TOWARDS UNDERSTANDING THE SPECTRAL BIAS OF DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

An intriguing phenomenon observed during training neural networks is the spectral bias, where neural networks are biased towards learning less complex functions. The priority of learning functions with low complexity might be at the core of explaining generalization ability of neural network, and certain efforts have been made to provide theoretical explanation for spectral bias. However, there is still no satisfying theoretical results justifying the existence of spectral bias. In this work, we give a comprehensive and rigorous explanation for spectral bias and relate it with the neural tangent kernel function proposed in recent work. We prove that the training process of neural networks can be decomposed along different directions defined by the eigenfunctions of the neural tangent kernel, where each direction has its own convergence rate and the rate is determined by the corresponding eigenvalue. We then provide a case study when the input data is uniformly distributed over the unit shpere, and show that lower degree spherical harmonics are easier to be learned by over-parameterized neural networks.

1 INTRODUCTION

Over-parameterized neural networks have achieved great success in many applications such as computer vision (He et al., 2016), language processing (Collobert and Weston, 2008) and speech recognition (Hinton et al., 2012). It has been shown that over-parameterized neural networks can fit complicated target function or even randomly labeled data (Zhang et al., 2016) and still exhibit good generalization performance when trained with real labels. Intuitively, this is at odds with the traditional notions of generalization ability such as model complexity. In lack of enough justification for over-parameterized training, efforts have been made towards the perspective of "implicit bias" (Gunasekar et al., 2018b; Soudry et al., 2017; Gunasekar et al., 2018a), which states that the training algorithms for deep learning implicitly pose an inductive bias onto the training process and lead to a solution with low complexity.

In many attempts to establish implicit bias, an intriguing phenomenon called *spectral bias* was first presented in (Rahaman et al., 2018). The *spectral bias* means that during training, neural networks tend to learn components of lower complexity faster. The concept of spectral bias is appealing because this may intuitively explain why over-parameterized neural networks can achieve great accuracy without overfitting. During training, the networks only fit the low complexity components first and thus lie in the concept class of low complexity. Arguments like this may guide rigorous guarantee for generalization.

Among many works to seek and explain spectral bias, Rahaman et al. (2018) evaluates the Fourier spectrum of ReLU networks and empirically showed that the lower frequencies are learned first; also lower frequencies are more robust to random perturbation. Andoni et al. (2014) shows that for a sufficiently wide network, gradient descent with respect to the second layer can learn any low degree bounded polynomial. Xu (2018) gives Fourier analysis to two-layer networks and show similar empirical results on one-dimensional functions and real data. Nakkiran et al. (2019) uses information theoretical approach to shows that networks obtained by stochastic gradient descent can be explained by a linear classifier early during training. All these works provide convincing evidence that neural networks exhibit spectral bias in real tasks. But the theoretical explanations, if any, are to some extent restricted. For example, the popular Fourier analysis is usually constrained under the one-dimensional setting, and thus lacks generality.

Meanwhile, a recent line of works (Jacot et al., 2018; Du et al., 2018b; Li and Liang, 2018) have shed lights on new approaches to analyze neural networks. In particular, they show that under certain over-parameterized condition, the learning dynamics of neural networks will be characterized by the kernel gradient with respect to the Neural Tangent Kernel (NTK). Du et al. (2018b) at the same time shows that under NTK regime, the convergence is provably guaranteed by the smallest eigenvalue of NTK. Arora et al. (2019) further gives a finer characterization of error convergence based on the eigenvalues of NTK's gram matrix. Su and Yang (2019) improves the convergence guarantee to the k-th largest eigenvalue, given that the target function admits k-order approximation.

Inspired by these works mentioned above, we can present a theoretical explanation for spectral bias. Under NTK regime, we can give a precise characterization for the training process of neural networks. More specifically, we theoretically prove that over-parameterized neural networks' training process can be controlled by the eigenvalues of the integrating operator defined by the NTK. Under the specific case of uniform distribution on unit sphere, we give an exact calculation for these eigenvalues and show that the lower frequencies have larger eigenvalues, which thus leads to faster convergence. We also conduct experiments to corroborate the theory we establish.

Our contributions are highlighted as follows:

1. We prove a general, distribution-independent theorem which states that under sufficient samples and over-parameterization conditions, the error term's convergence along different directions actually relies on the corresponding eigenvalues. This theorem gives finer-grained control on error term's than Su and Yang (2019), in which it also controls the error term's projection onto certain directions.
2. We present a more general results about the spectra of the neural tangent kernel. Both layers of the neural network with ReLU activation are trained. In particular, we show that the order of eigenvalues appears as $\mu_k = \mathcal{O}(\min\{k^{-d-1}, d^{-k+1}\})$, which is better than the bound $\mathcal{O}(k^{-d-1})$ derived in Bietti and Mairal (2019) when $d \gg k$. It is clear that our bound is closer to the natural data's setting.
3. We establish a rigorous explanation for the spectral bias, based on the aforementioned theoretical results. The error terms from different frequencies are provably controlled by the eigenvalues of the NTK, and it is shown that the lower-frequency components enjoys faster convergence rate. As far as the authors know, this is the first attempt to give a comprehensive theory justifying the existence of spectral bias.

1.1 Additional Related Work

Recently, there is a rich literature about the property of neural tangent kernel. Jacot et al. (2018) first shows that during training, the network function follows a descent along the kernel gradient with respect to the Neural Tangent Kernel (NTK) under infinity width setting. Li and Liang (2018) and Du et al. (2018b) implicitly build connection between Neural Tangent Kernel and gradient descent by showing that GD can provably optimize sufficiently wide two-layer neural networks. In Du et al. (2018b), it is shown gradient descent can achieve zero training loss, at a linear convergence rate, for training two-layer ReLU network with square loss. However, these works only consider the smallest eigenvalue λ_0 of the Gram matrix. Allen-Zhu et al. (2018); Du et al. (2018a); Zou et al. (2018); Cao and Gu (2019b); Zou and Gu (2019); Cao and Gu (2019a) further study the optimization and generalization of deep neural networks, and are all either implicitly or explicitly related to the neural tangent kernel. Later, Su and Yang (2019) shows that the smallest eigenvalue actually scales in the number of samples n and will eventually converges to 0. In order to obtain constant convergence rate, Su and Yang (2019) assumes the target function f^* can be approximated by the first few eigenfunctions of the integrating operator $L_\kappa f(s) := \int_{\mathbb{S}^d} \kappa(x, s) f(s) \tau(s)$ and gives constant rate as $\log(1 - \frac{3}{4}\eta\mu_l)$, where η is the step size and μ_l is the l-th eigenvalue of L_κ. Although they achieve non-zero convergence rate, they still depend on the minimal eigenvalue in the eigenspace.

A few theoretical studies have been done towards understanding the spectra of neural tangent kernels. Bach (2017), though not relevant to neural tangent kernel, gives a harmonic decomposition for any α-homogeneous function (including ReLU) and the precise coefficient. Based on the technique in Bach (2017), Bietti and Mairal (2019) shows eigenvalue decay of integrating operator $L_\kappa f(x)$ defined by NTK on unit sphere by using spherical harmonics. In Vempala and Wilmes (2018), they consider two-layer neural networks with sigmoid activation function and the eigenvalues are computed. Basri et al. (2019) gives similar results as Bietti and Mairal (2019) where the one-hidden layer network is equipped with bias, but only the first layer is optimized. Exact eigenvalues of integral operator with respect to the NTK on Boolean Cube is presented in Yang and Salman (2019) by Fourier Analysis.

The rest of the paper is organized as follows. We state the notation, problem setup and other preliminaries in Section 2 and present our main results in Section 3. In Section 4, we present experimental results to support our theory. Proofs of our main results can be found in Appendix.

2 Preliminaries

In this section we introduce the basic problem setup including the neural network structure and the training algorithm, as well as some background on the neural tangent kernel proposed recently in Jacot et al. (2018) and the corresponding integral operator.

2.1 Notation

We use lower case, lower case bold face, and upper case bold face letters to denote scalars, vectors and matrices respectively. For a vector $\mathbf{v} = (v_1, \dots, v_d)^T \in \mathbb{R}^d$ and a number $1 \leqslant p < \infty$, we denote its $p-$norm by $\|\mathbf{v}\|_p = (\sum_{i=1}^d |v_i|^p)^{1/p}$. We also define infinity norm by $\|\mathbf{v}\|_\infty = \max_i |v_i|$. For a matrix $\mathbf{A} = (A_{i,j})_{m \times n}$, we use $\|\mathbf{A}\|_0$ to denote the number of non-zero entries of $\mathbf{A}$, and use $\|\mathbf{A}\|_F = (\sum_{i,j=1}^d A_{i,j}^2)^{1/2}$ to denote its Frobenius norm. Let $\|\mathbf{A}\|_p = \max_{\|\mathbf{v}\|_p \leqslant 1} \|\mathbf{A}\mathbf{v}\|_p$ for $p \geqslant 1$, and $\|\mathbf{A}\|_{\max} = \max_{i,j} |A_{i,j}|$. For two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, we define $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^\top \mathbf{B})$. We use $\mathbf{A} \succeq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is positive semi-definite. In addition, we define the asymptotic notations $\mathcal{O}(\cdot)$, $\widetilde{\mathcal{O}}(\cdot)$, $\Omega(\cdot)$ and $\widetilde{\Omega}(\cdot)$ as follows. Suppose that a_n and b_n be two sequences. We write $a_n = \mathcal{O}(b_n)$ if $\limsup_{n \to \infty} |a_n/b_n| < \infty$, and $a_n = \Omega(b_n)$ if $\liminf_{n \to \infty} |a_n/b_n| > 0$. We use $\widetilde{\mathcal{O}}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to hide the logarithmic factors in $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$.

