0$ large enough, $\left\|\widehat{f}-f^*\right\|_2^2 = O(\sqrt{t/n})$ with probability at least $1-\exp(-t)$. Utilizing Lemma C.3 again with $R = O(\sqrt{t/n})$ we have for some $C > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{G}(R)}\left|\|f-f^*\|_n^2 - \|f-f^*\|_2^2\right| \le \frac{Ct}{n}\right) \ge 1-e^{-t},$$

where $\mathcal{G}(R) := \{f \in \mathcal{N}(1) : \|f-f^*\|_2 \le R\}$. Notice that $\widehat{f} \in \mathcal{G}(R)$ with probability at least $1-\exp(-t)$. Therefore, $\left\|\widehat{f}-f^*\right\|_2^2 = O(n^{-d/(2d-1)} + t/n)$ with probability at least $1-2\exp(-t)$. □

D Proofs of main theorems in Section 4

For brevity, let $\widehat{f}_k = f_{\boldsymbol{W}(k),\boldsymbol{a}}$. For two positive semidefinite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we write $\boldsymbol{A} \ge \boldsymbol{B}$ to denote that $\boldsymbol{A}-\boldsymbol{B}$ is positive semidefinite and $\boldsymbol{A} > \boldsymbol{B}$ to denote that $\boldsymbol{A}-\boldsymbol{B}$ is positive definite. This partial order of positive semidefinite matrices is also known as Loewner order. We focus on the L_2 loss of our estimator $\widehat{f}_k$ after k GD updates. Let $\widetilde{f}$ denote the kernel regression solution with kernel $h(\cdot,\cdot)$ that interpolates all $\{(\boldsymbol{x}_i, f^*(\boldsymbol{x}_i))\}_{i=1}^n$, i.e.,

$$g(\boldsymbol{x}) = h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}^*. \tag{D.1}$$

We first provide some lemmas used in this section. The proofs of lemmas are presented in Appendix F. Lemma D.1 states some basic inequalities that are also used in the proof of Theorem 5.1. Lemma D.2 provides the convergence rate of interpolant using NTK. Lemmas D.3 can be found in [9]. Lemma D.4 is implied by the proof in [9]. Lemma D.5 provides some bounds on the related quantities used in the proofs of Theorems 4.1 and 5.2. Lemma D.6 provide some properties of Loewner order.

Lemma D.1. Let μ be as in Theorem 3.2. Then we have

$$h(\boldsymbol{s},\boldsymbol{s}) - h(\boldsymbol{s},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X},\boldsymbol{s}) \ge 0,$$
$$\int_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty+\mu\boldsymbol{I})^{-2}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x} = O_{\mathbb{P}}(n^{-\frac{d}{2d-1}}),$$
$$\int_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x},\boldsymbol{x}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x} = O_{\mathbb{P}}(n^{-\frac{1}{2d-1}}),$$

where $h(\boldsymbol{x},\boldsymbol{X}) = (h(\boldsymbol{x},\boldsymbol{x}_1),...,h(\boldsymbol{x},\boldsymbol{x}_n))$ and $h(\boldsymbol{X},\boldsymbol{x}) = h(\boldsymbol{x},\boldsymbol{X})^\top$.

Lemma D.2. Assume the true function $f^* \in \mathcal{N}$ with finite RKHS norm, then $g(\boldsymbol{x})$ defined (D.1) satisfies

$$\|g-f^*\|_2 = O_{\mathbb{P}}\left(n^{-1/2}\right).$$

Lemma D.3 (Lemma C.1 in [9])**.** If $\lambda_0 = \lambda_{min}(\boldsymbol{H}^\infty) > 0$, $m = \Omega\left(\frac{n^6}{\lambda_0^4\tau^2\delta^3}\right)$ and $\eta = O\left(\frac{\lambda_0}{n^2}\right)$, with probability at least $1-\delta$ over the random initialization, we have

$$\|\boldsymbol{w}_r(k)-\boldsymbol{w}_r(0)\|_2 \le R_0,\ \ \forall\, r\in[m], \forall\, k\ge 0,$$

where $R_0 = \frac{4\sqrt{n}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2}{\sqrt{m}\lambda_0}$.

Lemma D.4 ([9])**.** Denote $u_i(k) = f_{\boldsymbol{W}(k),\boldsymbol{a}}(\boldsymbol{x}_i)$ to be the network's prediction on the i-th input and let $\boldsymbol{u}(k) = (u_1(k),...,u_n(k))^\top \in \mathbb{R}^n$ denote all n predictions on the points $\boldsymbol{x}_1,...,\boldsymbol{x}_n$ at iteration k. We have

$$\boldsymbol{u}(k)-\boldsymbol{y} = (\boldsymbol{I}-\eta\boldsymbol{H}^\infty)^k(\boldsymbol{u}(0)-\boldsymbol{y}) + \boldsymbol{e}(k)$$

where

$$\|\boldsymbol{e}(k)\|_2 = O\left(k\left(1-\frac{\eta\lambda_0}{4}\right)^{k-1}\frac{\eta n^{5/2}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^2}{\sqrt{m}\lambda_0\tau\delta}\right).$$

Lemma D.5. With probability at least $1-\delta$, we have

(a) $\|\boldsymbol{Z}(k)-\boldsymbol{Z}(0)\|_F = O\left(\frac{n^{3/4}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^{1/2}}{\sqrt{m^{1/2}\lambda_0\tau\delta}}\right)$;

(b) $\|\boldsymbol{H}(0)-\boldsymbol{H}^\infty\|_F = O\left(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}\right)$;

(c) $\left\|\boldsymbol{z}_0(\cdot)^\top\boldsymbol{Z}(0)-h(\cdot,\boldsymbol{X})\right\|_2 = O\left(\frac{\sqrt{n}\sqrt{\log(n/\delta)}}{\sqrt{m}}\right)$;

(d) $\left\|z_0(\cdot)^\top\mathrm{vec}(\boldsymbol{W}(0))\right\|_2 = O\left(\tau\sqrt{\log(1/\delta)}\right)$.

Lemma D.6 (Properties of Loewner order)**.** For two positive semidefinite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

(a). Suppose $\boldsymbol{A}$ is nonsingular, then $\boldsymbol{A}\geq\boldsymbol{B}\iff\lambda_{max}(\boldsymbol{B}\boldsymbol{A}^{-1})\leq 1$ and $\boldsymbol{A}>\boldsymbol{B}\iff\lambda_{\max}(\boldsymbol{B}\boldsymbol{A}^{-1})>1$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of the input matrix.

(b). Suppose $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{Q}$ are positive definite, $\boldsymbol{A}$ and $\boldsymbol{B}$ are exchangeable, then $\boldsymbol{A}\geq\boldsymbol{B}\Longrightarrow\boldsymbol{A}\boldsymbol{Q}\boldsymbol{A}\geq\boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}$.

D.1 Proof of Theorem 4.1

For notational simplification, we use $\widehat{f}_k = f_{\boldsymbol{W}(k),\boldsymbol{a}}$. Define

$$\widetilde{f}_k(\boldsymbol{x}) = \mathrm{vec}(\boldsymbol{W}(k))^\top\boldsymbol{z}_0(\boldsymbol{x}), \tag{D.2}$$

where $\boldsymbol{z}_0(\boldsymbol{x}) = \boldsymbol{z}(\boldsymbol{x})|_{\boldsymbol{W}=\boldsymbol{W}(0)}$. Then we can write the following decomposition

$$\widehat{f}_k - f^* = (\widehat{f}_k-\widetilde{f}_k)+(\widetilde{f}_k-g)+(g-f^*) = \Delta_1+\Delta_2+\Delta_3, \tag{D.3}$$

where g is as in (D.1). It follows from Lemma D.2 that

$$\|\Delta_3\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{1}{n}}\right). \tag{D.4}$$

For Δ_1, under the assumptions of Lemma D.3, with high probability, we have $\|\boldsymbol{w}_r(k)-\boldsymbol{w}_r(0)\|_2\leq R_0$. Thus, for fixed $\boldsymbol{x}$, we have

$$|\boldsymbol{w}_r(k)^\top\boldsymbol{x}-\boldsymbol{w}_r(0)^\top\boldsymbol{x}|\leq\|\boldsymbol{w}_r(k)-\boldsymbol{w}_r(0)\|_2\,\|\boldsymbol{x}\|_2\leq R_0.$$

Define event

$$B_r(\boldsymbol{x}) = \{|\boldsymbol{w}_r(0)^\top\boldsymbol{x}|\leq R_0\}, \forall r\in[m].$$

If $\mathbb{I}\{B_r(\boldsymbol{x})\}=0$, then we have $\mathbb{I}_{r,k}(\boldsymbol{x})=\mathbb{I}_{r,0}(\boldsymbol{x})$, where $\mathbb{I}_{r,k}(\boldsymbol{x})=\mathbb{I}\{\boldsymbol{w}_r(k)^\top\boldsymbol{x}\geq 0\}$. Therefore, for any fixed $\boldsymbol{x}$, we have

$$\begin{aligned}
|\widehat{f}_k(\boldsymbol{x})-\widetilde{f}_k(\boldsymbol{x})| &= \left|\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r(\mathbb{I}_{r,k}(\boldsymbol{x})-\mathbb{I}_{r,0}(\boldsymbol{x}))\boldsymbol{w}_r(k)^\top\boldsymbol{x}\right| \\
&= \left|\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\mathbb{I}\{B_r(\boldsymbol{x})\}(\mathbb{I}_{r,k}(\boldsymbol{x})-\mathbb{I}_{r,0}(\boldsymbol{x}))\boldsymbol{w}_r(k)^\top\boldsymbol{x}\right| \\
&\leq \frac{1}{\sqrt{m}}\sum_{r=1}^m\mathbb{I}\{B_r(\boldsymbol{x})\}|\boldsymbol{w}_r(k)^\top\boldsymbol{x}| \\
&\leq \frac{1}{\sqrt{m}}\sum_{r=1}^m\mathbb{I}\{B_r(\boldsymbol{x})\}\left(|\boldsymbol{w}_r(0)^\top\boldsymbol{x}|+|\boldsymbol{w}_r(k)^\top\boldsymbol{x}-\boldsymbol{w}_r(0)^\top\boldsymbol{x}|\right) \\
&\leq \frac{2R_0}{\sqrt{m}}\sum_{r=1}^m\mathbb{I}\{B_r(x)\}
\end{aligned}$$

Recall that $\|\boldsymbol{x}\|_2 = 1$, which implies that $\boldsymbol{w}_r(0)^\top \boldsymbol{x}$ is distributed as $N(0, \tau^2)$. Therefore, we have

$$\mathbb{E}[\mathbb{I}\{B_r(x)\}] = \mathbb{P}\left(|\boldsymbol{w}_r(0)^\top \boldsymbol{x}| \le R_0\right) = \int_{-R_0}^{R_0} \frac{1}{\sqrt{2\pi\tau}} \exp\left\{-\frac{u^2}{2\tau^2}\right\} du \le \frac{2R_0}{\sqrt{2\pi\tau}}.$$

By Markov's inequality, with probability at least $1-\delta$, we have

$$\sum_{r=1}^{m} \mathbb{I}\{B_r(x)\} \le \frac{2mR_0}{\sqrt{2\pi\tau\delta}}.$$

Thus, we have

$$\|\Delta_1\|_2 \le \frac{2R_0}{\sqrt{m}} \left\|\sum_{r=1}^{m} \mathbb{I}\{B_r(\cdot)\}\right\|_2 \le \frac{4\sqrt{m}R_0^2}{\sqrt{2\pi\tau\delta}} = O\left(\frac{n\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^2}{\sqrt{m}\tau\lambda_0^2\delta}\right). \tag{D.5}$$

Next, we evaluate Δ_2. Recall that the GD update rule is

$$\mathrm{vec}(\boldsymbol{W}(j+1)) = \mathrm{vec}(\boldsymbol{W}(j)) - \eta \boldsymbol{Z}(j)(\boldsymbol{u}(j) - \boldsymbol{y}), j \ge 0.$$

Applying Lemma D.4, we can get

$$\begin{aligned}
&\mathrm{vec}(\boldsymbol{W}(k)) - \mathrm{vec}(\boldsymbol{W}(0)) \\
=&\sum_{j=0}^{k-1}(\mathrm{vec}(\boldsymbol{W}(j+1)) - \mathrm{vec}(\boldsymbol{W}(j))) \\
=&-\sum_{j=0}^{k-1}\eta \boldsymbol{Z}(j)(\boldsymbol{u}(j) - \boldsymbol{y}) \\
=&\sum_{j=0}^{k-1}\eta \boldsymbol{Z}(j)(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^j(\boldsymbol{y} - \boldsymbol{u}(0)) - \sum_{j=0}^{k-1}\eta\boldsymbol{Z}(j)\boldsymbol{e}(j) \\
=&\sum_{j=0}^{k-1}\eta \boldsymbol{Z}(0)(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^j(\boldsymbol{y} - \boldsymbol{u}(0)) + \sum_{j=0}^{k-1}\eta(\boldsymbol{Z}(j) - \boldsymbol{Z}(0))(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^j(\boldsymbol{y} - \boldsymbol{u}(0)) - \sum_{j=0}^{k-1}\eta\boldsymbol{Z}(j)\boldsymbol{e}(j) \\
=&\sum_{j=0}^{k-1}\eta \boldsymbol{Z}(0)(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^j(\boldsymbol{y} - \boldsymbol{u}(0)) + \zeta(k).
\end{aligned}$$