2.2 Problem Setup

Here we introduce the basic problem setup. We consider two-layer fully connected neural networks of the form

$$f_{\mathbf{W}}(\mathbf{x}) = \sqrt{m} \cdot \mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}),$$

where $\mathbf{W}_1 \in \mathbb{R}^{m \times (d+1)}$, $\mathbf{W}_2 \in \mathbb{R}^{1 \times m}$[1] are the first and second layer weight matrices respectively, and $\sigma(\cdot) = \max\{0, \cdot\}$ is the entry-wise ReLU activation function. The network is trained according to the square loss on n training examples $S = \{(\mathbf{x}_i, y_i) | i \in [n]\}$

$$L_S(\mathbf{W}) = \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in S} (y_i - \theta f_{\mathbf{W}}(\mathbf{x}_i))^2,$$

where θ is a small coefficient to control the effect of initialization, and the sample $\{\mathbf{x}_i\}_{i=1}^n$ is assumed to follow some unknown distribution τ on the unit sphere $\mathbb{S}^d \in \mathbb{R}^{d+1}$. Without loss of generality, we also assume that $y_i \leqslant 1$.

We first use random initialization to parameters in our networks and then apply gradient descent to optimize both layers. We present our detailed neural network training algorithm in Algorithm 1.

[1] Here the dimension of input is $d + 1$ since throughout this paper we assume that all training data lie in the d-dimension unit sphere $\mathbb{S}^d \in \mathbb{R}^{d+1}$.

Algorithm 1 GD for DNNs starting at Gaussian initialization

```
Input: Number of iterations $T$, step size $\eta$.
Generate each entry of $\mathbf{W}_1^{(0)}$ and $\mathbf{W}_2^{(0)}$ from $N(0, 2/m)$ and $N(0, 1/m)$ respectively.
for $t = 0, 1, \ldots, T-1$ do
    Update $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \cdot \nabla_{\mathbf{W}} L_S(\mathbf{W}^{(t)})$.
end for
Output: $\mathbf{W}^{(T)}$.
```

The initialization scheme for $\mathbf{W}^{(0)}$ given in Algorithm 1 is known as He initialization (He et al., 2015). This scheme generates each entry of the weight matrices from a Gaussian distribution with mean zero. The variance follows the principal that the initialization do not change the magnitudes of inputs in each layer. The second layer parameter is not associated with the ReLU activation function, thus it is initialized with variance $1/m$ instead of $2/m$.

2.3 Neural Tangent Kernel

Attempts are made to study the convergence of gradient descent assuming the width of the network is infinity (Du et al., 2018b; Li and Liang, 2018). When the width of the network goes to infinity, with certain initialization on parameters, inner product of gradients of the output function would converge to a limiting kernel, the Neural Tangent Kernel (Jacot et al., 2018). In this paper, we denote it as $\kappa(\mathbf{x}, \mathbf{x}') = \lim_{m\to\infty} m^{-1}\langle \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}), \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}')\rangle$ and we have

$$\kappa(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}'\rangle \cdot \kappa_1(\mathbf{x}, \mathbf{x}') + 2 \cdot \kappa_2(\mathbf{x}, \mathbf{x}'), \tag{2.1}$$

with

$$\begin{aligned}\kappa_1(\mathbf{x}, \mathbf{x}') &= \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}[\sigma'(\langle \mathbf{w}, \mathbf{x}\rangle)\sigma'(\langle \mathbf{w}, \mathbf{x}'\rangle)],\\ \kappa_2(\mathbf{x}, \mathbf{x}') &= \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}[\sigma(\langle \mathbf{w}, \mathbf{x}\rangle)\sigma(\langle \mathbf{w}, \mathbf{x}'\rangle)].\end{aligned} \tag{2.2}$$

Since we apply gradient descent to both layers, the Neural Tangent Kernel is the sum of two different kernel functions and clearly it can be reduced to one layer training setting. From Cho and Saul (2009) we know that these two kernels are arc-cosine kernels of degree 0 and 1. Explicit expressions are given with $t = \langle \mathbf{x}, \mathbf{x}'\rangle / (\|\mathbf{x}\| \, \|\mathbf{x}'\|)$

$$\kappa_1(t) = \frac{1}{2\pi}\left(\pi - \arccos(t)\right) \quad \kappa_2(t) = \frac{1}{2\pi}\left(t \cdot (\pi - \arccos(t)) + \sqrt{1 - t^2}\right). \tag{2.3}$$

2.4 Integral Operator

The theory of integral operator with respect to kernel function has been well studied in machine learning (Smale and Zhou, 2007; Rosasco et al., 2010) thus we only give a brief introduction here. Let $L^2_\tau(X)$ be the Hilbert space of square-integrable functions with respect to a Borel measure τ from $X \to \mathbb{R}$. For any continuous kernel function $\kappa : X \times X \to \mathbb{R}$ and τ we can define an integral operator L_κ on $L^2_\tau(X)$ by

$$L_\kappa(f)(\mathbf{x}) = \int_X \kappa(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\tau(\mathbf{y}), \quad \mathbf{x} \in X. \tag{2.4}$$

It has been pointed out in Cho and Saul (2009) that arc-cosine kernels are positive semi-definite. Thus the kernel function κ defined by (2.1) is positive semi-definite being a product and a sum of positive semi-definite kernels. Clearly this kernel is also continuous and symmetric. Thus we know that this Neural Tangent Kernel is a Mercer kernel and we will present Mercer decomposition in next section.

3 Main Results

In this section we present our main results. In Section 3.1, we give a general result on the convergence rate of gradient descent along different eigendirections of neural tangent kernel. Motivated by this result, in Section 3.2, we give a case study on the spectrum of L_κ when the input data are uniformly distributed over the unit sphere $\mathbb{S}^d$. In Section 3.3, we combine the spectrum analysis with the general convergence result to give explicit convergence rate for uniformly distributed data on the unit sphere.

3.1 Convergence Analysis of Gradient Descent

In this section we study the convergence of Algorithm 1. Instead of studying the standard convergence of loss function value, we aim to provide a refined analysis on the speed of convergence along different directions defined by the eigenfunction of L_κ. We first introduce the following definitions and notations.

Let $\{\lambda_i\}_{i\geqslant 1}$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots$ be the strictly positive eigenvalues of L_κ, and $\phi_1(\cdot), \phi_2(\cdot), \cdots$ be the corresponding orthornormal eigenfunctions. Set $\mathbf{v}_i = n^{-1/2}(\phi_i(\mathbf{x}_1), \ldots, \phi_i(\mathbf{x}_n))$, $i \in [n]$. Note that L_κ may have eigenvalues with multiplicities larger then 1 and λ_i, $i \geqslant 1$ are not distinct. Therefore for any integer k, we define r_k as the sum of the multiplicities of the first k distinct eigenvalues of L_κ. Define $\mathbf{V}_{r_k} = (\mathbf{v}_1, \ldots, \mathbf{v}_{r_k})$. By definition, $\mathbf{v}_i$, $i \in [r_k]$ are rescaled restrictions of orthornormal functions in $L^2_\tau(\mathbb{S}^d)$. Therefore we can expect them to form a set of almost orthornomal bases in the vector space $\mathbb{R}^n$. The following lemma follows by standard concentration inequality.

Lemma 3.1. Suppose that $|\phi_i(\mathbf{x})| \leqslant M$ for all $\mathbf{x} \in \mathbb{S}^d$. For any $\delta > 0$, with probability at least $1 - \delta$,

$$\|\mathbf{V}_{r_k}^\top \mathbf{V}_{r_k} - \mathbf{I}\|_{\max} \leqslant C\sqrt{M \log(r_k/\delta)/n},$$

where C is an absolute constant.

Denote $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\widehat{\mathbf{y}}^{(t)} = \theta \cdot (f_{\mathbf{W}^{(t)}}(\mathbf{x}_1), \ldots, f_{\mathbf{W}^{(t)}}(\mathbf{x}_n))$, $t = 0, \ldots, T$. Then Lemma 3.1 shows that the convergence rate of $\|\mathbf{V}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(t)})\|_2$ roughly represents the speed gradient descent learns the components of the target function corresponding to the first r_k eigenvalues. The following theorem gives the convergence guarantee of $\|\mathbf{V}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(t)})\|_2$.