For the first term of $\zeta(k)$, applying Lemma D.5 (a), with probability at least $1-\delta$, we get

$$\begin{aligned}
&\left\|\sum_{j=0}^{k-1}\eta(\boldsymbol{Z}(j) - \boldsymbol{Z}(0))(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^j(\boldsymbol{y} - \boldsymbol{u}(0))\right\|_2 \\
\le&\sum_{j=0}^{k-1} O\left(\frac{n^{3/4}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^{1/2}}{\sqrt{m^{1/2}\lambda_0\tau\delta}}\right)\eta\|\boldsymbol{I} - \eta\boldsymbol{H}^\infty\|_2^j\|(\boldsymbol{y}-\boldsymbol{u}(0))\|_2 \\
\le&O\left(\frac{n^{3/4}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^{3/2}}{\sqrt{m^{1/2}\lambda_0\tau\delta}}\right)\sum_{j=0}^{k-1}\eta(1-\eta\lambda_0)^j \\
=&O\left(\frac{n^{3/4}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2^{3/2}}{m^{1/4}\tau^{1/2}\lambda_0^{3/2}\delta^{1/2}}\right).
\end{aligned}$$

Denote that $z_i(j) = z(\boldsymbol{x}_i)|_{\boldsymbol{W}=\boldsymbol{W}(j)}$. By (A.1), we have $\|\boldsymbol{z}_i(j)\|_2 \le 1$. Thus,

$$\|\boldsymbol{Z}(j)\|_F = \left(\sum_{i=1}^{n}\|\boldsymbol{z}_i(j)\|_2^2\right)^{\frac{1}{2}} \le \sqrt{n}\ , \forall\, j \ge 0. \tag{D.6}$$

For the second term of $\zeta(k)$, we have

$$\begin{aligned}
&\left\|\sum_{j=0}^{k-1} \eta \boldsymbol{Z}(j)\boldsymbol{e}(j)\right\|_2 \\
\leq &\sum_{j=0}^{k-1} \eta \left\|\boldsymbol{Z}(j)\right\|_F \left\|\boldsymbol{e}(j)\right\|_2 \\
\leq &\sum_{j=0}^{k-1} \eta\sqrt{n} O\left(j\left(1-\frac{\eta\lambda_0}{4}\right)^{j-1} \frac{\eta n^{5/2}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^2}{\sqrt{m}\tau\lambda_0\delta}\right) \\
= &O\left(\frac{n^3\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^2}{\sqrt{m}\lambda_0^3\tau\delta}\right).
\end{aligned}$$

Therefore,

$$\left\|\zeta(k)\right\|_2 = O\left(\frac{n^{3/4}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^{3/2}}{m^{1/4}\tau^{1/2}\lambda_0^{3/2}\delta^{1/2}}\right) + O\left(\frac{n^3\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^2}{\sqrt{m}\lambda_0^3\tau\delta}\right). \tag{D.7}$$

Define $\boldsymbol{G}_k = \sum_{j=0}^{k-1}\eta(\boldsymbol{I}-\eta\boldsymbol{H}^\infty)^j$. Recalling that $\boldsymbol{y}=\boldsymbol{y}^*+\boldsymbol{\epsilon}$, for fixed $\boldsymbol{x}$, we have

$$\begin{aligned}
\widetilde{f}_k(\boldsymbol{x}) - g(\boldsymbol{x}) =& \boldsymbol{z}_0(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{W}(k)) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}^* \\
=& \boldsymbol{z}_0(\boldsymbol{x})^\top\left[\boldsymbol{Z}(0)\boldsymbol{G}_k(\boldsymbol{y}-\boldsymbol{u}(0)) + \zeta(k) + \mathrm{vec}(\boldsymbol{W}(0))\right] \\
=& \left[h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{G}_k - (\boldsymbol{H}^\infty)^{-1})\boldsymbol{y}^* + h(\boldsymbol{x},\boldsymbol{X})\boldsymbol{G}_k\boldsymbol{\epsilon}\right] + \left[\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0) - h(\boldsymbol{x},\boldsymbol{X})\right]\boldsymbol{G}_k\boldsymbol{y} \\
&+ \left[\boldsymbol{z}_0(\boldsymbol{x})^\top\mathrm{vec}(\boldsymbol{W}(0)) + \boldsymbol{z}_0(\boldsymbol{x})^\top\zeta(k) - \boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0)\boldsymbol{G}_k\boldsymbol{u}(0)\right] \\
=& \Delta_{21}(\boldsymbol{x}) + \Delta_{22}(\boldsymbol{x}) + \Delta_{23}(\boldsymbol{x}).
\end{aligned} \tag{D.8}$$

Using Lemma D.5 (c), we can bound Δ_{22} as

$$\begin{aligned}
\left\|\Delta_{22}\right\|_2 \leq& \left\|\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0) - h(\boldsymbol{x},\boldsymbol{X})\right\|_2 \left\|\boldsymbol{G}_k\boldsymbol{y}\right\|_2 \\
\leq& O\left(\frac{\sqrt{n}\sqrt{\log(n/\delta)}}{\sqrt{m}}\right)\left\|(\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}\right\|_2 \\
=& O\left(\frac{\sqrt{n}\sqrt{\log(n/\delta)}\left\|\boldsymbol{y}\right\|_2}{\sqrt{m}\lambda_0}\right).
\end{aligned} \tag{D.9}$$

Since the i-th coordinate of $\boldsymbol{u}(0)$ is

$$u_i(0) = \boldsymbol{z}_0(\boldsymbol{x}_i)^\top\mathrm{vec}(\boldsymbol{W}(0)) = \sum_{r=1}^m a_r \boldsymbol{w}(0)^\top\boldsymbol{x}_i\mathbb{I}\{\boldsymbol{w}(0)^\top\boldsymbol{x}_i\},$$

where $a_r \sim \mathrm{unif}\{1,-1\}$ and $\boldsymbol{w}(0)^\top\boldsymbol{x}_i \sim N(0,\tau^2)$, it is easy to prove that $u_i(0)$ has zero mean and variance τ^2. This implies $\mathbb{E}[\left\|\boldsymbol{u}(0)\right\|_2^2] = O(n\tau^2)$. By Markov's inequality, with probability at least $1-\delta$, we have $\left\|\boldsymbol{u}(0)\right\|_2 = O\left(\frac{\sqrt{n}\tau}{\delta}\right)$. Similar to (D.6), we can obtain $\left\|\boldsymbol{Z}(0)\right\|_F = O(\sqrt{n})$. Thus,

$$\left|\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0)\boldsymbol{G}_k\boldsymbol{u}(0)\right| \leq \left\|\boldsymbol{z}_0(\boldsymbol{x})\right\|_2\left\|\boldsymbol{Z}(0)\right\|_F\left\|\boldsymbol{G}_k\boldsymbol{u}(0)\right\|_2 \leq \sqrt{n}\left\|(\boldsymbol{H}^\infty)^{-1}\boldsymbol{u}(0)\right\|_2 = O\left(\frac{n\tau}{\lambda_0\delta}\right). \tag{D.10}$$

Combining Lemma D.5 (d), (D.7) and (D.10), we obtain

$$\begin{aligned}
\left\|\Delta_{23}\right\|_2 \leq& \left\|\boldsymbol{z}_0(\cdot)^\top\mathrm{vec}(\boldsymbol{W}(0))\right\|_2 + \left\|\boldsymbol{z}_0(\cdot)\right\|_2\left\|\zeta(k)\right\|_2 + \left\|\boldsymbol{z}_0(\cdot)^\top\boldsymbol{Z}(0)\boldsymbol{G}_k\boldsymbol{u}(0)\right\|_2 \\
=& O\left(\tau\sqrt{\log(1/\delta)}\right) + O\left(\frac{n^{3/4}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^{3/2}}{m^{1/4}\tau^{1/2}\lambda_0^{3/2}\delta^{1/2}}\right) + O\left(\frac{n^3\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^2}{\sqrt{m}\lambda_0^3\tau\delta}\right) + O\left(\frac{n\tau}{\lambda_0\delta}\right) \\
=& O\left(\frac{n^{3/4}\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^{3/2}}{m^{1/4}\tau^{1/2}\lambda_0^{3/2}\delta^{1/2}}\right) + O\left(\frac{n^3\left\|\boldsymbol{y}-\boldsymbol{u}(0)\right\|_2^2}{\sqrt{m}\lambda_0^3\tau\delta}\right) + O\left(\frac{n\tau}{\lambda_0\delta}\right).
\end{aligned} \tag{D.11}$$

By (D.3) and (D.8), we can rewrite $\widehat{f}_k - f^*$ as

$$\widehat{f}_k - f^* = \Delta_{21} + (\Delta_1 + \Delta_3 + \Delta_{22} + \Delta_{23}) := \Delta_{21} + \Xi,$$

Next we bound the expected value of $\|\Xi\|_2^2$ over noise, $\mathbb{E}_{\boldsymbol{\epsilon}} \|\Xi\|_2^2$. Note that we have

$$\mathbb{E}_{\boldsymbol{\epsilon}} \|\boldsymbol{y}\|_2^2 = \mathbb{E}_{\boldsymbol{\epsilon}} \|\boldsymbol{y}^* + \boldsymbol{\epsilon}\|_2^2 \leq 2\boldsymbol{y}^{*\top}\boldsymbol{y}^* + 2\mathbb{E}_{\boldsymbol{\epsilon}}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} = O(n). \tag{D.12}$$

By Markov's inequality, with probability $1-\delta$ over random initialization, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} \|\boldsymbol{y} - \boldsymbol{u}(0)\|_2 \leq & \left(\mathbb{E}_{\boldsymbol{\epsilon}} \|\boldsymbol{y} - \boldsymbol{u}(0)\|_2^2\right)^{\frac{1}{2}} \\
\leq & \left(\frac{3\mathbb{E}_{\boldsymbol{W}(0),\boldsymbol{a}}\left[\boldsymbol{u}(0)^\top\boldsymbol{u}(0) + \boldsymbol{y}^{*\top}\boldsymbol{y}^* + \mathbb{E}_{\boldsymbol{\epsilon}}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}\right]}{\delta}\right)^{\frac{1}{2}} \\
= & O\left(\sqrt{\frac{n(1+\tau^2)}{\delta}}\right) = O\left(\sqrt{\frac{n}{\delta}}\right),
\end{aligned} \tag{D.13}$$

where the last equality of D.13 is because $\tau^2 \lesssim 1$. By (D.4), (D.5), (D.9), (D.11), (D.12) and (D.13), $\mathbb{E}_{\boldsymbol{\epsilon}} \|\Xi\|_2^2$ can be upper bounded as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} \|\Xi\|_2^2 \leq & 4\mathbb{E}_{\boldsymbol{\epsilon}}(\|\Delta_1\|_2^2 + \|\Delta_3\|_2^2 + \|\Delta_{22}\|_2^2 + \|\Delta_{23}\|_2^2) \\
= & \mathbb{E}_{\boldsymbol{\epsilon}}\left[O\left(\frac{n^2\|\boldsymbol{y} - \boldsymbol{u}(0)\|_2^4}{m\tau^2\lambda_0^4\delta^2}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{n\log(n/\delta)\|\boldsymbol{y}\|_2^2}{m\lambda_0^2}\right)\right] + 4\mathbb{E}_{\boldsymbol{\epsilon}}\|\Delta_{23}\|_2^2 \\
\leq & O\left(\frac{n^4}{m\tau^2\lambda_0^4\delta^4}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{n^2\log(n/\delta)}{m\lambda_0^2\delta}\right) + O\left(\frac{n^2\tau^2}{\lambda_0^2\delta^2}\right) + \\
& + \mathbb{E}_{\boldsymbol{\epsilon}}\left[O\left(\frac{n^{3/2}\|\boldsymbol{y} - \boldsymbol{u}(0)\|_2^3}{m^{1/2}\tau\lambda_0^3\delta}\right) + O\left(\frac{n^6\|\boldsymbol{y} - \boldsymbol{u}(0)\|_2^4}{m\tau^2\lambda_0^6\delta^2}\right)\right] \\
= & O\left(\frac{n^4}{m\tau^2\lambda_0^4\delta^4}\right) + O\left(\frac{1}{n}\right) + O\left(\frac{n^2\log(n/\delta)}{m\lambda_0^2\delta}\right) + O\left(\frac{n^2\tau^2}{\lambda_0^2\delta^2}\right) \\
& + O\left(\frac{n^3}{\sqrt{m}\tau\lambda_0^3\delta^{5/2}}\right) + O\left(\frac{n^8}{m\tau^2\lambda_0^6\delta^4}\right) \\
= & O\left(\frac{1}{n}\right) + O\left(\frac{n^2\tau^2}{\lambda_0^2\delta^2}\right) + \frac{\text{poly}\left(n, \frac{1}{\lambda_0}, \frac{1}{\delta}\right)}{m^{\frac{1}{2}}\tau}.
\end{aligned}$$

In the following, we will evaluate Δ_{21} and discuss how the iteration number k would affect the L_2 estimation error $\left\|\widehat{f}_k - f^*\right\|_2^2$.

Case 1: The iteration number k cannot be too small By taking expectation of $\|\Delta_{21}\|_2^2$ over the noise, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}} \|\Delta_{21}\|_2^2 = & \int_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x}, \boldsymbol{X})\big[((\boldsymbol{H}^\infty)^{-1} - \boldsymbol{G}_k)\boldsymbol{y}^*\boldsymbol{y}^{*\top}((\boldsymbol{H}^\infty)^{-1} - \boldsymbol{G}_k) + \boldsymbol{G}_k^2\big]h(\boldsymbol{X}, \boldsymbol{x})d\boldsymbol{x} \\
= & \int_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}\boldsymbol{M}_k(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X}, \boldsymbol{x})d\boldsymbol{x},
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{M}_k = & (\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k\boldsymbol{S}(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k + (\boldsymbol{I} - (\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k)^2 \\
= & [(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k - (\boldsymbol{S} + \boldsymbol{I})^{-1}](\boldsymbol{S} + \boldsymbol{I})[(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k - (\boldsymbol{S} + \boldsymbol{I})^{-1}] + \boldsymbol{I} - (\boldsymbol{S} + \boldsymbol{I})^{-1}
\end{aligned} \tag{D.14}$$

and $\boldsymbol{S} = \boldsymbol{y}^*\boldsymbol{y}^{*\top}$. If $k \geq C_0\left(\frac{\log n}{\eta\lambda_0}\right)$ for some constant $C_0 > 1$, we have

$$(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^k \leq (1 - \eta\lambda_0)^k\boldsymbol{I} \leq \exp\{-\eta\lambda_0 k\}\boldsymbol{I} \leq \exp\{-C_0\log n\}\boldsymbol{I} = \frac{1}{n^{C_0}}\boldsymbol{I},$$

Since $1+\|\boldsymbol{y}^*\|_2^2 \leq C_1 n$ for some constant C_1, we have

$$\lambda_{\max}\left(\frac{1}{n^{C_0}}(\boldsymbol{S}+\boldsymbol{I})\right) = \frac{1+\|\boldsymbol{y}^*\|_2^2}{n^{C_0}} \leq \frac{C_1}{n^{C_0-1}} < 1.$$

By Lemma D.6 (a), we have

$$(\boldsymbol{I}-\eta\boldsymbol{H}^\infty)^k \leq \frac{1}{n^{C_0}}\boldsymbol{I} < (\boldsymbol{S}+\boldsymbol{I})^{-1}.$$

Therefore, we have

$$(\boldsymbol{S}+\boldsymbol{I})^{-1} - (\boldsymbol{I}-\eta\boldsymbol{H}^\infty)^k \geq (\boldsymbol{S}+\boldsymbol{I})^{-1} - \frac{1}{n^{C_0}}\boldsymbol{I},$$

where $(\boldsymbol{S}+\boldsymbol{I})^{-1} - (\boldsymbol{I}-\eta\boldsymbol{H}^\infty)^k$ and $(\boldsymbol{S}+\boldsymbol{I})^{-1} - n^{-C_0}\boldsymbol{I}$ are positive definite matrices. It is also obvious that the two matrices are exchangeable. By Lemma D.6 (b) and (D.14), we have

$$\boldsymbol{M}_k \geq \left(1-\frac{1}{n^{C_0}}\right)^2 \boldsymbol{I} + \frac{1}{n^{2C_0}}\boldsymbol{S}.$$

Then we have

$$\mathbb{E}_{\boldsymbol{\epsilon}}\|\Delta_{21}\|_2^2 \geq \left(1-\frac{1}{n^{C_0}}\right)^2 I_1 + \frac{1}{n^{2C_0}} I_2 \geq c_0 I_1$$

where $c_0 \in (0,1)$ is a constant,

$$I_1 = \int h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-2}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x}, \quad \text{and} \quad I_2 = \int [h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}^*]^2 d\boldsymbol{x}.$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\widehat{f}_k - f^*\right\|_2^2 =& \mathbb{E}_{\boldsymbol{\epsilon}}\|\Delta_{21}+\Xi\|_2^2 \\ \geq& \frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}}\|\Delta_{21}\|_2^2 - \mathbb{E}_{\boldsymbol{\epsilon}}\|\Xi\|_2^2 \\ \geq& \frac{c_0}{2}I_1 - O\left(\frac{1}{n}\right) - O\left(\frac{n^2\tau^2}{\lambda_0^2\delta^2}\right) - \frac{\operatorname{poly}\left(n, \frac{1}{\lambda_0}, \frac{1}{\delta}\right)}{m^{\frac{1}{2}}\tau}. \end{aligned} \tag{D.15}$$

Let $\tau \leq C_3\frac{\lambda_0\delta}{n}\left\|(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X},\cdot)\right\|_2$ for some constant $C_3>0$ such that the third term of (D.15) is bounded by $\frac{c_0}{4}\left\|(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X},\cdot)\right\|_2^2$. Therefore, $\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\widehat{f}_k - f^*\right\|_2^2$ can be lower bounded as

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\widehat{f}_k - f^*\right\|_2^2 \geq C_1^*\left\|(\boldsymbol{H}^\infty)^{-1}h(\boldsymbol{X},\cdot)\right\|_2^2 - O\left(\frac{1}{n}\right), \tag{D.16}$$

where $C_1^*>0$ is a constant. Note that I_1 is $\mathbb{E}_\epsilon\left\|\widehat{f}_\infty - g^*\right\|_2^2$, where $g^*\equiv 0$ and $\widehat{f}_\infty$ is the interpolated estimator of g^*, as in Theorem 4.2. Therefore, by Theorem 4.2, there exists a constant c_1 such that $\mathbb{E}_\epsilon\left\|\widehat{f}_\infty - g^*\right\|_2^2 \geq c_1$, which implies $I_1 \geq c_1$. Taking n large enough such that the second term in (D.16) is smaller than $C_1^*c_1$, we finish the proof of the case that k is large.

Case 2: The iteration number k cannot be too large We can rewrite Δ_{21} as

$$\begin{aligned}\Delta_{21} =& h(\boldsymbol{x},\boldsymbol{X})\boldsymbol{G}_k(\boldsymbol{y}^*+\boldsymbol{\epsilon}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}^* \\ =& \Delta_{21}^* - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}^*.\end{aligned}$$

Since

$$\boldsymbol{G}_k = \sum_{j=0}^{k-1}\eta(\boldsymbol{I}-\eta\boldsymbol{H}^\infty)^j = \sum_{j=0}^{k-1}\eta\sum_{i=1}^n(1-\eta\lambda_i)^j\boldsymbol{v}_i\boldsymbol{v}_i^\top \leq \eta k\boldsymbol{I},$$

we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\Delta_{21}^*\right\|_2^2 =& \int_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x},\boldsymbol{X})\boldsymbol{G}_k(\boldsymbol{S}+\boldsymbol{I})\boldsymbol{G}_k h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x}\\ \le&\eta^2k^2\int_{\boldsymbol{x}\in\Omega} h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{S}+\boldsymbol{I})h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x}\\ =&\eta^2k^2\left(\int_{\boldsymbol{x}\in\Omega}\left[h(\boldsymbol{x},\boldsymbol{X})\boldsymbol{y}^*\right]^2d\boldsymbol{x}+\left\|h(\cdot,\boldsymbol{X})\right\|_2^2\right)\\ =&O\left(\eta^2k^2n^2\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\widehat{f}_k-f^*\right\|_2^2 =&\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\Delta_{21}^*+\Xi-h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}^*\right\|_2^2\\ \ge&\frac{1}{2}\left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}^*\right\|_2^2-\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\Delta_{21}^*+\Xi\right\|_2^2\\ \ge&\frac{1}{2}\left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}^*\right\|_2^2-2\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\Delta_{21}^*\right\|_2^2-2\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\Xi\right\|_2^2\\ \ge&\frac{1}{2}\left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}^*\right\|_2^2-O\left(\eta^2k^2n^2\right)\\ &-O\left(\frac{1}{n}\right)-O\left(\frac{n^2\tau^2}{\lambda_0^2\delta^2}\right)-\frac{\text{poly}\left(n,\frac{1}{\lambda_0},\frac{1}{\delta}\right)}{m^{\frac{1}{2}}\tau}.\end{aligned}\tag{D.17}$$

Let $k\le C_1\left(\frac{1}{\eta n}\right)$ for some constant $C_1>0$ such that the the second term of (D.17) can be bounded by $\frac{1}{8}\left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}^*\right\|_2^2$. Let $\tau\le C_2\left(\frac{\delta\lambda_0}{n}\right)$ for some constant $C_2>0$ such that the fourth term in (D.17) can be bounded by $\frac{1}{8}\left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}^*\right\|_2^2$. Note that we can also choose m such that the fifth term in (D.17) is bounded by $\frac{1}{8}\left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}^*\right\|_2^2$. Therefore, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\widehat{f}_k-f^*\right\|_2^2\ge&C_2^*\left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}\boldsymbol{y}^*\right\|_2^2-O\left(\frac{1}{n}\right)\\ \ge&C_3^*\left\|f^*\right\|_2^2-O\left(\frac{1}{n}\right),\end{aligned}\tag{D.18}$$

where the last inequality is because of Lemma D.2, and $C_2^*>0$ is a constant. By taking n large enough such that the second term in (D.18) is smaller than $C_3^*\left\|f^*\right\|_2^2/2$, we finish the proof.

D.2 Proof of Theorem 4.2

Let's first introduce the GD update for the kernel ridge regression. By the representer theorem [43], the kernel estimator can be written as

$$\widehat{f}(\boldsymbol{x})=\sum_{i=1}^n\omega_i h(\boldsymbol{x},\boldsymbol{x}_i):=h(\boldsymbol{x},\boldsymbol{X})\boldsymbol{\omega},$$

where $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_n)$ is the coefficient vector. Consider using the squared loss

$$\Phi(\boldsymbol{\omega})=\frac{1}{2}\sum_{i=1}^n(\widehat{f}(\boldsymbol{x}_i)-y_i)^2.$$

Let $\boldsymbol{\omega}_k$ be the $\boldsymbol{\omega}$ at the k-th GD iteration and choose $\boldsymbol{\omega}_0=\mathbf{0}$. Then, the GD update rule for estimating $\boldsymbol{\omega}$ can be expressed as

$$\boldsymbol{\omega}_{k+1}=\boldsymbol{\omega}_k-\eta\left((\boldsymbol{H}^{\infty})^2\boldsymbol{\omega}-\boldsymbol{H}^{\infty}\boldsymbol{y}\right)\tag{D.19}$$

In the formulation of the stopping rule, two quantities play an important role: first, the running sum of the step sizes $\alpha_j:=\sum_{i=0}^j\eta_i$, and secondly, the eigenvalues $\widehat{\lambda}_1\ge\widehat{\lambda}_2\ge\cdots\ge\widehat{\lambda}_n\ge0$ of the empirical kernel matrix H^{∞}, which are computable from the data. Recall the definition of the optimal stopping time k^* as in (4.2). The following lemma establishes the L_2 estimation results for $\widehat{f}_{k^*}$ for kernels with polynomial eigendecay.

Lemma D.7 (Corollary 1 in [26])**.** Suppose that variables $\{\boldsymbol{x}_i\}_{i=1}^n$ are sampled i.i.d. and the kernel class $\mathcal{N}$ satisfies the polynomial eigenvalue decay $\lambda_j \lesssim j^{-2\nu}$ for some $\nu > 1/2$. Then there is a universal constant C such that

$$\mathbb{E}\left\|\widehat{f}_{k^*} - f^*\right\|_2^2 \le C\left(\frac{\sigma^2}{n}\right)^{\frac{2\nu}{2\nu+1}}.$$

Moreover, if $\lambda_j \asymp j^{-2\nu}$ for all $j = 1, 2, \ldots$, then for all iterations $k = 1, 2, \ldots$,

$$\mathbb{E}\left\|\widehat{f}_{k^*} - f^*\right\|_2^2 \ge \frac{\sigma^2}{4}\min\big\{1,\ \frac{(\alpha_k)^{\frac{1}{2\nu}}}{n}\big\}.$$

By Lemma 3.1, apply Lemma D.7 with $2\nu = d/(d-1)$ and the running sum of the step sizes $\alpha_k = k\eta$ gives the convergence rate.

Moreover, if $k \to \infty$, i.e., interpolation of training data, the lower bound result in Lemma D.7 implies $\mathbb{E}\left\|f_{\widehat{T}} - f^*\right\|_2^2 \gtrsim \sigma^2$ that doesn't converge to 0.

E Proofs of main theorems in Section 5

E.1 Proof of Theorem 5.1

Consider event

$$A_{ir} = \{\exists \boldsymbol{w} \in \mathbb{R}^d : \left\|\boldsymbol{w} - (1-\eta_2\mu)^k \boldsymbol{w}_r(0)\right\|_2 \le R, \mathbb{I}\{\boldsymbol{x}_i^\top \boldsymbol{w}_r(0) \ge 0\} \ne \mathbb{I}\{\boldsymbol{x}_i^\top \boldsymbol{w} \ge 0\}\},$$

where R will be determined later. Set $S_i = \{r \in [m] : \mathbb{I}\{A_{ir}\} = 0\}$ and $S_i^\perp = [m]\backslash S_i$. Then A_{ir} happens if and only if $|\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i| < R/(1-\eta_2\mu)^k$. By concentration inequality of Gaussian, we have $\mathbb{P}(A_{ir}) = \mathbb{P}(|\boldsymbol{w}_r(0)^\top \boldsymbol{x}_i| < R/(1-\eta_2\mu)^k \le \frac{2R}{\sqrt{2\pi}\tau(1-\eta_2\mu)^k}$. Thus, it follows the union bound inequality that with probability at least $1-\delta$ we have

$$\sum_{i=1}^n |S_i^\perp| \le \frac{CmnR}{\delta(1-\eta_2\mu)^k}, \tag{E.1}$$

where C is a positive constant.