Theorem 3.2. Suppose $|\phi_j(\mathbf{x})| \leqslant M$ for $j \in [r_k]$ and $\mathbf{x} \in \mathbb{S}^d$. For any $\epsilon, \delta > 0$ and integer k, if $n \geqslant \widetilde{\Omega}(\max\{\epsilon^{-1}(\lambda_{r_k} - \lambda_{r_k+1})^{-1}, \epsilon^{-2}M^2 r_k^2\})$, $m \geqslant \widetilde{\Omega}(\mathrm{poly}(T, \lambda_{r_k}, \epsilon^{-1}))$, then with probability at least $1 - \delta$, Algorithm 1 with $\eta = \widetilde{\mathcal{O}}((m\theta^2)^{-1})$, $\theta = \epsilon/16$ satisfies

$$n^{-1/2} \cdot \|\mathbf{V}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 \leqslant 2(1 - \lambda_{r_k})^T \cdot n^{-1/2} \cdot \|\mathbf{V}_{r_k}^\top \mathbf{y}\|_2 + \epsilon.$$

Remark 3.3. Theorem 3.2 theoretically reveals the spectral bias of deep learning. Specifically, as long as the network is wide enough and the sample size is large enough, gradient descent first learns the target function along the the eigendirections of neural tangent kernel with larger eigenvalues, and then learns the rest components corresponding to smaller eigenvalues. Therefore, Theorem 3.2 theoretically explains the empirical observations given in Rahaman et al. (2018), and demonstrates that the difficulty of a function to be learned by neural network should be studied in the eigenspace of neural tangent kernel: if the target function has a component corresponding to a small eigenvalue of neural tangent kernel, then learning this function to good accuracy takes longer, and requires more samples and wider network.

3.2 Spectral analysis of Neural Tangent Kernel for Uniform Distribution

After presenting a general theorem (without assuming data distribution) in above subsection, we restrict sample distribution to be uniform distribution on the unit sphere. We present our results (an extension of **Proposition 5** in Bietti and Mairal (2019)) of spectral analysis of Neural Tangent Kernel. We show Mercer decomposition of Neural Tangent Kernel for two layers setting. We give explicit expression of eigenvalues and show orders of eigenvalues in both cases when $d \gg k$ and $k \gg d$.

Theorem 3.4. For any $\mathbf{x}, \mathbf{x}' \in \mathbb{S}^d \subset \mathbb{R}^{d+1}$, we have the Mercer decomposition of the Neural Tangent Kernel $\kappa : \mathbb{S}^d \times \mathbb{S}^d \to \mathbb{R}$,

$$\kappa\left(\mathbf{x}, \mathbf{x}'\right) = \sum_{k=0}^{\infty} \mu_k \sum_{j=1}^{N(d,k)} Y_{k,j}\left(\mathbf{x}\right) Y_{k,j}\left(\mathbf{x}'\right), \tag{3.1}$$

where $Y_{k,j}$ for $j = 1, \cdots, N(d,k)$ are linearly independent spherical harmonics of degree k in $d+1$ variables with $N(d,k) = \frac{2k+d-1}{k}\binom{k+d-2}{d-1}$ and orders of μ_k are explicitly given by

$$\mu_0 = \frac{d-1}{2d\pi} + \frac{d-1}{2d\pi}\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+2}{2}\right)}, \qquad \mu_{k'} = 0,\ k' = 2j+1,\ j \in \mathbb{N}^+,$$

$$\mu_k = \mathcal{O}\left(\max\left\{d^{d+1}k^{k-1}(k+d)^{-k-d}, d^{d+1}k^k(k+d)^{-k-d-1}, d^{d+2}k^{k-2}(k+d)^{-k-d-1}\right\}\right),$$

for $k \geqslant 1,\ k \neq k'$. More specifically, we have $\mu_k = \mathcal{O}\left(k^{-d-1}\right)$ when $k \gg d$ and $\mu_k = \mathcal{O}\left(d^{-k+1}\right)$ when $d \gg k$.

Remark 3.5. In the above theorem, the coefficients μ_k are actually different eigenvalues of the integral operator L_κ on $L^2_{\tau_d}(\mathbb{S}^d)$ defined by

$$L_\kappa(f)(\mathbf{y}) = \int_{\mathbb{S}^d} \kappa(\mathbf{x},\mathbf{y})f(\mathbf{x})d\tau_d(\mathbf{x}), \quad f \in L^2_{\tau_d}(\mathbb{S}^d),$$

where τ_d is the uniform probability measure on unit sphere $\mathbb{S}^d$. Therefore the λ_{r_k} in Theorem 3.2 is just μ_k given in Theorem 3.4 when τ_d is uniform distribution.

Remark 3.6. In Vempala and Wilmes (2018), they consider two layers neural networks with sigmoid activation function, and present explicit order of $m = (d+1)^{\mathcal{O}(k)\text{poly}\frac{\|g\|_2}{\epsilon}}$ and iteration times $T = (d+1)^{\mathcal{O}(k)\log\frac{\|g\|_2}{\epsilon}}$ to achieve $\epsilon_0 + \epsilon$ error by restricting that $\left\|g - g^{(\leqslant k)}\right\|_2 \leqslant \epsilon_0$. Another highly related work is Bietti and Mairal (2019), which gives $\mu_k = \mathcal{O}(k^{-d-1})$. The order of eigenvalues we present appears as $\mu_k = \min(\mathcal{O}(k^{-d-1}), \mathcal{O}(d^{-k+1}))$. This is better when $d \gg k$, which is closer to the practical setting.

3.3 Explicit Convergence Rate for Uniformly Distributed Data

In this section, we combine our results in above two subsections and give explicit convergence rate for uniformly distributed data on the unit sphere. From the following corollaries, we can see how spectral bias is revealed.

Corollary 3.7. Suppose that $k \gg d$, the sample $\{\mathbf{x}_i\}_{i=1}^n$ follows the uniform distribution τ_d on the unit sphere $\mathbb{S}^d$ and $|\phi_j(\mathbf{x})| \leqslant M$ for $j \in [r_k]$. For any $\epsilon, \delta > 0$ and integer k, if $n \geqslant \widetilde{\Omega}(\max\{\epsilon^{-1}, k^{d+1}, \epsilon^{-2}M^2r_k^2\})$, $m \geqslant \widetilde{\Omega}(\text{poly}(T, k^{-d-1}, \epsilon^{-1}))$, then with probability at least $1-\delta$, Algorithm 1 with $\eta = \widetilde{\mathcal{O}}((m\theta^2)^{-1})$, $\theta = \epsilon/16$ satisfies

$$n^{-1/2}\cdot\|\mathbf{V}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 \leqslant 2\left(1 - \mathcal{O}\left(k^{-d-1}\right)\right)^T \cdot n^{-1/2}\cdot\|\mathbf{V}_{r_k}^\top\mathbf{y}\|_2 + \epsilon.$$

Corollary 3.8. Suppose that $d \gg k$, the sample $\{\mathbf{x}_i\}_{i=1}^n$ follows the uniform distribution τ_d on the unit sphere $\mathbb{S}^d$ and $|\phi_j(\mathbf{x})| \leqslant M$ for $j \in [r_k]$. For any $\epsilon, \delta > 0$ and integer k, if $n \geqslant \widetilde{\Omega}(\max\{\epsilon^{-1}, d^{k-1}, \epsilon^{-2}M^2r_k^2\})$, $m \geqslant \widetilde{\Omega}(\text{poly}(T, d^{-k+1}, \epsilon^{-1}))$, then with probability at least $1-\delta$, Algorithm 1 with $\eta = \widetilde{\mathcal{O}}((m\theta^2)^{-1})$, $\theta = \epsilon/16$ satisfies

$$n^{-1/2}\cdot\|\mathbf{V}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 \leqslant 2\left(1 - \mathcal{O}\left(d^{-k+1}\right)\right)^T \cdot n^{-1/2}\cdot\|\mathbf{V}_{r_k}^\top\mathbf{y}\|_2 + \epsilon.$$

Remark 3.9. Here we give a further explanation for these corollaries: when $k = 1$, the results imply that the convergence rates would be controlled by the first eigenvalue of the integral operator of Neural Tangent Kernel. The directions would be degree 1 spherical harmonics. When $k = 2$, we know that the convergence rate of this algorithm is controlled by the second largest eigenvalue of the integral operator and the direction is the set of degree 2 spherical harmonics. The convergence rate is clearly slower than that when $k = 1$. And k can go to infinity with the same phenomenon that the convergence rate would be slower when we project the target function to more complex basis. In this way we give exact illustration how spectral bias is formed.

4 Experiments

In this section we illustrate our results by training neural networks on synthetic data. Across all tasks, we train a two-layer hidden neural networks with 4096 neurons and initialize it exactly as defined in the setup. The optimization method is vanilla full gradient descent. We sample 1000 training data which is uniformly sampled from the unit sphere in $\mathbb{R}^{10}$.

4.1 LEARNING COMBINATION OF SPHERICAL HARMONICS

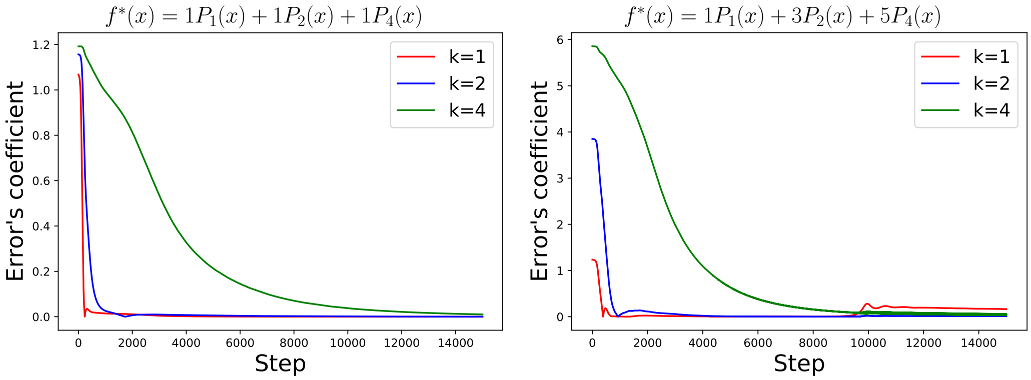
First, we show a result when the target function is exactly linear combination of spherical harmonics. The target function is explicitly defined as

$$f^*(x) = a_1 \times P_1(\langle \eta_1, x\rangle) + a_2 \times P_2(\langle \eta_2, x\rangle) + a_4 \times P_4(\langle \eta_4, x\rangle),$$

where the $P_k(t)$ is the Gegenbauer polynomial. Note that according to the addition formula $\sum_{j=1}^{N(d,k)} Y_{k,j}(\mathbf{x})Y_{k,j}(\mathbf{y}) = N(d,k)P_k(\langle \mathbf{x}, \mathbf{y}\rangle)$, every normalized Gegenbauer polynomial is a spherical harmonic, so $f^*(x)$ is a linear combination of spherical harmonics of order 1,2 and 4.The odd-order Gegenbauer polynomial is omitted because the spectral analysis showed that $\mu_k = 0, k = 3, 5, 7 \ldots$

We mainly focus on how fast different components of the residual function $f^*(x) - f_{\mathbf{W}^{(t)}}(x)$ descend during training. The coefficient for a give component is calculated by integrating along the spherical harmonics which is $\int_{\mathbb{S}^d} f^*(x)P_k(\langle \eta, x\rangle)d\tau(x)$. By Nystrom method we discretize this measure to

$$\hat{a}_k = \frac{\sum_{i=1}^{n} f^*(x_i)P_k(\langle \eta, x_i\rangle)}{\sum_{i=1}^{n} P_k^2(\langle \eta, x_i\rangle)}$$



(a) components with the same scale

(b) components with different scale

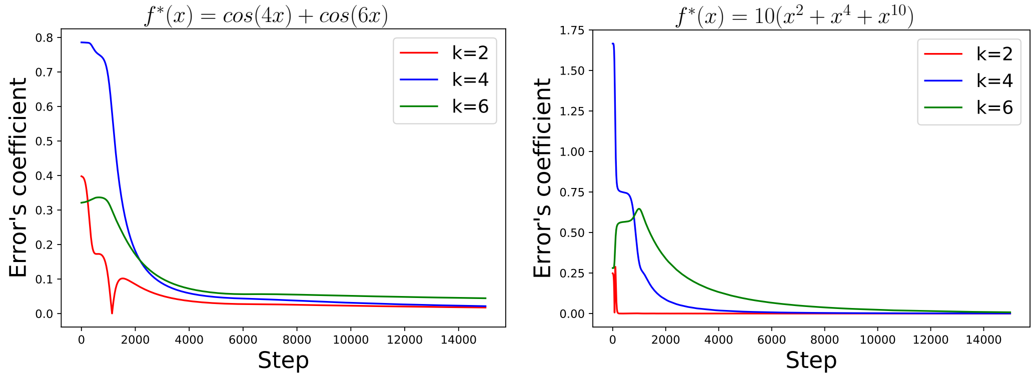
Figure 1: Convergence curve for different component. (a) shows the curve when the target function have different component with the same scale. (b) shows the curve when the higner-order components have larger scale. Both illustrate that the lower-order components are learned first.

Figure 1 shows that the convergence rates of different components are exactly predicted by our theory in a qualitative sense. At the beginning of training, the lowest frequency($k = 1$) are converging to zero fast and then the second lowest($k = 2$). The last frequency is converged latest. Following the setting of Rahaman et al. (2018) we assign high frequencies a larger scale, expecting that larger scale will introduce a better descending speed. Still, the low frequencies are regressed first.

4.2 LEARNING FUNCTIONS OF SIMPLE FORM

Apart from the synthesized low frequency function, we also showed the dynamics of normal functions' projection to $P_k(x)$. These functions, though in a simple form, are composed of almost all frequencies. In this subsection we further show our results still apply when all frequencies exists. The target function is given as $f^*(x) = \sum_i \cos(a_i\langle \eta, x\rangle)$ or $f^*(x) = \sum_i \langle \eta, x\rangle^{p_i}$, where η is a fixed unit vector. The coefficient of given components is calculated in the same way as in 4.1.

Figure 2 shows that even for arbitrarily chosen functions of simple form, the networks can still first learn the low frequency components of the target function. Notice that early in training not all the curves may descend, we believe this is due to the unseen components' influence on the gradient. Again, as the training proceeds, the convergence is controlled at the predicted rate.



(a) cosine function along one direction

(b) even polynomial along one direction

Figure 2: Convergence curve for different component. (a) shows the curve of a trigonometric function. (b) shows the curve of a polynomial with even degrees. Both exhibits similar tendency as combination of spherical harmonics.

Remark 4.1. The reason why we only use cosine function and even polynomial is that the only odd basis function with non-zero eigenvalue is $P_1(x)$. To show a general tendency it is better to restrict the target function in the even function space.

5 Conclusion and Discussion

In this paper, we give theoretical justification for spectral bias through a detailed analysis of the convergence behavior of two-layer neural networks with ReLU activation function. We show that the convergence of gradient descent in different directions depends on the corresponding eigenvalues and essentially exhibits different convergence rates. We show Mercer decomposition of Neural Tangent Kernel and give explicit order of eigenvalues of integral operator with respect to the Neural Tangent Kernel when the data is uniformly distributed on the unit sphere $\mathbb{S}^d$. Combined with the convergence analysis, we give exact order of convergence rate on different directions. We also conduct experiments on synthetic data to support our theoretical work.

So far, we have considered the upper bound for convergence with respect to low frequency components and present comprehensive theorem to explain the spectral bias. One desired improvement is to give the lower bound of convergence with respect to high frequency components, which is essential to establish tighter characterization of spectral-biased optimization. Another potential improvement is to generalize the result to multi-layer neural networks, which might require different techniques since our analysis heavily rely on exactly computing the eigenvalues of the neural tangent kernel.

References

Allen-Zhu, Z., Li, Y. and Song, Z. (2018). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962* .

Andoni, A., Panigrahy, R., Valiant, G. and Zhang, L. (2014). Learning polynomials with neural networks. In *International Conference on Machine Learning*.

Arora, S., Du, S. S., Hu, W., Li, Z. and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584* .

Atkinson, K. and Han, W. (2012). *Spherical harmonics and approximations on the unit sphere: an introduction*, vol. 2044. Springer Science & Business Media.

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research* **18** 1–53.

BASRI, R., JACOBS, D., KASTEN, Y. and KRITCHMAN, S. (2019). The convergence rate of neural networks for learned functions of different frequencies. *arXiv preprint arXiv:1906.00425* .

BIETTI, A. and MAIRAL, J. (2019). On the inductive bias of neural tangent kernels. *arXiv preprint arXiv:1905.12173* .

CAO, Y. and GU, Q. (2019a). Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210* .

CAO, Y. and GU, Q. (2019b). A generalization theory of gradient descent for learning over-parameterized deep relu networks. *arXiv preprint arXiv:1902.01384* .

CHO, Y. and SAUL, L. K. (2009). Kernel methods for deep learning. In *Advances in neural information processing systems*.

COLLOBERT, R. and WESTON, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. ACM.

DU, S. S., LEE, J. D., LI, H., WANG, L. and ZHAI, X. (2018a). Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804* .

DU, S. S., ZHAI, X., POCZOS, B. and SINGH, A. (2018b). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054* .

FRYE, C. and EFTHIMIOU, C. J. (2012). Spherical harmonics in p dimensions. *arXiv preprint arXiv:1205.3548* .

GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018a). Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246* .

GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018b). Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468* .

HE, K., ZHANG, X., REN, S. and SUN, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

HINTON, G., DENG, L., YU, D., DAHL, G., MOHAMED, A.-R., JAITLY, N., SENIOR, A., VANHOUCKE, V., NGUYEN, P., KINGSBURY, B. ET AL. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine* **29**.

JACOT, A., GABRIEL, F. and HONGLER, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572* .

LI, Y. and LIANG, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. *arXiv preprint arXiv:1808.01204* .

NAKKIRAN, P., KAPLUN, G., KALIMERIS, D., YANG, T., EDELMAN, B. L., ZHANG, F. and BARAK, B. (2019). Sgd on neural networks learns functions of increasing complexity. *arXiv preprint arXiv:1905.11604* .

RAHAMAN, N., BARATIN, A., ARPIT, D., DRAXLER, F., LIN, M., HAMPRECHT, F. A., BENGIO, Y. and COURVILLE, A. (2018). On the spectral bias of neural networks. *arXiv preprint arXiv:1806.08734* .