Let $\boldsymbol{u}_D(l) = (u_{D,1}(l), ..., u_{D,n}(l))^\top \in \mathbb{R}^n$ be the predictions on the points $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ using the modified GD at the k-th iteration. We first study the difference between two predictions $\boldsymbol{u}_D(l+1)$ and $\boldsymbol{u}_D(l)$. For any $i \in [n]$, we have

$$\begin{aligned} u_{D,i}(l+1) - (1-\eta_2\mu)u_{D,i}(l) =& \frac{1}{\sqrt{m}}\sum_{r=1}^m a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\ =& \frac{1}{\sqrt{m}}\sum_{r\in S_i^\perp} a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\ &+ \frac{1}{\sqrt{m}}\sum_{r\in S_i} a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\ =& I_{1,i}(l) + I_{2,i}(l). \end{aligned} \tag{E.2}$$

The first term $I_{1,i}(l)$ can be bounded by

$$\begin{aligned}
I_{1,i}(l) =& \frac{1}{\sqrt{m}} \sum_{r \in S_i^\perp} a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\
\leq& \frac{1}{\sqrt{m}} \sum_{r \in S_i^\perp} \left|(\boldsymbol{w}_{D,r}(l+1) - (1-\eta_2\mu)\boldsymbol{w}_{D,r}(l))^\top \boldsymbol{x}_i\right| \\
\leq& \frac{1}{\sqrt{m}} \sum_{r \in S_i^\perp} \left\|\boldsymbol{w}_{D,r}(l+1) - (1-\eta_2\mu)\boldsymbol{w}_{D,r}(l)\right\|_2 \\
=& \frac{1}{\sqrt{m}} \sum_{r \in S_i^\perp} \left\| \frac{\eta_1}{\sqrt{m}} a_r \sum_{j=1}^n (u_{D,j}(l) - y_j)\mathbb{I}_{r,j}(l)\boldsymbol{x}_j \right\|_2 \\
\leq& \frac{\eta_1}{m} \sum_{r \in S_i^\perp} \sum_{j=1}^n |u_{D,j}(l) - y_j| \\
\leq& \frac{\eta_1 \sqrt{n}|S_i^\perp|}{m} \left\|\boldsymbol{u}_D(l) - \boldsymbol{y}\right\|_2. \qquad \text{(E.3)}
\end{aligned}$$

In (E.3), the second and the last inequalities are by the Cauchy-Schwarz inequality. The second term $I_{2,i}(l)$ can be bounded by

$$\begin{aligned}
I_{2,i}(l) =& \frac{1}{\sqrt{m}} \sum_{r \in S_i} a_r(\sigma(\boldsymbol{w}_{D,r}(l+1)^\top \boldsymbol{x}_i) - (1-\eta_2\mu)\sigma(\boldsymbol{w}_{D,r}(l)^\top \boldsymbol{x}_i)) \\
=& \frac{1}{\sqrt{m}} \sum_{r \in S_i} a_r \mathbb{I}_{r,i}(l)(\boldsymbol{w}_{D,r}(l+1) - (1-\eta_2\mu)\boldsymbol{w}_{D,r}(l))^\top \boldsymbol{x}_i \\
=& -\frac{1}{\sqrt{m}} \sum_{r \in S_i} a_r \mathbb{I}_{r,i}(l) \left(\frac{\eta_1}{\sqrt{m}} a_r \sum_{j=1}^n (u_{D,j}(l) - y_j)\mathbb{I}_{r,j}(l)\boldsymbol{x}_j \right)^\top \boldsymbol{x}_i \\
=& -\frac{\eta_1}{m} \sum_{j=1}^n (u_{D,j}(l) - y_j)\boldsymbol{x}_j^\top \boldsymbol{x}_i \sum_{r \in S_i} \mathbb{I}_{r,i}(l)\mathbb{I}_{r,j}(l) \\
=& -\eta_1 \sum_{j=1}^n (u_{D,j}(l) - y_j)\boldsymbol{H}_{ij}(l) + I_{3,i}(l), \qquad \text{(E.4)}
\end{aligned}$$

where

$$I_{3,i}(l) = \frac{\eta_1}{m} \sum_{j=1}^n (u_{D,j}(l) - y_j)\boldsymbol{x}_j^\top \boldsymbol{x}_i \sum_{r \in S_i^\perp} \mathbb{I}_{r,i}(l)\mathbb{I}_{r,j}(l).$$

The term $I_{3,i}(l)$ in (E.4) can be bounded by

$$\begin{aligned}
|I_{3,i}(l)| \leq& \left| \frac{\eta_1}{m} \sum_{j=1}^n (u_{D,j}(l) - y_j)\boldsymbol{x}_j^\top \boldsymbol{x}_i \sum_{r \in S_i^\perp} \mathbb{I}_{r,i}(l)\mathbb{I}_{r,j}(l) \right| \\
\leq& \frac{\eta_1}{m}|S_i^\perp| \sum_{j=1}^n |u_{D,j}(l) - y_j| \\
\leq& \frac{\eta_1 \sqrt{n}|S_i^\perp|}{m} \left\|\boldsymbol{u}_D(l) - \boldsymbol{y}\right\|_2. \qquad \text{(E.5)}
\end{aligned}$$

Plugging (E.3) and (E.4) into (E.2), we have

$$u_{D,i}(l+1) - (1-\eta_2\mu)u_{D,i}(l) = -\eta_1 \sum_{j=1}^n (u_{D,j}(l) - y_j)\boldsymbol{H}_{ij}(l) + I_{1,i}(l) + I_{3,i}(l),$$

which leads to

$$\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l) = -\eta_1 \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y}) + \boldsymbol{I}(l), \tag{E.6}$$

where $\boldsymbol{I}(l) = (I_{1,1}(l) + I_{3,1}(l), ..., I_{1,n}(l) + I_{3,n}(l))^\top$. By the triangle inequality, we have

$$\|\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l)\|_2 \leq \|\eta_1 \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y})\|_2 + \|\boldsymbol{I}(l)\|_2 . \tag{E.7}$$

By (E.1), (E.3), and (E.5), the term $\|\boldsymbol{I}(l)\|_2$ in (E.7) can be bounded by

$$\begin{aligned}\|\boldsymbol{I}(l)\|_2 \leq & \sum_{i=1}^n |I_{3,i}(l)| + |I_{1,i}(l)| \leq \sum_{i=1}^n \frac{2\eta_1\sqrt{n}|S_i^\perp|}{m} \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2 \\ \leq & \frac{2\eta_1\sqrt{n}}{m}\frac{CmnR}{\delta(1-\eta_2\mu)^k} \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2 = \frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k} \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2 . \end{aligned} \tag{E.8}$$

Gershgorin's theorem [44] implies

$$\lambda_{\max}(H(l)) \leq \max_j \sum_{i=1}^n H_{ij}(l) \leq n.$$

Therefore, the term $\|\eta_1 \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y})\|_2$ in (E.7) can be bounded by

$$\|\eta_1 \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y})\|_2 \leq \eta_1 \lambda_{\max}(H(l)) \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2 \leq \eta_1 n \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2 . \tag{E.9}$$

By (E.7) and (E.8), $\|\boldsymbol{y} - \boldsymbol{u}_D(l+1)\|_2$ can be bounded by

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{u}_D(l+1)\|_2^2 = & \|\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l)\|_2^2 - 2(\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l))^\top (\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l)) \\ & + \|\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l)\|_2^2 \\ = & \|\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l)\|_2^2 + 2\eta_1(\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l))^\top \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y}) \\ & - 2\eta_1(\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l))^\top \boldsymbol{I}(l) + \|\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l)\|_2^2 \\ = & T_1 + T_2 + T_3 + T_4. \end{aligned} \tag{E.10}$$

The first term T_1 can be bounded by

$$\begin{aligned} T_1 = & \|\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l)\|_2^2 \\ = & \eta_2^2\mu^2 \|\boldsymbol{y}\|_2^2 + (1-\eta_2\mu)^2 \|\boldsymbol{y} - \boldsymbol{u}_D(l)\|_2^2 + 2\eta_2\mu(1-\eta_2\mu)\boldsymbol{y}^\top(\boldsymbol{y} - \boldsymbol{u}_D(l)) \\ \leq & (\eta_2^2\mu^2 + \eta_2\mu)\|\boldsymbol{y}\|_2^2 + (1+\eta_2\mu)(1-\eta_2\mu)^2 \|\boldsymbol{y} - \boldsymbol{u}_D(l)\|_2^2 . \end{aligned} \tag{E.11}$$

The second term T_2 can be bounded by

$$\begin{aligned} T_2 = & 2\eta_1(\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l))^\top \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y}) \\ = & 2\eta_1(1-\eta_2\mu)(\boldsymbol{y} - \boldsymbol{u}_D(l))^\top \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y}) + 2\eta_1\eta_2\mu \boldsymbol{y}^\top \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y}) \\ = & -2\eta_1(1-\eta_2\mu)(\boldsymbol{y} - \boldsymbol{u}_D(l))^\top \boldsymbol{H}(l)(\boldsymbol{y} - \boldsymbol{u}_D(l)) + 2\eta_1\eta_2\mu \boldsymbol{y}^\top \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y}) \\ \leq & 4\eta_1\eta_2\mu n \|\boldsymbol{y}\|_2^2 + 4\eta_1\eta_2\mu n \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2^2 . \end{aligned} \tag{E.12}$$

Using (E.8), the third term T_3 can be bounded by

$$\begin{aligned} T_3 = & -2\eta_1(\boldsymbol{y} - (1-\eta_2\mu)\boldsymbol{u}_D(l))^\top \boldsymbol{I}(l) \\ = & -2\eta_1(1-\eta_2\mu)(\boldsymbol{y} - \boldsymbol{u}_D(l))^\top \boldsymbol{I}(l) + 2\eta_1\eta_2\mu \boldsymbol{y}^\top \boldsymbol{I}(l) \\ \leq & 2\eta_1(1-\eta_2\mu)\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k} \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2 + 4\eta_1\eta_2\mu \|\boldsymbol{y}\|_2^2 + 4\eta_1\eta_2\mu \|\boldsymbol{I}(l)\|_2^2 \\ \leq & 2\eta_1(1-\eta_2\mu)\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k} \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2^2 + 4\eta_1\eta_2\mu \|\boldsymbol{y}\|_2^2 + 4\eta_1\eta_2\mu \left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2 \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2^2 . \end{aligned} \tag{E.13}$$

The fourth term T_4 can be bounded by

$$\begin{aligned} T_4 = & \|\boldsymbol{u}_D(l+1) - (1-\eta_2\mu)\boldsymbol{u}_D(l)\|_2^2 \\ \leq & 2\|\eta_1 \boldsymbol{H}(l)(\boldsymbol{u}_D(l) - \boldsymbol{y})\|_2^2 + 2\|\boldsymbol{I}(l)\|_2^2 \\ \leq & 2\eta_1^2 n^2 \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2^2 + 2\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2 \|\boldsymbol{u}_D(l) - \boldsymbol{y}\|_2^2 . \end{aligned} \tag{E.14}$$

Plugging (E.11) - (E.14) into (E.10), we have

$$\begin{aligned}
&\|\boldsymbol{y}-\boldsymbol{u}_D(l+1)\|_2^2\\
\leq&(\eta_2^2\mu^2+\eta_2\mu)\|\boldsymbol{y}\|_2^2+(1+\eta_2\mu)(1-\eta_2\mu)^2\|\boldsymbol{y}-\boldsymbol{u}_D(l)\|_2^2+4\eta_1\eta_2\mu n\|\boldsymbol{y}\|_2^2+4\eta_1\eta_2\mu n\|\boldsymbol{u}_D(l)-\boldsymbol{y}\|_2^2\\
&+2\eta_1(1-\eta_2\mu)\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\|\boldsymbol{u}_D(l)-\boldsymbol{y}\|_2^2+4\eta_1\eta_2\mu\|\boldsymbol{y}\|_2^2+4\eta_1\eta_2\mu\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2\|\boldsymbol{u}_D(l)-\boldsymbol{y}\|_2^2\\
&+2\eta_1^2n^2\|\boldsymbol{u}_D(l)-\boldsymbol{y}\|_2^2+2\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2\|\boldsymbol{u}_D(l)-\boldsymbol{y}\|_2^2\\
=&a_1\|\boldsymbol{y}\|_2^2+a_2\|\boldsymbol{u}_D(l)-\boldsymbol{y}\|_2^2, \qquad \text{(E.15)}
\end{aligned}$$

where

$$\begin{aligned}
a_1=&(\eta_2^2\mu^2+\eta_2\mu)+4\eta_1\eta_2\mu n+4\eta_1\eta_2\mu\leq 2\eta_2\mu+8\eta_1\eta_2\mu n,\\
a_2=&(1+\eta_2\mu)(1-\eta_2\mu)^2+4\eta_1\eta_2\mu n+2\eta_1(1-\eta_2\mu)\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\\
&+4\eta_1\eta_2\mu\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2+2\eta_1^2n^2+2\left(\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}\right)^2\\
\leq&1-\left(\eta_2\mu-4\eta_1\eta_2\mu n-2\eta_1\frac{2C\eta_1 n^{3/2}R}{\delta(1-\eta_2\mu)^k}-2\eta_1^2n^2\right)\\
=&1-\nu_0.
\end{aligned}$$

By the conditions imposed on η_1,η_2,μ,m, the dominating terms in a_1 and ν_0 are both $\eta_2\mu$. Thus $a_1=o(1/n)$, $\nu_0=o(1/n)$ and $a_1/\nu_0=O(1)$. Using (E.15) iteratively, we have

$$\begin{aligned}
\|\boldsymbol{y}-\boldsymbol{u}_D(l+1)\|_2^2\leq&a_1\|\boldsymbol{y}\|_2^2+a_2\|\boldsymbol{u}_D(l)-\boldsymbol{y}\|_2^2\\
\leq&...\leq\sum_{i=0}^{l}(1-\nu_0)^i(a_1\|\boldsymbol{y}\|_2^2)+(1-\nu_0)^{l+1}\|\boldsymbol{y}-\boldsymbol{u}_D(0)\|_2^2 \qquad \text{(E.16)}\\
\leq&\frac{a_1\|\boldsymbol{y}\|_2^2}{\nu_0}+(1-\nu_0)^{l+1}\|\boldsymbol{y}-\boldsymbol{u}_D(0)\|_2^2. \qquad \text{(E.17)}
\end{aligned}$$

By the modified GD rule, we have

$$\boldsymbol{w}_{D,r}(l+1)-(1-\eta_2\mu)\boldsymbol{w}_{D,r}(l)=-\frac{\eta_1}{\sqrt{m}}a_r\sum_{j=1}^{n}(u_{D,j}(l)-y_j)\mathbb{I}_{r,j}(l)\boldsymbol{x}_j,$$

which implies

$$\|\boldsymbol{w}_{D,r}(l+1)-(1-\eta_2\mu)\boldsymbol{w}_{D,r}(l)\|_2\leq\frac{\eta_1\sqrt{n}}{\sqrt{m}}\|\boldsymbol{u}_D(l)-\boldsymbol{y}\|_2\leq\frac{C\eta_1 n}{\sqrt{m}} \qquad \text{(E.18)}$$

for some constant C. Using (E.18) iteratively yields

$$\begin{aligned}
&\left\|\boldsymbol{w}_{D,r}(l+1)-(1-\eta_2\mu)^{l+1}\boldsymbol{w}_{D,r}(0)\right\|_2\\
\leq&\|\boldsymbol{w}_{D,r}(l+1)-(1-\eta_2\mu)\boldsymbol{w}_{D,r}(l)\|_2+\left\|(1-\eta_2\mu)\boldsymbol{w}_{D,r}(0)-(1-\eta_2\mu)^{l+1}\boldsymbol{w}_{D,r}(l)\right\|_2\\
\leq&\frac{C\eta_1 n}{\sqrt{m}}+(1-\eta_2\mu)\left\|\boldsymbol{w}_{D,r}(l)-(1-\eta_2\mu)^l\boldsymbol{w}_{D,r}(0)\right\|_2\\
\leq&...\leq\sum_{i=0}^{l}(1-\eta_2\mu)^i\frac{C\eta_1 n}{\sqrt{m}}\leq\frac{C\eta_1 n}{\eta_2\mu\sqrt{m}}. \qquad \text{(E.19)}
\end{aligned}$$

By similar approach as in the proof of Lemma C.2 of [7], we can show that with probability at least $1-\delta$ with respect to random initialization,

$$\|\boldsymbol{Z}(l)-\boldsymbol{Z}(0)\|_F^2\leq\frac{2nR}{\sqrt{2\pi}\tau\delta(1-\eta_2\mu)^k}+\frac{n}{m}=O\left(\frac{\eta_1 n^2}{(1-\eta_2\mu)^k\eta_2\mu\sqrt{m}\delta^{3/2}\tau}\right),\forall l\in[k],$$

and

$$\|\boldsymbol{H}(l)-\boldsymbol{H}(0)\|_F \le \frac{4n^2R}{\sqrt{2\pi}\tau}+\frac{2n^2\delta}{m}=O\left(\frac{\eta_1 n^3}{(1-\eta_2\mu)^k\eta_2\mu\sqrt{m}\delta^{3/2}\tau}\right), \forall l\in[k].$$

By Lemma C.3 of [7], we have with probability at least $1-\delta$ with respect to random initialization,

$$\|\boldsymbol{H}(0)-\boldsymbol{H}^\infty\|_F=O\left(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}\right). \tag{E.20}$$

By (E.6), we have

$$\begin{aligned}\boldsymbol{u}_D(l+1)-(1-\eta_2\mu)\boldsymbol{u}_D(l)&=-\eta_1\boldsymbol{H}(l)(\boldsymbol{u}_D(l)-\boldsymbol{y})+\boldsymbol{I}(l)\\&=-\eta_1\boldsymbol{H}^\infty(\boldsymbol{u}_D(l)-\boldsymbol{y})+\boldsymbol{I}(l)-\eta_1(\boldsymbol{H}(l)-\boldsymbol{H}^\infty)(\boldsymbol{u}_D(l)-\boldsymbol{y}),\end{aligned}$$

which yields

$$\boldsymbol{u}_D(l+1)-B=\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)(\boldsymbol{u}_D(l)-B)+\boldsymbol{I}(l)-\eta_1(\boldsymbol{H}(l)-\boldsymbol{H}^\infty)(\boldsymbol{u}_D(l)-\boldsymbol{y}), \tag{E.21}$$

where

$$B=(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\eta_1\boldsymbol{H}^\infty\boldsymbol{y}=\eta_1\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}. \tag{E.22}$$

Iteratively using (E.21), we have

$$\begin{aligned}\boldsymbol{u}_D(l+1)-B=&\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l+1}(\boldsymbol{u}_D(0)-B)\\&+\sum_{i=0}^{l}\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^i\left(\boldsymbol{I}(l-i)-\eta_1(\boldsymbol{H}(l-i)-\boldsymbol{H}^\infty)(\boldsymbol{u}_D(l-i)-\boldsymbol{y})\right)\\=&\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l+1}(\boldsymbol{u}_D(0)-B)+e_l,\end{aligned} \tag{E.23}$$

where

$$e_l=\sum_{i=0}^{l}\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^i\left(\boldsymbol{I}(l-i)-\eta_1(\boldsymbol{H}(l-i)-\boldsymbol{H}^\infty)(\boldsymbol{u}_D(l-i)-\boldsymbol{y})\right). \tag{E.24}$$

The term e_l can be bounded by

$$\begin{aligned}\|e_l\|_2=&\left\|\sum_{i=0}^{l}\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^i\left(\boldsymbol{I}(l-i)-\eta_1(\boldsymbol{H}(l-i)-\boldsymbol{H}^\infty)(\boldsymbol{u}_D(l-i)-\boldsymbol{y})\right)\right\|_2\\\le&\sum_{i=0}^{l}\|(1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\|_2^i\left(\|\boldsymbol{I}(l-i)\|_2+\eta_1\|\boldsymbol{H}(l-i)-\boldsymbol{H}^\infty\|_2\|\boldsymbol{u}_D(l-i)-\boldsymbol{y}\|_2\right)\\\le&\sum_{i=0}^{l}(1-\eta_2\mu)^iO\bigg(\frac{2C\eta_1^2n^{5/2}}{\eta_2\mu\sqrt{m}\delta^{3/2}(1-\eta_2\mu)^k}+\frac{\eta_1^2n^{7/2}}{(1-\eta_2\mu)^k\eta_2\mu\sqrt{m}\delta^2\tau}\bigg)\\=&O\bigg(\frac{\eta_1^2n^{7/2}}{\eta_2^2\mu^2\sqrt{m}\delta^2(1-\eta_2\mu)^k\tau}\bigg).\end{aligned} \tag{E.25}$$

By (E.23) and taking $l=k-1$, with probability at least $1-\delta$ with respect to the random initialization, the difference $\boldsymbol{u}_D(k)-B$ can be bounded by

$$\begin{aligned}\|\boldsymbol{u}_D(k)-B\|_2\le&\left\|\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^k(\boldsymbol{u}_D(0)-B)\right\|_2+\|e_k\|_2\\=&O\left(\sqrt{n}(1-\eta_2\mu-\eta_1\lambda_0)^k+\frac{n^{7/2}}{\mu^2\sqrt{m}\delta^2(1-\eta_2\mu)^k\tau}\right)\\=&O\left(\sqrt{n}(1-\eta_2\mu)^k+\frac{n^{7/2}}{\mu^2\sqrt{m}\delta^2(1-\eta_2\mu)^k\tau}\right).\end{aligned}$$

This implies that

$$\|\boldsymbol{u}_D(k) - B\|_2 = O_{\mathbb{P}}\left(\sqrt{n}(1-\eta_2\mu)^k + \frac{n^{7/2}}{\mu^2\sqrt{m}(1-\eta_2\mu)^k\tau}\right).$$

By choosing $m = \mathrm{poly}(n, 1/\tau, 1/\lambda_0)$ such that $\frac{n^{7/2}}{\mu^2\sqrt{m}(1-\eta_2\mu)^k\tau} \le \sqrt{n}(1-\eta_2\mu)^k$, we finish the proof of (5.3).

Now consider $\mathrm{vec}(\boldsymbol{W}_D(l+1))$. Direct calculation shows that

$$\begin{aligned}\mathrm{vec}(\boldsymbol{W}_D(l+1)) =&(1-\eta_2\mu)\mathrm{vec}(\boldsymbol{W}_D(l)) - \eta_1\boldsymbol{Z}(l)(\boldsymbol{u}_D(l)-\boldsymbol{y})\\ =&(1-\eta_2\mu)\mathrm{vec}(\boldsymbol{W}_D(l)) - \eta_1\boldsymbol{Z}(0)(\boldsymbol{u}_D(l)-\boldsymbol{y}) - \eta_1(\boldsymbol{Z}(l)-\boldsymbol{Z}(0))(\boldsymbol{u}_D(l)-\boldsymbol{y})\\ =&(1-\eta_2\mu)^{l+1}\mathrm{vec}(\boldsymbol{W}_D(0)) - \eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i(\boldsymbol{u}_D(l-i)-\boldsymbol{y})\\ &-\sum_{i=0}^{l}(1-\eta_2\mu)^i\eta_1(\boldsymbol{Z}(l)-\boldsymbol{Z}(0))(\boldsymbol{u}_D(l)-\boldsymbol{y}).\end{aligned} \tag{E.26}$$

Plugging

$$\boldsymbol{u}_D(l+1) = \left((1-\eta_2\mu)I - \eta_1\boldsymbol{H}^\infty\right)^{l+1}(\boldsymbol{u}_D(0)-B) + e_l + B$$

into (E.26), we have

$$\begin{aligned}&\mathrm{vec}(\boldsymbol{W}_D(l+1)) - (1-\eta_2\mu)^{l+1}\mathrm{vec}(\boldsymbol{W}_D(0))\\ =&-\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l-i}(\boldsymbol{u}_D(0)-B)\\ &-\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i(e_{l-i-1}+B-\boldsymbol{y}) - \sum_{i=0}^{l}(1-\eta_2\mu)^i\eta_1(\boldsymbol{Z}(l)-\boldsymbol{Z}(0))(\boldsymbol{u}_D(l)-\boldsymbol{y})\\ =&\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l-i}\eta_1\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}\\ &-\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l-i}\boldsymbol{u}_D(0)\\ &-\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i e_{l-i-1} - \eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i(B-\boldsymbol{y})\\ &-\sum_{i=0}^{l}(1-\eta_2\mu)^i\eta_1(\boldsymbol{Z}(l)-\boldsymbol{Z}(0))(\boldsymbol{u}_D(l)-\boldsymbol{y})\\ =&E_1 - E_2 + E_3 - T_5 - E_4.\end{aligned} \tag{E.27}$$

Let

$$\begin{aligned}\boldsymbol{T}_l =&\sum_{i=0}^{l}(1-\eta_2\mu)^i\left((1-\eta_2\mu)I-\eta_1\boldsymbol{H}^\infty\right)^{l-i}\\ =&(1-\eta_2\mu)^l\sum_{i=0}^{l}\left(I-\frac{\eta_1}{(1-\eta_2\mu)}\boldsymbol{H}^\infty\right)^i\end{aligned} \tag{E.28}$$

and

$$\boldsymbol{a}_1 = \eta_1\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}. \tag{E.29}$$

The first term E_1 can be bounded by

$$\begin{aligned}\|E_1\|_2^2 =& \|\eta_1 \boldsymbol{Z}(0)\boldsymbol{T}_l \boldsymbol{a}_1\|_2^2\\ =&\eta_1^2 \boldsymbol{a}_1^\top \boldsymbol{T}_l \boldsymbol{Z}(0)^\top \boldsymbol{Z}(0)\boldsymbol{T}_l \boldsymbol{a}_1\\ =&\eta_1^2 \boldsymbol{a}_1^\top \boldsymbol{T}_l \boldsymbol{H}^\infty \boldsymbol{T}_l \boldsymbol{a}_1 + \eta_1^2 \boldsymbol{a}_1^\top \boldsymbol{T}_l (\boldsymbol{H}(0) - \boldsymbol{H}^\infty)\boldsymbol{T}_l \boldsymbol{a}_1\\ =&\eta_1^2 \boldsymbol{a}_1^\top \boldsymbol{T}_l \boldsymbol{H}^\infty \boldsymbol{T}_l \boldsymbol{a}_1 + \eta_1^2 O\left(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}\right) \boldsymbol{a}_1^\top \boldsymbol{T}_l^2 \boldsymbol{a}_1. \end{aligned} \tag{E.30}$$

By (E.28), we have

$$\boldsymbol{T}_l =(1-\eta_2\mu)^l \sum_{j=1}^n \frac{1-(1-\frac{\eta_1}{(1-\eta_2\mu)}\lambda_j)^{l+1}}{\frac{\eta_1}{(1-\eta_2\mu)}\lambda_j} \boldsymbol{v}_j \boldsymbol{v}_j^\top \preceq \frac{(1-\eta_2\mu)^l}{\eta_1\lambda_0}\boldsymbol{I},$$

and

$$\boldsymbol{T}_l \boldsymbol{H}^\infty \boldsymbol{T}_l =(1-\eta_2\mu)^{2l} \sum_{j=1}^n \left(\frac{1-(1-\frac{\eta_1}{(1-\eta_2\mu)}\lambda_j)^{2l+2}}{\frac{\eta_1}{(1-\eta_2\mu)}\lambda_j}\right)^2 \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top \preceq \frac{(1-\eta_2\mu)^{l+1}}{\eta_1^2}(\boldsymbol{H}^\infty)^{-1}.$$

Therefore,

$$\begin{aligned}\eta_1^2 \boldsymbol{a}_1^\top \boldsymbol{T}_l \boldsymbol{H}^\infty \boldsymbol{T}_l \boldsymbol{a}_1 \le& (1-\eta_2\mu)^{2l+2} \boldsymbol{a}_1^\top (\boldsymbol{H}^\infty)^{-1} \boldsymbol{a}_1,\\ \eta_1^2 O\left(\frac{n\sqrt{\log(n/\delta)}}{\sqrt{m}}\right) \boldsymbol{a}_1^\top \boldsymbol{T}_l^2 \boldsymbol{a}_1 \le& O\left(\frac{n^2(1-\eta_2\mu)^{2l}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right).\end{aligned}$$

Together with (E.30), we have

$$\|E_1\|_2^2 = (1-\eta_2\mu)^{2l+2} \boldsymbol{a}_1^\top (\boldsymbol{H}^\infty)^{-1} \boldsymbol{a}_1 + O\left(\frac{n^2(1-\eta_2\mu)^{2l}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right). \tag{E.31}$$

By similar approach, the second term E_2 can be bounded by

$$\begin{aligned}\|E_2\|_2^2 =& \left\|\eta_1 \boldsymbol{Z}(0) \sum_{i=0}^l (1-\eta_2\mu)^i \left((1-\eta_2\mu)I - \eta_1 \boldsymbol{H}^\infty\right)^{l-i} \boldsymbol{u}_D(0)\right\|_2^2\\ =&\eta_1^2 \boldsymbol{u}_D(0)^\top \boldsymbol{T}_1(l) \boldsymbol{Z}(0)^\top \boldsymbol{Z}(0) \boldsymbol{T}_1(l) \boldsymbol{u}_D(0)\\ =&\eta_1^2 \boldsymbol{u}_D(0)^\top \boldsymbol{T}_1(l) \boldsymbol{H}^\infty \boldsymbol{T}_1(l) \boldsymbol{u}_D(0) + \eta_1^2 \boldsymbol{u}_D(0)^\top \boldsymbol{T}_1(l)(\boldsymbol{H}(0)-\boldsymbol{H}^\infty)\boldsymbol{T}_1(l)\boldsymbol{u}_D(0)\\ =&(1-\eta_2\mu)^{2l+2} \boldsymbol{u}_D(0)^\top (\boldsymbol{H}^\infty)^{-1} \boldsymbol{u}_D(0) + O\left(\frac{n^2(1-\eta_2\mu)^{2l}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right).\end{aligned} \tag{E.32}$$