ROSASCO, L., BELKIN, M. and VITO, E. D. (2010). On learning with integral operators. *Journal of Machine Learning Research* **11** 905–934.

SMALE, S. and ZHOU, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive approximation* **26** 153–172.

SMALE, S. and ZHOU, D.-X. (2009). Geometry on probability spaces. *Constructive Approximation* **30** 311.

SOUDRY, D., HOFFER, E. and SREBRO, N. (2017). The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345* .

SU, L. and YANG, P. (2019). On learning over-parameterized neural networks: A functional approximation prospective. *arXiv preprint arXiv:1905.10826* .

VEMPALA, S. and WILMES, J. (2018). Gradient descent for one-hidden-layer neural networks: Polynomial convergence and sq lower bounds. *arXiv preprint arXiv:1805.02677* .

XU, Z. J. (2018). Understanding training and generalization in deep learning by fourier analysis. *arXiv preprint arXiv:1808.04295* .

YANG, G. and SALMAN, H. (2019). A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599* .

ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530* .

ZOU, D., CAO, Y., ZHOU, D. and GU, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888* .

ZOU, D. and GU, Q. (2019). An improved analysis of training over-parameterized deep neural networks. *arXiv preprint arXiv:1906.04688* .

A Review on spherical harmonics

In this section, we give a brief review on relevant concepts in spherical harmonics. For more detials, see Bach (2017), Bietti and Mairal (2019), Frye and Efthimiou (2012) and Atkinson and Han (2012) for references.

We consider the unit sphere $\mathbb{S}^d = \left\{\mathbf{x} \in \mathbb{R}^{d+1} : ||\mathbf{x}|| = 1\right\}$, whose surface area is given by $\omega_d = 2\pi^{(d+1)/2}/\Gamma((d+1)/2)$ and denote τ_d the uniform measure on the sphere.

For any $k \geqslant 1$, we consider a set of spherical harmonics $\left\{Y_{k,j} : \mathbb{S}^d \to \mathbb{R} | 1 \leqslant j \leqslant N(d,k) = \frac{2k+d-1}{k}\binom{k+d-2}{d-1}\right\}$. They form an orthonormal basis and satisfy the following equation $\langle Y_{ki}, Y_{sj}\rangle_{\mathbb{S}^d} = \int_{\mathbb{S}^d} Y_{ki}(x) Y_{sj}(x) d\tau_d(x) = \delta_{ij}\delta_{sk}$. Moreover, since they are homogeneous functions of degree k, it is clear that for any $Y_k(x)$, this harmonics has the same parity as k.

We have the addition formula

$$\sum_{j=1}^{N(d,k)} Y_{k,j}(\mathbf{x}) Y_{k,j}(\mathbf{y}) = N(d,k) P_k(\langle \mathbf{x}, \mathbf{y}\rangle), \tag{A.1}$$

where $P_k(t)$ is the Legendre polynomial of degree k in $d+1$ dimensions, explicitly given by (Rodrigues' formula)

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

We can also see that $P_k(t)$, the Legendre polynomial of degree k shares the same parity with k. By the orthogonality and the addition formula (A.1) we have,

$$\int_{\mathbb{S}^d} P_j(\langle \mathbf{w}, \mathbf{x}\rangle) P_k(\langle \mathbf{w}, \mathbf{y}\rangle) d\tau_d(\mathbf{w}) = \frac{\delta_{jk}}{N(d,k)} P_k(\langle \mathbf{x}, \mathbf{y}\rangle). \tag{A.2}$$

Further we have the recurrence relation for the Legendre polynomials,

$$tP_k(t) = \frac{k}{2k+d-1} P_{k-1}(t) + \frac{k+d-1}{2k+d-1} P_{k+1}(t), \tag{A.3}$$

for $k \geqslant 1$ and $tP_0(t) = P_1(t)$ for $k = 0$.

The Hecke-Funk formula is given for a spherical harmonic Y_k of degree k

$$\int_{\mathbb{S}^d} f(\langle \mathbf{x}, \mathbf{y}\rangle) Y_k(\mathbf{y}) dt'_d(\mathbf{y}) = \frac{\omega_{d-1}}{\omega_d} Y_k(\mathbf{x}) \int_{-1}^{1} f(t) P_k(t) (1-t^2)^{(d-2)/2} dt.$$

B Proof of Main Theorems

B.1 Proof of Theorem 3.2

In this section we give the proof of Theorem 3.2. We first introduce the following definitions and notations.

Define $\mathbf{K}^{(0)} = m^{-1}(\langle \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_j)\rangle)_{n\times n}$, $\mathbf{K}^{(\infty)} = (\kappa(\mathbf{x}_i, \mathbf{x}_j))_{n\times n} = \lim_{m\to\infty} \mathbf{K}^{(0)}$. Let $\{\widehat{\lambda}_i\}_{i=1}^n$, $\widehat{\lambda}_1 \geqslant \cdots \geqslant \widehat{\lambda}_n$ be the eigenvalues of $n^{-1}\mathbf{K}^{\infty}$, and $\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_n$ be the corresponding eigenvectors. Set $\widehat{\mathbf{V}}_{r_k} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_{r_k})$, $\widehat{\mathbf{V}}_{r_k}^{\perp} = (\widehat{\mathbf{v}}_{r_k+1}, \ldots, \widehat{\mathbf{v}}_n)$.

For notation simplicity, we denote $\nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i) = [\nabla_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{x}_i)]\big|_{\mathbf{W}=\mathbf{W}^{(0)}}$, $\nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i) = [\nabla_{\mathbf{W}_l} f_{\mathbf{W}}(\mathbf{x}_i)]\big|_{\mathbf{W}=\mathbf{W}^{(0)}}$, $l = 1, 2$.

The following lemma is a direct application of Proposition 1 in Smale and Zhou (2009) or Proposition 10 in Rosasco et al. (2010).

Lemma B.1. For any $\delta > 0$, with probability at least $1-\delta$,

$$|\lambda_i - \widehat{\lambda}_i| \leqslant \mathcal{O}(\sqrt{\log(1/\delta)/n}).$$

The following lemma is partly summarized from the proof of equation (44) in Su and Yang (2019).

Lemma B.2. Suppose that $|\phi_i(\mathbf{x})| \leqslant M$ for all $\mathbf{x} \in S^{d-1}$. There exist absolute constants $C, C', c'' > 0$, such that for any $\delta > 0$ and integer k with $r_k \leqslant n$, if $n \geqslant C(\lambda_{r_k} - \lambda_{r_k+1})^{-2}\log(1/\delta)$, then with probability at least $1-\delta$,

$$\|\mathbf{V}_{r_k}^\top \widehat{\mathbf{V}}_{r_k}^\perp\|_F \leqslant C' \frac{1}{\lambda_{r_k} - \lambda_{r_k+1}} \cdot \sqrt{\frac{\log(1/\delta)}{n}},$$
$$\|\mathbf{V}_{r_k}\mathbf{V}_{r_k}^\top - \widehat{\mathbf{V}}_{r_k}\widehat{\mathbf{V}}_{r_k}^\top\|_2 \leqslant C''\bigg[\frac{1}{(\lambda_{r_k+1} - \lambda_{r_k})^2}\frac{\log(1/\delta)}{n} + Mr_k\sqrt{\frac{\log(r_k/\delta)}{n}}\bigg].$$

Lemma B.3. Suppose that the iterates of gradient descent $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(t)}$ are inside the ball $\mathcal{B}(\mathbf{W}^{(0)}, \omega)$. If $\omega \leqslant \mathcal{O}([\log(m)]^{-3/2})$, then with probability at least $1 - \mathcal{O}(n)\cdot\exp[-\Omega(m\omega^{2/3})]$,

$$\mathbf{y} - \widehat{\mathbf{y}}^{(t'+1)} = [\mathbf{I} - (\eta m\theta^2/n)\mathbf{K}^\infty](\mathbf{y} - \widehat{\mathbf{y}}^{(t')}) + \mathbf{e}^{(t)},\ \|\mathbf{e}^{(t')}\|_2 \leqslant \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2)\cdot\|\mathbf{y} - \widehat{\mathbf{y}}^{(t')}\|_2$$

for all $t' = 0, \ldots, t-1$, where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\widehat{\mathbf{y}}^{(t')} = \theta\cdot(f_{\mathbf{W}^{(t')}}(\mathbf{x}_1), \ldots, f_{\mathbf{W}^{(t')}}(\mathbf{x}_n))^\top$.