By (E.25), the third term E_3 can be bounded by

$$\begin{aligned}\|E_3\|_2^2 =& \left\|\eta_1 \boldsymbol{Z}(0) \sum_{i=0}^l (1-\eta_2\mu)^i e_{l-i-1}\right\|_2^2\\ =&\eta_1^2 \left(\sum_{i=0}^l (1-\eta_2\mu)^i e_{l-i-1}\right)^\top \boldsymbol{H}(0) \left(\sum_{i=0}^l (1-\eta_2\mu)^i e_{l-i-1}\right)\\ =&O\Big(\frac{\eta_1^6 n^8}{\eta_2^6 \mu^6 m \delta^4 (1-\eta_2\mu)^{2k}\tau^2}\Big).\end{aligned} \tag{E.33}$$

The fourth term E_4 can be bounded by

$$\begin{aligned}\|E_4\|_2^2 =& \left\|\sum_{i=0}^l (1-\eta_2\mu)^i \eta_1 (\boldsymbol{Z}(l) - \boldsymbol{Z}(0))(\boldsymbol{u}_D(l) - \boldsymbol{y})\right\|_2^2\\ =&O\left(\frac{\eta_1^3 n^3}{(1-\eta_2\mu)^k \eta_2^3 \mu^3 \sqrt{m}\delta^{3/2}\tau}\right).\end{aligned} \tag{E.34}$$

Note that

$$\begin{aligned}B-\boldsymbol{y}=&\eta_1\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}-\boldsymbol{y}\\=&(\eta_1\boldsymbol{H}^\infty-\eta_2\mu I-\eta_1\boldsymbol{H}^\infty)(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}\\=&-\eta_2\mu(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}.\end{aligned}$$

Therefore, the remaining term T_5 can be bounded by

$$\begin{aligned}\|T_5\|_2^2=&\left\|\eta_1\boldsymbol{Z}(0)\sum_{i=0}^{l}(1-\eta_2\mu)^i(B-\boldsymbol{y})\right\|_2^2\\\leq&\eta_1^2\boldsymbol{y}^\top(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}\\\leq&\boldsymbol{y}^\top(\eta_2\mu/\eta_1 I+\boldsymbol{H}^\infty)^{-1}\boldsymbol{H}^\infty(\eta_2\mu/\eta_1 I+\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}.\end{aligned}$$

By the assumption that $\eta_2\asymp\eta_1$, the term T_5 can be further bounded by

$$\|T_5\|_2^2\leq\boldsymbol{y}^\top(C\mu I+\boldsymbol{H}^\infty)^{-1}\boldsymbol{H}^\infty(C\mu I+\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}. \tag{E.35}$$

The right-hand side of (E.35) is $\left\|\widehat{f}\right\|_{\mathcal{N}}^2$, where $\widehat{f}$ is defined in (3.4). The term $\left\|\widehat{f}\right\|_{\mathcal{N}}^2$ can be bounded by some constant as in Theorem 3.2. This also implies

$$\boldsymbol{a}_1^\top(\boldsymbol{H}^\infty)^{-1}\boldsymbol{a}_1=\eta_1^2\boldsymbol{y}^\top(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{H}^\infty(\eta_2\mu I+\eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}=O(1). \tag{E.36}$$

Note also that

$$\boldsymbol{u}_D(0)^\top(\boldsymbol{H}^\infty)^{-1}\boldsymbol{u}_D(0)=O\left(\frac{n\tau^2}{\lambda_0}\right). \tag{E.37}$$

By the assumptions of Theorem 5.1, plugging (E.30)-(E.37) into (E.27), and taking the iteration number at k, we can conclude that

$$\begin{aligned}&\left\|\mathrm{vec}(\boldsymbol{W}_D(k))-(1-\eta_2\mu)^k\mathrm{vec}(\boldsymbol{W}_D(0))\right\|_2^2\\=&O((1-\eta_2\mu)^{2k})+O\left(\frac{n^2(1-\eta_2\mu)^{2k-2}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right)\\&+O\left(\frac{n\tau^2}{\lambda_0}(1-\eta_2\mu)^{2k}\right)+O\left(\frac{n^2(1-\eta_2\mu)^{2k-2}\sqrt{\log(n/\delta)}}{\sqrt{m}\lambda_0^2}\right)\\&+O\biggl(\frac{n^8}{\mu^6m\delta^4(1-\eta_2\mu)^{2k}\tau^2}\biggr)+O\left(\frac{n^3}{(1-\eta_2\mu)^k\mu^3\sqrt{m}\delta^{3/2}\tau}\right)+O(1)\\=&O(1),\end{aligned} \tag{E.38}$$

where the last equality is because we can select some polynomials such that all the terms in (E.38) except the $O(1)$ term converge to zero, and $\exp(-2\eta_2\mu k)\leq(1-\eta_2\mu)^k\leq\exp(-\eta_2\mu k)$ for sufficiently large n. This finishes the proof of (5.4) in Theorem 5.1.

E.2 Proof of Theorem 5.2

For notational simplification, we use $\widehat{f}_k=f_{\boldsymbol{W}(k),\boldsymbol{a}}$. Similar to the proof of Theorem 4.1, we define

$$\widetilde{f}_k(\boldsymbol{x})=\mathrm{vec}(\boldsymbol{W}_D(k))^\top\boldsymbol{z}_0(\boldsymbol{x}), \tag{E.39}$$

where $\boldsymbol{z}_0(\boldsymbol{x})=\boldsymbol{z}(\boldsymbol{x})|_{\boldsymbol{W}_D=\boldsymbol{W}_D(0)}$. Then we can write the following decomposition

$$\begin{aligned}\widehat{f}_k(\boldsymbol{x})-f^*(\boldsymbol{x})=&(\widehat{f}_k(\boldsymbol{x})-\widetilde{f}_k(\boldsymbol{x}))+(\widetilde{f}_k(\boldsymbol{x})-\widehat{f}(\boldsymbol{x}))+(\widehat{f}(\boldsymbol{x})-f^*(\boldsymbol{x}))\\=&\Delta_1(\boldsymbol{x})+\Delta_2(\boldsymbol{x})+\Delta_3(\boldsymbol{x}),\end{aligned} \tag{E.40}$$

where $\widehat{f}$ is as in (3.4). It follows from Theorem 3.2 that

$$\|\Delta_3\|_2^2=O_{\mathbb{P}}\left(n^{-\frac{d}{2d-1}}\right). \tag{E.41}$$

Next, we consider Δ_1. From (E.19), it can be seen that

$$\left\|\boldsymbol{w}_{D,r}(k) - (1-\eta_2\mu)^k \boldsymbol{w}_{D,r}(0)\right\|_2 \leq \frac{C\eta_1 n}{\eta_2\mu\sqrt{m}}. \tag{E.42}$$

Define event

$$B_{D,r}(\boldsymbol{x}) = \{|(1-\eta_2\mu)^k \boldsymbol{w}_{D,r}(0)^\top \boldsymbol{x}| \leq R_1\}, \forall r \in [m],$$

where $R_1 = \frac{C\eta_1 n}{\eta_2\mu\sqrt{m}}$. If $\mathbb{I}\{B_{D,r}(\boldsymbol{x})\} = 0$, then we have $\mathbb{I}_{r,k}(\boldsymbol{x}) = \mathbb{I}_{r,0}(\boldsymbol{x})$, where $\mathbb{I}_{r,k}(\boldsymbol{x}) = \mathbb{I}\{\boldsymbol{w}_{D,r}(k)^\top \boldsymbol{x} \geq 0\}$. Therefore, for any fixed $\boldsymbol{x}$,

$$\begin{aligned}
|\Delta_1(\boldsymbol{x})| &= |\widehat{f}_k(\boldsymbol{x}) - \widetilde{f}_k(\boldsymbol{x})| \\
&= \left|\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r(\mathbb{I}_{r,k}(\boldsymbol{x}) - \mathbb{I}_{r,0}(\boldsymbol{x}))\boldsymbol{w}_{D,r}(k)^\top \boldsymbol{x}\right| \\
&= \left|\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r \mathbb{I}\{B_{D,r}(\boldsymbol{x})\}(\mathbb{I}_{r,k}(\boldsymbol{x}) - \mathbb{I}_{r,0}(\boldsymbol{x}))\boldsymbol{w}_{D,r}(k)^\top \boldsymbol{x}\right| \\
&\leq \frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{I}\{B_{D,r}(\boldsymbol{x})\}|\boldsymbol{w}_{D,r}(k)^\top \boldsymbol{x}| \\
&\leq \frac{1}{\sqrt{m}}\sum_{r=1}^m \mathbb{I}\{B_{D,r}(\boldsymbol{x})\}\left(|(1-\eta_2\mu)^k \boldsymbol{w}_{D,r}(0)^\top \boldsymbol{x}| + |\boldsymbol{w}_{D,r}(k)^\top \boldsymbol{x} - (1-\eta_2\mu)^k \boldsymbol{w}_r(0)^\top \boldsymbol{x}|\right) \\
&\leq \frac{2R_1}{\sqrt{m}}\sum_{r=1}^m \mathbb{I}\{B_{D,r}(x)\}.
\end{aligned}$$

Note that $\|\boldsymbol{x}\|_2 = 1$, which implies that $\boldsymbol{w}_{D,r}(0)^\top \boldsymbol{x}$ is distributed as $N(0,\tau^2)$. Therefore, we have

$$\begin{aligned}
\mathbb{E}[\mathbb{I}\{B_{D,r}(x)\}] &= \mathbb{P}\left(|(1-\eta_2\mu)^k \boldsymbol{w}_{D,r}(0)^\top \boldsymbol{x}| \leq R_1\right) \\
&= \int_{-R_1/(1-\eta_2\mu)^k}^{R_1/(1-\eta_2\mu)^k} \frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{u^2}{2\tau^2}\right\} du \leq \frac{2R_1}{\sqrt{2\pi}(1-\eta_2\mu)^k\tau}.
\end{aligned}$$

By Markov's inequality, with probability at least $1-\delta$, we have

$$\sum_{r=1}^m \mathbb{I}\{B_{D,r}(x)\} \leq \frac{2mR_1}{\sqrt{2\pi}(1-\eta_2\mu)^k\tau\delta}.$$

Thus, we have with probability at least $1-\delta$,

$$\|\Delta_1\|_2 \leq \frac{2R_1}{\sqrt{m}}\left\|\sum_{r=1}^m \mathbb{I}\{B_{D,r}(\cdot)\}\right\|_2 \leq \frac{4\sqrt{m}R_1^2}{\sqrt{2\pi}(1-\eta_2\mu)^k\tau\delta} = O\left(\frac{n^2}{\sqrt{m}\lambda_0^2\delta^2(1-\eta_2\mu)^k\tau}\right),$$

which implies

$$\|\Delta_1\|_2 = O_{\mathbb{P}}\left(\frac{n^2}{\sqrt{m}\lambda_0^2(1-\eta_2\mu)^k\tau}\right). \tag{E.43}$$

Now we bound Δ_2. Note that Define $\boldsymbol{G}_k = \sum_{j=0}^{k-1}\eta(\boldsymbol{I} - \eta\boldsymbol{H}^\infty)^j$. Recalling that $\boldsymbol{y} = \boldsymbol{y}^* + \boldsymbol{\epsilon}$, for fixed $\boldsymbol{x}$, we have

$$\begin{aligned}
\Delta_2(\boldsymbol{x}) =& \widetilde{f}_k(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x}) \\
=& \boldsymbol{z}_0(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{W}_D(k)) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y} \\
=& \boldsymbol{z}_0(\boldsymbol{x})^\top E_1 - \boldsymbol{z}_0(\boldsymbol{x})^\top E_2 + \boldsymbol{z}_0(\boldsymbol{x})^\top E_3 - \boldsymbol{z}_0(\boldsymbol{x})^\top T_5 - \boldsymbol{z}_0(\boldsymbol{x})^\top E_4 \\
&+ (1-\eta_2\mu)^k \boldsymbol{z}_0(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{W}_D(0)) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y},
\end{aligned} \tag{E.44}$$

where E_1, E_2, E_3, T_5, E_4 are as in (E.27). Noting that $\|z_0(\boldsymbol{x})\|_2 = O_{\mathbb{P}}(1)$, we have that

$$|\boldsymbol{z}_0(\boldsymbol{x})^\top E_1|^2 \le \|\boldsymbol{z}_0(\boldsymbol{x})\|_2^2 \|E_1\|_2^2 = O_{\mathbb{P}}((1-\eta_2\mu)^{2k}) + O_{\mathbb{P}}\left(\frac{n^2(1-\eta_2\mu)^{2k-2}\sqrt{\log(n)}}{\sqrt{m}\lambda_0^2}\right), \tag{E.45}$$

$$|\boldsymbol{z}_0(\boldsymbol{x})^\top E_2|^2 \le \|\boldsymbol{z}_0(\boldsymbol{x})\|_2^2 \|E_2\|_2^2 = O_{\mathbb{P}}\left(\frac{n\tau^2}{\lambda_0}(1-\eta_2\mu)^{2k}\right) + O_{\mathbb{P}}\left(\frac{n^2(1-\eta_2\mu)^{2k-2}\sqrt{\log(n)}}{\sqrt{m}\lambda_0^2}\right), \tag{E.46}$$

$$|\boldsymbol{z}_0(\boldsymbol{x})^\top E_3|^2 \le \|\boldsymbol{z}_0(\boldsymbol{x})\|_2^2 \|E_3\|_2^2 = O_{\mathbb{P}}\left(\frac{\eta_1^6 n^8}{\eta_2^6\mu^6 m(1-\eta_2\mu)^{2k}\tau^2}\right), \tag{E.47}$$

$$|\boldsymbol{z}_0(\boldsymbol{x})^\top E_4|^2 \le \|\boldsymbol{z}_0(\boldsymbol{x})\|_2^2 \|E_4\|_2^2 = O_{\mathbb{P}}\left(\frac{n^3}{(1-\eta_2\mu)^k\mu^3\sqrt{m}\delta^{3/2}\tau}\right), \tag{E.48}$$

where (E.45) is because of (E.31) and (E.36), (E.46) is because of (E.32) and (E.37), (E.47) is because of (E.33), and (E.48) is because of (E.34). By Lemma D.5 (d), the term $(1-\eta_2\mu)^k\boldsymbol{z}_0(\boldsymbol{x})^\top\mathrm{vec}(\boldsymbol{W}_D(0))$ in (E.44) can be bounded by

$$\left\|(1-\eta_2\mu)^k\boldsymbol{z}_0(\cdot)^\top\mathrm{vec}(\boldsymbol{W}_D(0))\right\|_2 = O_{\mathbb{P}}((1-\eta_2\mu)^k\tau). \tag{E.49}$$

Define

$$B = \eta_1\boldsymbol{H}^\infty(\eta_2\mu I + \eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}.$$

Note that

$$\begin{aligned}
B - \boldsymbol{y} =& \eta_1\boldsymbol{H}^\infty(\eta_2\mu I + \eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y} - \boldsymbol{y}\\
=&(\eta_1\boldsymbol{H}^\infty - \eta_2\mu I - \eta_1\boldsymbol{H}^\infty)(\eta_2\mu I + \eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}\\
=&-\eta_2\mu(\eta_2\mu I + \eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}.
\end{aligned}$$

Therefore, the remaining term in (E.44) $-\boldsymbol{z}_0(\boldsymbol{x})^\top T_5 - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}$ can be bounded by

$$\begin{aligned}
&-\boldsymbol{z}_0(\boldsymbol{x})^\top T_5 - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\\
=&-\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0)\sum_{i=0}^{k-1}\eta_1(1-\eta_2\mu)^i(B-\boldsymbol{y}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\\
=&-\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0)\eta_1\frac{1-(1-\eta_2\mu)^k}{\eta_2\mu}(B-\boldsymbol{y}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\\
=&\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0)\eta_1(1-(1-\eta_2\mu)^k)(\eta_2\mu I + \eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y} - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\\
=&(\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0) - h(\boldsymbol{x},\boldsymbol{X}))(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y} - \eta_1(1-\eta_2\mu)^k\boldsymbol{z}_0(\boldsymbol{x})^\top\boldsymbol{Z}(0)(\eta_2\mu I + \eta_1\boldsymbol{H}^\infty)^{-1}\boldsymbol{y}.
\end{aligned} \tag{E.50}$$

The first term in (E.50) can be bounded by

$$\begin{aligned}
&\left\|(\boldsymbol{z}_0(\cdot)^\top\boldsymbol{Z}(0) - h(\cdot,\boldsymbol{X}))(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2\\
\le&\left\|(\boldsymbol{z}_0(\cdot)^\top\boldsymbol{Z}(0) - h(\cdot,\boldsymbol{X}))\right\|_2\left\|(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2\\
=&O_{\mathbb{P}}\left(\frac{n\sqrt{\log(n)}\eta_1}{\sqrt{m}\eta_2\mu}\right),
\end{aligned} \tag{E.51}$$

where we utilize

$$\left\|(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2^2 = \boldsymbol{y}^\top(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-2}\boldsymbol{y} \le \frac{\eta_1^2}{\eta_2^2\mu^2}\|\boldsymbol{y}\|_2^2 = O_{\mathbb{P}}\left(\frac{\eta_1^2}{\eta_2^2\mu^2}n\right),$$

and Lemma D.5 (c).

The second term in (E.50) can be bounded by

$$\begin{aligned}
&\left\|(1-\eta_2\mu)^k \boldsymbol{z}_0(\cdot)^\top \boldsymbol{Z}(0)(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2\\
\leq &(1-\eta_2\mu)^k \left\|(\boldsymbol{z}_0(\cdot)^\top \boldsymbol{Z}(0) - h(\cdot,\boldsymbol{X}))(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2\\
&+ (1-\eta_2\mu)^k \left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_2\\
\leq &O_{\mathbb{P}}\left(\frac{n\sqrt{\log(n)}\eta_1}{\sqrt{m}\eta_2\mu}\right) + (1-\eta_2\mu)^k \left\|h(\cdot,\boldsymbol{X})(\boldsymbol{H}^\infty + \eta_2\mu/\eta_1 I)^{-1}\boldsymbol{y}\right\|_{\mathcal{N}}\\
= &O_{\mathbb{P}}((1-\eta_2\mu)^k), \qquad\qquad (E.52)
\end{aligned}$$

where the second inequality is because of (E.51) and the last equality is because of Theorem 3.2 and the assumption $\eta_1 \asymp \eta_2$. Plugging (E.45)-(E.52) to (E.44), we can conclude that

$$\|\Delta_2\|_2 = o_{\mathbb{P}}(n^{-\frac{d}{2d-1}}), \qquad (E.53)$$

by choosing k and m as in Theorem 5.2. Combining (E.43), (E.53), and (E.41) finishes the proof.

F Proof of lemmas in the Appendix

F.1 Proof of Lemma B.1

The proof of Lemma B.1 mainly from Appendix C of [24] and Appendix D of [45], with some modification.

We first review some background of spherical harmonic analysis [46, 47]. Let $Y_{k,j}$ be the spherical harmonics of degree k on $\mathcal{S}^{d-1}$, where $N(p.k) = \frac{2k+d-2}{k}\begin{pmatrix} k+d-3 \\ d-2 \end{pmatrix}$. Then $Y_{k,j}$ is an orthonormal basis of $L_2(\mathcal{S}^{p-1}, d\xi)$, where $d\xi$ is the uniform measure on the sphere. Then we have

$$\sum_{j=1}^{N(d,k)} Y_{k,j}(\boldsymbol{s})Y_{k,j}(\boldsymbol{t}) = N(d,k)P_k(\boldsymbol{s}^\top\boldsymbol{t}), \qquad (F.1)$$

where P_k is the k-th Legendre polynomial in dimension d, given by

$$P_k(t) = (-1/2)^k \frac{\Gamma(\frac{d-1}{2})}{\Gamma(k+\frac{d-1}{2})}(1-t^2)^{(3-d)/2}\left(\frac{d}{dt}\right)^k (1-t^2)^{k+(d-3)/2}. \qquad (F.2)$$

The polynomials P_k are orthogonal in $L_2([-1,1])d\nu$, where the measure $d\nu = (1-t^2)^{(d-3)/2}dt$ with Lebesgue measure dt, and

$$\int_{[-1,1]} P_k^2(t)(1-t^2)^{(d-3)/2}dt = \frac{w_{d-1}}{w_{d-2}}\frac{1}{N(d,k)}, \qquad (F.3)$$

where $w_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}$. Furthermore, it can be shown that [46]

$$tP_k(t) = \frac{k}{2k+d-2}P_{k-1}(t) + \frac{k+d-2}{2k+d-2}P_{k+1}(t), \qquad (F.4)$$

for $k \geq 1$, and for $j = 0$ we have $tP_0(t) = P_1(t)$. This implies that for large k enough, we have

$$\mu_k = \frac{k}{2k+d-2}\mu_{0,k-1} + \frac{k+d-2}{2k+d-2}\mu_{0,k+1},$$

where $\mu_{0,k-1}$ and $\mu_{0,k+1}$ are as in Lemma 17 of [24]. By Lemma 17 of [24], we have $\mu_{0,k} \asymp k^{-d}$ for large k, if $k = 1$ mod 2. This finish the proof of Lemma B.1.

F.2 Proof of Lemma C.1

By Theorem 1 of [48] and Lemma B.1, we can see that the function space $\mathcal{N}$ is a subspace of the Sobolev space $H^s(\mathcal{S}^{d-1})$. Therefore, the entropy of $\mathcal{N}(1)$ can be bounded if the entropy of $H^{d/2}(\mathcal{S}^{d-1})(1)$ can be bounded. By Theorem 1.2 of [49], we have that the k-th entropy number $e_k(T)$ can be bounded by $k^{-d/(2(d-1))}$. This implies that

$$H(\delta, \mathcal{N}(1), \|\cdot\|_{L_\infty}) \leq A\delta^{-\frac{2(d-1)}{d}}.$$

F.3 Proof of Lemma D.1

The first inequality follows the fact that h is positive definite, which implies the inverse of

$$\begin{pmatrix} h(\boldsymbol{s},\boldsymbol{s}) & h(\boldsymbol{X},\boldsymbol{s}) \\ h(\boldsymbol{s},\boldsymbol{X}) & \boldsymbol{h}^{\infty} \end{pmatrix}$$

is positive definite. By block matrix inverse, we have the first inequality in Lemma D.1 holds.

The second inequality and third inequality are direct results of Theorem 3.2 implies

$$\begin{aligned}
&\mathbb{E}_{\epsilon,\boldsymbol{X}}(\|\widehat{g}_n - g^*\|_2^2)\\
=&\int_{\mathbb{S}^{d-1}} (g^*(\boldsymbol{x}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-1}\boldsymbol{y}^*)^2 + h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-2}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x} = O_{\mathbb{P}}(n^{-\frac{d}{2d-1}})
\end{aligned}$$

for any function g^* with $\|g^*\|_{\mathcal{N}} \le 1$. Then we have

$$\int_{\mathbb{S}^{d-1}} h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-2}h(\boldsymbol{X},\boldsymbol{x})d\boldsymbol{x} = O_{\mathbb{P}}(n^{-\frac{d}{2d-1}}),$$

which finishes the proof of the second equality. Let $g^*(\boldsymbol{x}) = h(\boldsymbol{s},\boldsymbol{x})$, then we have

$$\int_{\mathbb{S}^{d-1}} (h(\boldsymbol{s},\boldsymbol{x}) - h(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))^2 d\boldsymbol{x} = O_{\mathbb{P}}(n^{-\frac{d}{2d-1}}).$$

By the interpolation inequality, we have

$$\begin{aligned}
&h(\boldsymbol{s},\boldsymbol{s}) - h(\boldsymbol{s},\boldsymbol{X})(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))\\
\le&\left\|h(\boldsymbol{s},\cdot) - h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))\right\|_{\infty}\\
\le& C\left\|h(\boldsymbol{s},\cdot) - h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))\right\|_2^{1-\frac{d-1}{d}} \left\|h(\boldsymbol{s},\cdot) - h(\cdot,\boldsymbol{X})(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s})\right\|_{\mathcal{N}}^{\frac{d-1}{d}}\\
=&O_{\mathbb{P}}(n^{-\frac{1}{2d-1}})(h(\boldsymbol{s},\boldsymbol{s}) + h(\boldsymbol{s},\boldsymbol{X})(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-1}\boldsymbol{H}^{\infty}(\boldsymbol{H}^{\infty}+\mu\boldsymbol{I})^{-1}h(\boldsymbol{X},\boldsymbol{s}))^{\frac{d-1}{d}}\\
\le&O_{\mathbb{P}}(n^{-\frac{1}{2d-1}})(h(\boldsymbol{s},\boldsymbol{s}) + h(\boldsymbol{s},\boldsymbol{X})(\boldsymbol{H}^{\infty})^{-1}h(\boldsymbol{X},\boldsymbol{s}))^{\frac{d-1}{d}} = O_{\mathbb{P}}(n^{-\frac{1}{2d-1}}),
\end{aligned}$$

where the last inequality follows the first inequality of Lemma D.1.

F.4 Proof of Lemma D.2

Given that g and f^* have the same value at all $\boldsymbol{x}_i$'s, the empirical norm $\|g - f^*\|_n = 0$. Notice that both g and f^* are in the RKHS generated by the NTK h, denoted by $\mathcal{N}$. Utilizing Lemma C.1 and C.3 similarly as in the proof of Theorem 3.2, we have $R, K = O(1)$ and $J_{\infty}(z,\mathcal{N}) \lesssim z^{1/d}$, which leads to

$$\sup_{h\in\mathcal{G}(R)} \left| \|h\|_n^2 - \|h\|_2^2 \right| = O_{\mathbb{P}}\left(\sqrt{\frac{1}{n}}\right),$$

where $\mathcal{G}(R) := \{g \in \mathcal{N}(1) : \|g - g^*\|_2 \le R\}$. Therefore, we can conclude that $\|g - f^*\|_2 = O_{\mathbb{P}}(n^{-1/2})$.

F.5 Proof of Lemma D.5

The proof of (a) and (b) can be found in [9].