Lemma B.4. Suppose that the iterates of gradient descent $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(t)}$ are inside the ball $\mathcal{B}(\mathbf{W}^{(0)}, \omega)$. If $\omega \leqslant \widetilde{\mathcal{O}}(\min\{[\log(m)]^{-3/2}, \lambda_{r_k}^{-3}, (\eta m)^{-3}\})$ and $n \geqslant \widetilde{\mathcal{O}}(\lambda_{r_k}^{-2})$, then with probability at least $1 - \mathcal{O}(n)\cdot\exp[-\Omega(m\omega^{2/3})]$

$$\|(\widehat{\mathbf{V}}_{r_k}^\perp)^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(t')})\|_2 \leqslant \|(\widehat{\mathbf{V}}_{r_k}^\perp)^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(0)})\|_2 + t'\cdot\omega^{1/3}\eta m\theta^2\cdot\sqrt{n}\cdot\widetilde{\mathcal{O}}(1 + \omega\sqrt{m}) \tag{B.1}$$
$$\begin{aligned}\|\widehat{\mathbf{V}}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(t')})\|_2 &\leqslant \sqrt{n}\cdot(1 - \eta m\theta^2\lambda_{r_k}/2)^{t'} + t'\lambda_{r_k}^{-1}\cdot\omega^{2/3}\eta m\theta^2\cdot\sqrt{n}\cdot\widetilde{\mathcal{O}}(1 + \omega\sqrt{m})\\ &\quad + \lambda_{r_k}^{-1}\cdot\widetilde{\mathcal{O}}(\omega^{1/3})\cdot\|(\widehat{\mathbf{V}}_{r_k}^\perp)^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(0)})\|_2\end{aligned} \tag{B.2}$$
$$\begin{aligned}\|\mathbf{y} - \widehat{\mathbf{y}}^{(t')}\|_2 &\leqslant \sqrt{n}\cdot(1 - \eta m\theta^2\lambda_{r_k}/2)^{t'} + \widetilde{\mathcal{O}}((\eta m\theta^2\lambda_{r_k})^{-1})\cdot\|(\widehat{\mathbf{V}}_{r_k}^\perp)^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(0)})\|_2\\ &\quad + \lambda_{r_k}^{-1}t'\omega^{1/3}\cdot\sqrt{n}\cdot\widetilde{\mathcal{O}}(1 + \omega\sqrt{m})\end{aligned} \tag{B.3}$$

for all $t' = 0, \ldots, t-1$.

Now we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. Define $\omega = \overline{C}T/(\lambda_{r_k}\sqrt{m})$ for some small enough absolute constant $\overline{C}$. Then by union bound, as long as $m \geqslant$, the conditions on ω given in Lemmas D.2, D.4, D.5, B.3 and B.4 are all satisfied.

We first show that all the iterates $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(T)}$ are inside the ball $\mathcal{B}(\mathbf{W}^{(0)}, \omega)$. We prove this result by inductively show that $\mathbf{W}^{(t)} \in \mathcal{B}(\mathbf{W}^{(0)}, \omega)$, $t = 0, \ldots, T$. First of all, it is clear that $\mathbf{W}^{(0)} \in \mathcal{B}(\mathbf{W}^{(0)}, \omega)$. Suppose that $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(t)} \in \mathcal{B}(\mathbf{W}^{(0)}, \omega)$. Then the results of Lemmas D.2, D.4, D.5, B.3 and B.4 hold for $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(t)}$. Denote $\mathbf{u}^{(t)} = \mathbf{y} - \widehat{\mathbf{y}}^{(t)}$, $t \in T$. Then we have

$$\begin{aligned}\|\mathbf{W}_l^{(t+1)} - \mathbf{W}_l^{(0)}\|_F &\leqslant \sum_{t'=0}^{t}\|\mathbf{W}_l^{(t'+1)} - \mathbf{W}_l^{(t')}\|_F\\ &= \eta\sum_{t'=0}^{t}\bigg\|\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta\cdot f_{\mathbf{W}^{(t)}}(\mathbf{x}_i))\cdot\theta\cdot\nabla_{\mathbf{W}_l}f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\bigg\|_F\\ &\leqslant \eta\theta\sum_{t'=0}^{t}\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta\cdot f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)|\cdot\|\nabla_{\mathbf{W}_l}f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\|_F\\ &\leqslant C_1\eta\theta\sqrt{m}\sum_{t'=0}^{t}\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta\cdot f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)|\\ &\leqslant C_1\eta\theta\sqrt{m/n}\sum_{t'=0}^{t}\|\mathbf{y} - \widehat{\mathbf{y}}^{(t')}\|_2,\end{aligned}$$

where the second inequality follows by Lemma D.4. By Lemma B.4, we have

$$\sum_{t'=0}^{t} \|\mathbf{y} - \widehat{\mathbf{y}}^{(t')}\|_2 \leqslant 2\sqrt{n}/(\eta m\theta^2\lambda_{r_k}) + \widetilde{\mathcal{O}}(T/(\eta m\theta^2\lambda_{r_k})) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(0)})\|_2 \\ + \lambda_{r_k}^{-1}T^2\omega^{1/3} \cdot \sqrt{n} \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m}).$$

It then follows by the choice $\omega = \overline{C}T/(\theta\lambda_{r_k}\sqrt{m})$, $\eta = \widetilde{\mathcal{O}}((m\theta^2\lambda_{r_k})^{-1})$, $\theta = \epsilon/16$ and the assumption $m \geqslant \widetilde{\mathcal{O}}(\text{poly}(\lambda_{r_k}, \epsilon^{-1}))$ that $\|\mathbf{W}_l^{(t+1)} - \mathbf{W}_l^{(0)}\|_F \leqslant \omega$, $l = 1, 2$. Therefore by induction, we see that $\mathbf{W}(0), \ldots, \mathbf{W}(T) \in \mathcal{B}(\mathbf{W}^{(0)}, \omega)$.

Applying Lemma B.4 then gives

$$n^{-1/2} \cdot \|\widehat{\mathbf{V}}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 \leqslant (1 - \eta m\theta^2\lambda_{r_k}/2)^T \cdot n^{-1/2} \cdot \|\widehat{\mathbf{V}}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(0)})\|_2 \\ + T\lambda_{r_k}^{-1} \cdot \omega^{2/3}\eta m\theta^2 \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m}) \\ + \lambda_{r_k}^{-1} \cdot \widetilde{\mathcal{O}}(\omega^{1/3}) \cdot n^{-1/2} \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(0)})\|_2.$$

Now by $\omega = \overline{C}T/(\lambda_{r_k}\sqrt{m})$, $\eta = \widetilde{\mathcal{O}}(\theta^2 m)^{-1}$ and the assumption that $m \geqslant m^* = \widetilde{\mathcal{O}}(\lambda_{r_k}^{-14} \cdot \epsilon^{-6})$, we obtain

$$n^{-1/2} \cdot \|\widehat{\mathbf{V}}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 \leqslant (1 - \lambda_{r_k})^T + \epsilon/16. \tag{B.4}$$

By Lemma 3.1, $\theta = \epsilon/16$ and the assumptions $n \geqslant \widetilde{\Omega}(\max\{\epsilon^{-1}(\lambda_{r_k} - \lambda_{r_k+1})^{-1}, \epsilon^{-2}M^2r_k^2\})$, the eigenvalues of $\mathbf{V}_{r_k}^\top\mathbf{V}_{r_k}$ are all between $1/\sqrt{2}$ and $\sqrt{2}$. Therefore by Lemma B.2 we have

$$\begin{aligned} \|\widehat{\mathbf{V}}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 &= \|\widehat{\mathbf{V}}_{r_k}\widehat{\mathbf{V}}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 \\ &\geqslant \|\mathbf{V}_{r_k}\mathbf{V}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 - \|(\mathbf{V}_{r_k}\mathbf{V}_{r_k}^\top - \widehat{\mathbf{V}}_{r_k}\widehat{\mathbf{V}}_{r_k}^\top)(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 \\ &\geqslant \|\widehat{\mathbf{V}}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2/\sqrt{2} - \mathcal{O}\bigg(\frac{1}{(\lambda_{r_k+1} - \lambda_{r_k})^2}\frac{\log(1/\delta)}{n} + Mr_k\sqrt{\frac{\log(r_k/\delta)}{n}}\bigg) \\ &\geqslant \|\widehat{\mathbf{V}}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2/\sqrt{2} - \epsilon\sqrt{n}/16. \end{aligned}$$

Similarly,

$$\|\widehat{\mathbf{V}}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 \leqslant \sqrt{2} \cdot \|\mathbf{V}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 + \epsilon\sqrt{n}/16 \leqslant \sqrt{2} \cdot \|\mathbf{V}_{r_k}^\top\mathbf{y}\|_2 + \epsilon\sqrt{n}/8$$

Plugging the above two inequalities into (B.4) gives

$$n^{-1/2} \cdot \|\mathbf{V}_{r_k}^\top(\mathbf{y} - \widehat{\mathbf{y}}^{(T)})\|_2 \leqslant 2(1 - \lambda_{r_k})^T \cdot n^{-1/2} \cdot \|\mathbf{V}_{r_k}^\top\mathbf{y}\|_2 + \epsilon.$$

This completes the proof. □

B.2 Proof of the Theorem 3.4

Proof of the Theorem 3.4. The idea of the proof is close to that of **Proposition 5** in (Bietti and Mairal, 2019) where they consider $k \gg d$ and we present a more general case including $k \gg d$ and $d \gg k$.