For (c), the i-th coordinates of $\boldsymbol{z}_0(\boldsymbol{x})^{\top}\boldsymbol{Z}(0)$ and $h(\boldsymbol{x},\boldsymbol{X})$ are

$$\frac{1}{m}\sum_{r=1}^{m} \boldsymbol{x}^{\top}\boldsymbol{x}_i\mathbb{I}\{\boldsymbol{w}_r^{\top}(0)\boldsymbol{x} \ge 0\}\mathbb{I}\{\boldsymbol{w}_r^{\top}(0)\boldsymbol{x}_i \ge 0\}, \text{ and } \mathbb{E}_{\boldsymbol{w}\sim N(0,\boldsymbol{I})}[\boldsymbol{x}^{\top}\boldsymbol{x}_i\mathbb{I}\{\boldsymbol{w}^{\top}\boldsymbol{x} \ge 0\}\mathbb{I}\{\boldsymbol{w}^{\top}\boldsymbol{x}_i \ge 0\}],$$

respectively. $\forall i \in [n]$, $(\boldsymbol{z}_0(\boldsymbol{x})^{\top}\boldsymbol{Z}(0))_i$ is the average of m i.i.d. random variables, which have expectation $h_i(\boldsymbol{x},\boldsymbol{X})$ and bounded in $[0,1]$. For any fixed $\boldsymbol{x}$, by Hoeffding's inequality, with probability at least $1-\delta^*$,

$$|(\boldsymbol{z}_0(\boldsymbol{x})^{\top}\boldsymbol{Z}(0))_i - h_i(\boldsymbol{x},\boldsymbol{X})| \le \sqrt{\frac{\log(2/\delta^*)}{2m}}$$

holds. By defining $\delta = n\delta^*$ and applying a union bound over all $i \in [n]$, with probability at least $1-\delta$, we have

$$\left\|\boldsymbol{z}_0(\boldsymbol{x})^\top \boldsymbol{Z}(0) - h(\boldsymbol{x}, \boldsymbol{X})\right\|_2^2 = O\left(n\frac{\log(2n/\delta)}{2m}\right)$$

For (d), since

$$\boldsymbol{z}_0(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{W}(0)) = \frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r \mathbb{I}\{\boldsymbol{w}_r(0)^\top \boldsymbol{x} \geq 0\}\boldsymbol{w}_r(0)^\top \boldsymbol{x}$$

Define random variables V_r, $r \in [m]$ as

$$V_r = a_r \mathbb{I}\{\boldsymbol{w}_r(0)^\top \boldsymbol{x} \geq 0\}\boldsymbol{w}_r(0)^\top \boldsymbol{x}$$

Since

$$\boldsymbol{w}_r(0)^\top \boldsymbol{x} \;\sim\; N(0, \tau^2) \quad \text{and} \quad a_r \;\sim\; \mathrm{unif}\{1, -1\}.$$

It's easy to prove that V_r, $r \in [m]$ are i.i.d. with mean 0 and sub-Gaussian parameter τ. By Hoeffding's inequality, at fixed bx, with probability at least $1-\delta$, we have

$$\left|\frac{1}{\sqrt{m}}\sum_{r=1}^{m} V_r\right| \leq \sqrt{2}\tau\sqrt{\log(2/\delta)}.$$

Thus $\left\|\boldsymbol{z}_0(\cdot)^\top \mathrm{vec}(\boldsymbol{W}(0))\right\|_2 = O\left(\tau\sqrt{\log(1/\delta)}\right)$.

G More details and results for numerical experiments

Neural network setup The neural network used in all experiments is a 2-layer ReLU neural network with $m = 500$ nodes in each hidden layer. All the weighs are initialized with the Glorot uniform initializer, also called as Xavier uniform initializer [50], which is the default choice in the TensorFlow Keras Sequential module. All the weights are trained by RMSProp [38] optimizer with the default setting, e.g. learning rate of 0.001, etc. All ONN experiments are conducted using TensorFlow 2 with Python API.

G.1 Simulated Data

The learning rate for NTK+ES is $\eta = 0.01$ and the GD update rule is as specified in (D.19). In the ℓ_2-regularized methods, the tuning parameter μ for each task is chosen by cross validation. The validation dataset is of size 100 that is also noiseless and follows the same generating mechanism as the test dataset. For NTK+ℓ_2, we use a grid search of interval $[0, 1]$ with $\mu = 0.01, 0.02, \ldots, 1$ and for ONN+ℓ_2, the μ candidates are $0.1, 0.2, \ldots, 10$. In both cases, we observe that the optimal μ increases with the noise level σ. For f_2^*, we plot the chosen μ and k^* for NTK+ℓ_2 and NTK+ES respectively vs. σ. For each σ value, the reported value is the average of 100 replications. The results are shown in Figure 3.

Figure 1 clearly demonstrates that ONN and NTK do not recover the true function well. As is explained in the paper, without regularization, overfitting the training data is harmful for the L_2 estimation. To illustrate this point, we show the trained estimators of f_2^* for all the methods in Figure 4 when $\sigma = 0.1$.

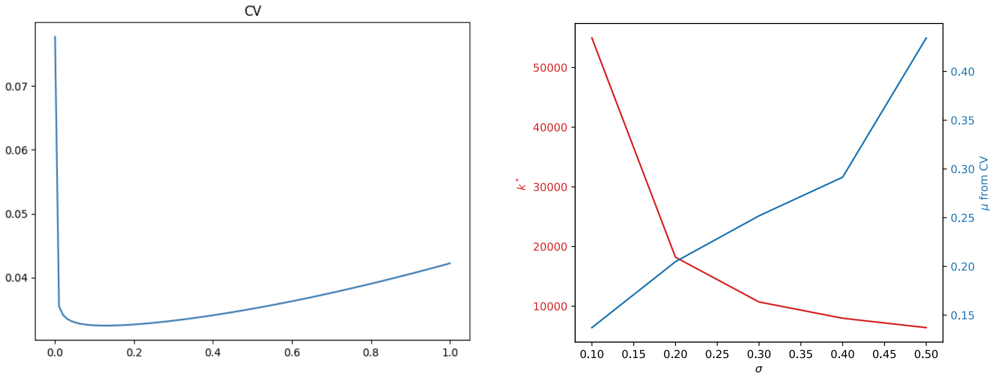


Figure 3: Left: Cross-validation of μ in NTK+ℓ_2 for fitting f_2^* when $\sigma = 0.1$. The horizontal axis is values of μ (100 points from 0.01 to 1) and the vertical axis is the validation mean squared error. The cross-validated μ in this case is 0.13. Right: Optimal stopping time k^* in NTK+ES and cross-validated μ in NTK+ℓ_2 for fitting f_2^* are shown vs. σ. The optimal GD stopping time decrease with noise level while the best μ increases with σ.

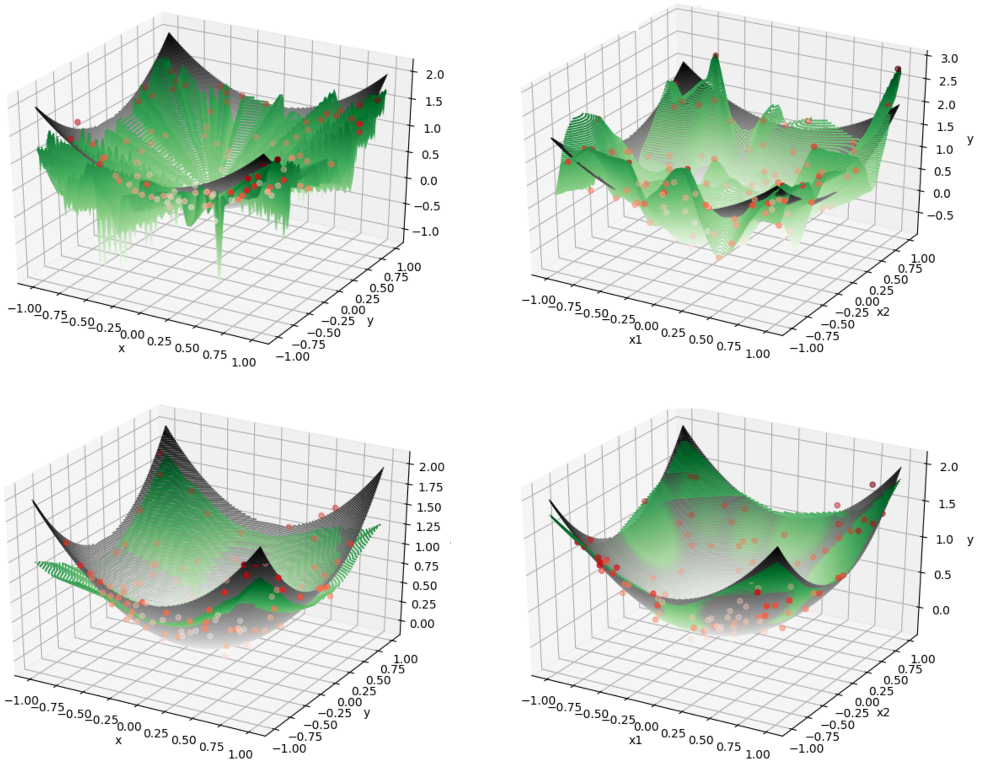


Figure 4: Visualizations for the trained estimators of NTK (top left), NTK+ℓ_2 (bottom left), ONN (top right) and ONN+ℓ_2 (bottom right). Training data are plotted as red dots. The green surface is the estimator and the grey surface is the true function f_2^*. Both surfaces are approximated by grid points $(i/100, j/100)$ for i, j from -100 to 100. As can be seen in the top row, without regularization, the estimators overfit training data. The fitted estimators are very rough and don't recover the true function well.

G.2 MNIST

For images 5 and 8, the training and test split are the default.[5] We change label 5 and 8 to -1 and 1 respectively. No further pre-processing is done to the dataset. For NTK+ES, the learning rate is $\eta = 0.0001$ and the GD update rule is as specified in (D.19). To account for the high data dimension, we divide the NTK matrix $\boldsymbol{H}^\infty$ by d. For the ONN+ℓ_2 and NTK+ℓ_2, we choose μ by cross-validation and the candidates are $\mu = 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000$ for ONN+ℓ_2 and $\mu = 1, 2, 3, \dots, 100$ for NTK+ℓ_2. The training/validation split is 80%/20% for cross-validation so the actual training data size is 9107 for all methods (ONN, NTK and NTK+ES do not use the validation dataset). The cross-validated μ for ONN+ℓ_2 and optimal stopping time k^* for NTK+ES are shown in Figure 5, together with the cross-validation results specifically for $\sigma = 1$.

NTK+ES The performance of NTK+ES is shown in Figure 6. Unlike in the simulated dataset where NTK+ES and NTK+ℓ_2 perform almost identically, NTK+ES performs noticeably worst for the MNIST dataset, especially when σ is small. One possible explanation lies in our additive label noise setting. Even though we treat the labels as continuous during training, the reported misclassification rate only depends on the sign of the label. If σ is small, the probability of changing signs is small. This may be one of the reasons that NTK, ONN perform relatively well for small σ's, since if the signs remain the same, it is not very harmful to overfit the labels. Note that NTK+ℓ_2 and ONN+ℓ_2 choose small μ's such that it is not very different from NTK and ONN. The stopping rule in NTK+ES, on the other hand, doesn't take the classification setting into consideration and tends to underestimate the stopping time when the additive label noises are small. Nonetheless, we don't recommend NTK+ES for handling large datasets. On one hand, the noise level is to be estimated, which brings extra instability to the algorithm. On the other hand, NTK+ES is very computationally intensive. The eigenvalue computation is of $O(n^3)$ complexity. The optimal stopping time is only for GD (not for adaptive gradient-based algorithms) and is often very large.

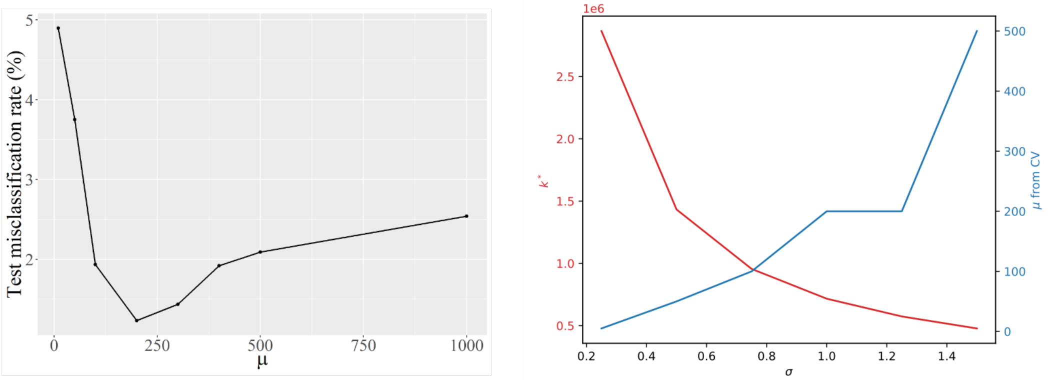


Figure 5: Left: Cross-validation result for μ in ONN+ℓ_2 when $\sigma = 1$ (with extra μ candidates of 300 and 400). In the range of $\mu = 5$ to $\mu = 1000$, we can clearly see a V-shape and the best μ in this case is 200. Right: Optimal stopping time k^* in NTK+ES and cross-validated μ in ONN+ℓ_2 for MNIST dataset are shown vs. σ. The optimal stopping time decreases with noise level while the best μ increases with σ.

[5] http://yann.lecun.com/exdb/mnist/

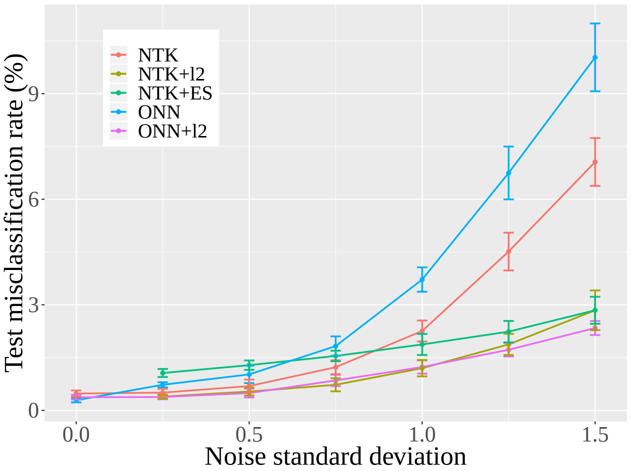


Figure 6: The test misclassification rates for all methods vs. σ are shown with their standard deviations plotted as vertical bars. NTK+ES for $\sigma = 0$ is omitted since k^* is not well-defined when $\sigma = 0$ and NTK+ES in this case should be the same as NTK, i.e. $k^* = \infty$. As σ increases, all misclassification rates increase but NTK+ℓ_2 and ONN+ℓ_2 perform significantly better than NTK and ONN with smaller misclassification rate and better stability, i.e., the standard deviation is smaller. The NTK+ES is the green line and it performs the worst when $\sigma \leq 0.5$ but better than NTK and ONN when $\sigma \geq 1$.