For any function $g : \mathbb{S}^d \to \mathbb{R}$, by denoting $g_0(\mathbf{x}) = \int_{\mathbb{S}^d} g(\mathbf{y})d\tau_d(\mathbf{y})$, it can be decomposed as

$$\begin{aligned} g(\mathbf{x}) = \sum_{k=0}^{\infty} g_k(\mathbf{x}) &= \sum_{k=0}^{\infty}\sum_{j=1}^{N(d,k)} \int_{\mathbb{S}^d} Y_{kj}(\mathbf{y})Y_{kj}(\mathbf{x})g(\mathbf{y})d\tau_d(\mathbf{y}) \\ &= \sum_{k=0}^{\infty} N(d,k)\int_{\mathbb{S}^d} g(\mathbf{y})P_k(\langle\mathbf{x},\mathbf{y}\rangle)d\tau_d(\mathbf{y}), \end{aligned} \tag{B.5}$$

where we project function g to spherical harmonics in the second equality and apply the addition equation in the last equality.

For a positive-definite dot-product kernel $\kappa(\mathbf{x}, \mathbf{x}') : \mathbb{S}^d \times \mathbb{S}^d \to \mathbb{R}$ which has the form $\kappa(\mathbf{x}, \mathbf{x}') =$

$\widehat{\kappa}(\langle \mathbf{x}, \mathbf{x}' \rangle)$ for $\widehat{\kappa} : [-1, 1] \to \mathbb{R}$, we can present a decomposition by (B.5) if we consider $g(\mathbf{x}) = \phi(\langle \mathbf{x}, \mathbf{z} \rangle)$ for $\mathbf{z} \in \mathbb{S}^d$ and $\phi : [-1, 1] \to \mathbb{R}$,

$$\begin{aligned} \kappa(\mathbf{x}, \mathbf{x}') &= \sum_{k=0}^{\infty} N(d,k) \int_{\mathbb{S}^d} \widehat{\kappa}(\langle \mathbf{y}, \mathbf{x}' \rangle) P_k(\langle \mathbf{y}, \mathbf{x} \rangle) d\tau_d(\mathbf{y}) \\ &= \sum_{k=0}^{\infty} N(d,k) \frac{\omega_{d-1}}{\omega_d} P_k(\langle \mathbf{x}, \mathbf{x}' \rangle) \int_{-1}^{1} \widehat{\kappa}(t) P_k(t) (1-t^2)^{(d-2)/2} dt, \end{aligned}$$

where we apply the Hecke-Funk formula and addition formula. By denoting $\lambda_k = (\omega_{d-1}/\omega_d) \int_{-1}^{1} \widehat{\kappa}(t) P_k(t)(1-t^2)^{(d-2)/2} dt$ and the addition formula, we have

$$\kappa(\mathbf{x}, \mathbf{x}') = \sum_{k=0}^{\infty} \mu_k N(d,k) P_k(\langle \mathbf{x}, \mathbf{x}' \rangle) = \sum_{k=0}^{\infty} \mu_k \sum_{j=1}^{N(p,k)} Y_{k,j}(\mathbf{x}) Y_{k,j}(\mathbf{x}'). \tag{B.6}$$

This formula (B.6) is the Mercer decomposition for the kernel function $\kappa(\mathbf{x}, \mathbf{x}')$ and μ_k is exactly the eigenvalue of the integral operator L_K on $L_2(\mathbb{S}^d)$ defined by

$$L_\kappa(f)(\mathbf{y}) = \int_{\mathbb{S}^d} \kappa(\mathbf{x}, \mathbf{y}) f(\mathbf{x}) d\tau_d(\mathbf{x}), \quad f \in L_2(\mathbb{S}^d).$$

By using same technique as $\kappa(\mathbf{x}, \mathbf{x}')$, we can derive a similar expression for $\sigma(\langle \mathbf{w}, \mathbf{x} \rangle) = \max\{\langle \mathbf{w}, \mathbf{x} \rangle, 0\}$ and $\sigma'(\langle \mathbf{w}, \mathbf{x} \rangle) = \mathbb{1}\{\langle \mathbf{w}, \mathbf{x} \rangle > 0\}$, since they are essentially dot-product function on $L_2(\mathbb{S}^d)$. We deliver the expression below without presenting proofs.

$$\sigma'(\langle \mathbf{w}, \mathbf{x} \rangle) = \sum_{k=0}^{\infty} \beta_{1,k} N(d,k) P_k(\langle \mathbf{w}, \mathbf{x} \rangle), \tag{B.7}$$

$$\sigma(\langle \mathbf{w}, \mathbf{x} \rangle) = \sum_{k=0}^{\infty} \beta_{2,k} N(d,k) P_k(\langle \mathbf{w}, \mathbf{x} \rangle), \tag{B.8}$$

where $\beta_{1,k} = (\omega_{d-1}/\omega_d) \int_{-1}^{1} \sigma(t) P_k(t)(1-t^2)^{(d-2)/2} dt$ and $\beta_{2,k} = (\omega_{d-1}/\omega_d) \int_{-1}^{1} \sigma'(t) P_k(t)(1-t^2)^{(d-2)/2} dt$. We add more comments on the values of $\beta_{1,k}$ and $\beta_{2,k}$. It has been pointed out in Bach (2017) that when $k > \alpha$ and when k and α have same parity, we have $\beta_{\alpha+1,k} = 0$. This is because the Legendre polynomial $P_k(t)$ is orthogonal to any other polynomials of degree less than k with respect to the density function $p(t) = (1-t^2)^{(d-2)/2}$. Then we clearly know that $\beta_{1,k} = 0$ for $k = 2j$ and $\beta_{2,k} = 0$ for $k = 2j+1$ with $j \in \mathbb{N}^+$.

For two kernel function defined in (2.2), we have

$$\begin{aligned} \kappa_1(\mathbf{x}, \mathbf{x}') &= \mathbb{E}_{\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})} \left[\sigma'(\langle \mathbf{w}, \mathbf{x} \rangle) \sigma'(\langle \mathbf{w}, \mathbf{x}' \rangle) \right] \\ &= \mathbb{E}_{\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})} \left[\sigma'(\langle \mathbf{w}/\|\mathbf{w}\|, \mathbf{x} \rangle) \sigma'(\langle \mathbf{w}/\|\mathbf{w}\|, \mathbf{x}' \rangle) \right] \\ &= \int_{\mathbb{S}^d} \sigma'(\langle \mathbf{v}, \mathbf{x} \rangle) \sigma'(\langle \mathbf{v}, \mathbf{x}' \rangle) d\tau_d(\mathbf{v}). \end{aligned} \tag{B.9}$$

The first equality holds because σ' is 0-homogeneous function and the second equality is true since the normalized direction of a multivariate Gaussian random variable satisfies uniform distribution on the unit sphere. Similarly we can derive

$$\kappa_2(\mathbf{x}, \mathbf{x}') = (d+1) \int_{\mathbb{S}^d} \sigma(\langle \mathbf{v}, \mathbf{x} \rangle) \sigma(\langle \mathbf{v}, \mathbf{x}' \rangle) d\tau_d(\mathbf{v}). \tag{B.10}$$

By combing (A.2), (B.7),(B.8), (B.9) and (B.10), we can get

$$\kappa_1(\mathbf{x}, \mathbf{x}') = \sum_{k=0}^{\infty} \beta_{1,k}^2 N(d,k) P_k(\langle \mathbf{x}, \mathbf{x}' \rangle), \tag{B.11}$$

and

$$\kappa_2(\mathbf{x}, \mathbf{x}') = (d+1) \sum_{k=0}^{\infty} \beta_{2,k}^2 N(d,k) P_k(\langle \mathbf{x}, \mathbf{x}' \rangle). \tag{B.12}$$

Comparing (B.6), (B.11) and (B.12), we can easily show that

$$\mu_{1,k} = \beta_{1,k}^2 \text{ and } \mu_{2,k} = (d+1)\beta_{2,k}^2. \tag{B.13}$$

In Bach (2017), explicit expressions for $\beta_{1,k}$ and $\beta_{2,k}$ for $k \geqslant \alpha + 1$ are presented by

$$\beta_{\alpha+1,k} = \frac{d-1}{2\pi} \frac{\alpha!(-1)^{(k-1-\alpha)/2}}{2^k} \frac{\Gamma(d/2)\Gamma(k-\alpha)}{\Gamma(\frac{k-\alpha+1}{2})\Gamma(\frac{k+d+\alpha+1}{2})}.$$

By Stirling formula $\Gamma(x) \approx x^{x-1/2}e^{-x}\sqrt{2\pi}$, we have following expression of $\beta_{\alpha+1,k}$ for $k \geqslant \alpha + 1$

$$\beta_{\alpha+1,k} = C(\alpha) \frac{(d-1)d^{\frac{d-1}{2}}(k-\alpha)^{k-\alpha-\frac{1}{2}}}{(k-\alpha+1)^{\frac{k-\alpha}{2}}(k+d+\alpha+1)^{\frac{k+d+\alpha}{2}}} = \mathcal{O}\left(d^{\frac{d+1}{2}} k^{\frac{k-\alpha-1}{2}} (k+d)^{\frac{-k-d-\alpha}{2}}\right)$$

where $C(\alpha) = \frac{\sqrt{2}\alpha!}{2\pi} \exp\{\alpha + 1\}$. Also $\beta_{\alpha+1,0} = \frac{d-1}{4\pi} \frac{\Gamma(\frac{\alpha+1}{2})\Gamma(\frac{d}{2})}{\Gamma(\frac{d+\alpha+2}{2})}$, $\beta_{1,1} = \frac{d-1}{2d\pi}$ and $\beta_{2,1} = \frac{d-1}{4\pi d} \frac{\Gamma(\frac{1}{2})\Gamma(\frac{d+2}{2})}{\Gamma(\frac{d+3}{2})}$. Thus combine (B.13) we know that $\mu_{\alpha+1,k} = \mathcal{O}\left(d^{d+1+\alpha} k^{k-\alpha-1} (k+d)^{-k-d-\alpha}\right)$.
By considering (A.3) and (B.6), we have

$$\mu_0 = \mu_{1,1} + 2\mu_{2,0},$$

$$\mu_{k'} = 0, \quad k' = 2j+1,\ j \in \mathbb{N}^+,$$

and

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

for $k \geqslant 1$ and $k \neq k'$. From the discussion above, we thus know exactly that for $k \geqslant 1$

$$\lambda_k = \mathcal{O}\left(\max\left\{d^{d+1}k^{k-1}(k+d)^{-k-d}, d^{d+1}k^k(k+d)^{-k-d-1}, d^{d+2}k^{k-2}(k+d)^{-k-d-1}\right\}\right). \tag{B.14}$$

This finishes the proof. □

C Proof of Technical Lemmas

C.1 Proof of Lemma B.2

Proof of Lemma B.2. The first inequality directly follows by equation (44) in Su and Yang (2019). To prove the second bound, we write $\mathbf{V}_{r_k} = \widehat{\mathbf{V}}_{r_k}\mathbf{A} + \widehat{\mathbf{V}}_{r_k}^{\perp}\mathbf{B}$, where $\mathbf{A} \in \mathbb{R}^{r_k \times r_k}$, $\mathbf{B} \in \mathbb{R}^{(n-r_k)\times r_k}$. Let $\xi_1 =\leqslant C'(\lambda_{r_k} - \lambda_{r_k+1})^{-1} \cdot \sqrt{\log(1/\delta)/n}$, $\xi_2 = C'''\sqrt{M\log(r_k/\delta)/n}$ be the bounds given in the first inequality and Lemma 3.1. By the first inequality, we have

$$\|\mathbf{B}\|_F = \|\mathbf{B}^\top\|_F = \|\mathbf{V}_{r_k}^\top \widehat{\mathbf{V}}_{r_k}^{\perp}\|_F \leqslant \xi_1.$$

Moreover, since $\mathbf{V}_{r_k}^\top \mathbf{V}_{r_k} = \mathbf{A}^\top\mathbf{A} + \mathbf{B}^\top\mathbf{B}$, by Lemma 3.1 we have

$$\|\mathbf{A}\mathbf{A}^\top - \mathbf{I}\|_2 = \|\mathbf{A}^\top\mathbf{A} - \mathbf{I}\|_2 \leqslant \|\mathbf{V}_{r_k}^\top\mathbf{V}_{r_k} - \mathbf{I}\|_2 + \|\mathbf{B}^\top\mathbf{B}\|_2 \leqslant r_k\xi_2 + \xi_1^2.$$

Therefore

$$\begin{aligned}
\|\mathbf{V}_{r_k}\mathbf{V}_{r_k}^\top - \widehat{\mathbf{V}}_{r_k}\widehat{\mathbf{V}}_{r_k}^\top\|_2 &= \|\widehat{\mathbf{V}}_{r_k}\mathbf{A}\mathbf{A}^\top\widehat{\mathbf{V}}_{r_k}^\top + \widehat{\mathbf{V}}_{r_k}^{\perp}\mathbf{B}\mathbf{B}^\top(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top - \widehat{\mathbf{V}}_{r_k}\widehat{\mathbf{V}}_{r_k}^\top\|_2 \\
&\leqslant \|\widehat{\mathbf{V}}_{r_k}(\mathbf{A}\mathbf{A}^\top - \mathbf{I})\widehat{\mathbf{V}}_{r_k}^\top\|_2 + \|\widehat{\mathbf{V}}_{r_k}^{\perp}\mathbf{B}\mathbf{B}^\top(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top\|_2 \\
&= \|\mathbf{A}\mathbf{A}^\top - \mathbf{I}\|_2 + \|\mathbf{B}\mathbf{B}^\top\|_2 \\
&\leqslant r_k\xi_2 + 2\xi_1^2
\end{aligned}$$

Plugging in the definition of ξ_1 and ξ_2 completes the proof. □

C.2 Proof of Lemma B.3

Proof of Lemma B.3. The gradient descent update formula gives

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \frac{\eta}{n}\sum_{i=1}^{n}(y_i - \theta f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)) \cdot \theta \nabla_{\mathbf{W}} f_{\mathbf{W}^{(t)}}(\mathbf{x}_i). \tag{C.1}$$

For any $j \in [n]$, subtracting $\mathbf{W}^{(t)}$ and applying inner product with $\theta \nabla_{\mathbf{W}} f_{\mathbf{W}^{(t)}}(\mathbf{x}_j)$ on both sides gives

$$\theta \langle \nabla_{\mathbf{W}} f_{\mathbf{W}^{(t)}}(\mathbf{x}_j), \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)} \rangle = \frac{\eta\theta^2}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i^{(t)}) \cdot \langle \nabla_{\mathbf{W}} f_{\mathbf{W}^{(t)}}(\mathbf{x}_j), \nabla_{\mathbf{W}} f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \rangle.$$

Further rearranging terms then gives

$$y_j - (\widehat{\mathbf{y}}^{(t+1)})_j = y_j - (\widehat{\mathbf{y}}^{(t)})_j - \frac{\eta m \theta^2}{n}\sum_{i=1}^{n}(y_i - f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)) \cdot \mathbf{K}_{i,j}^{\infty} + I_{1,j,t} + I_{2,j,t} + I_{3,j,t}, \tag{C.2}$$

where

$$\begin{aligned} I_{1,j,t} &= -\frac{\eta\theta^2}{n}\sum_{i=1}^{n}(y_i - f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)) \cdot [\langle \nabla_{\mathbf{W}} f_{\mathbf{W}^{(t)}}(\mathbf{x}_j), \nabla_{\mathbf{W}} f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \rangle - m\mathbf{K}_{i,j}^{(0)}], \\ I_{2,j,t} &= -\frac{\eta m \theta^2}{n}\sum_{i=1}^{n}(y_i - f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)) \cdot (\mathbf{K}_{i,j}^{(0)} - \mathbf{K}_{i,j}^{\infty}), \\ I_{3,j,t} &= -\theta \cdot [f_{\mathbf{W}^{(t+1)}}(\mathbf{x}_j) - f_{\mathbf{W}^{(t)}}(\mathbf{x}_j) - \langle \nabla f_{\mathbf{W}^{(t)}}(\mathbf{x}_j), \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)} \rangle]. \end{aligned}$$

For $I_{1,j,t}$, by Lemma D.6, we have

$$|I_{1,j,t}| \leqslant \widetilde{\mathcal{O}}(\omega^{1/3}\eta m \theta^2) \cdot \frac{1}{n}\sum_{i=1}^{n}|y_i - f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)| \leqslant \widetilde{\mathcal{O}}(\omega^{1/3}\eta m \theta^2) \cdot \|\mathbf{y} - \widehat{\mathbf{y}}^{(t)}\|_2/\sqrt{n}.$$

For $I_{2,j,t}$, by Bernstein inequality and union bound, with probability at least $1 - \mathcal{O}(n^2) \cdot \exp(-\Omega(m\omega^{2/3}))$, we have

$$\left|\mathbf{K}_{i,j}^{\infty} - \mathbf{K}_{i,j}^{(0)}\right| \leqslant \mathcal{O}(\omega^{1/3})$$

for all $i, j \in [n]$. Therefore

$$|I_{2,j,t}| \leqslant \mathcal{O}(\omega^{1/3}\eta m \theta^2) \cdot \frac{1}{n}\sum_{i=1}^{n}|y_i - f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)| \leqslant \mathcal{O}(\omega^{1/3}\eta m \theta^2) \cdot \|\mathbf{y} - \widehat{\mathbf{y}}^{(t)}\|_2/\sqrt{n}.$$

For $I_{3,j,t}$, we have

$$\begin{aligned} I_{3,j,t} &\leqslant \widetilde{\mathcal{O}}(\omega^{1/3}\sqrt{m}\theta) \cdot \|\mathbf{W}_1^{(t+1)} - \mathbf{W}_1^{(t)}\|_2 \\ &\leqslant \widetilde{\mathcal{O}}(\omega^{1/3}\sqrt{m}\theta) \cdot \frac{\eta}{n}\sum_{i=1}^{n}|y_i - \theta f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)| \cdot \theta \cdot \|\nabla_{\mathbf{W}_1} f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\|_2 \\ &\leqslant \widetilde{\mathcal{O}}(\omega^{1/3}\eta m \theta^2) \cdot \frac{1}{n}\sum_{i=1}^{n}|y_i - \theta f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)| \\ &\leqslant \widetilde{\mathcal{O}}(\omega^{1/3}\eta m \theta^2) \cdot \|\mathbf{y} - \widehat{\mathbf{y}}^{(t)}\|_2/\sqrt{n}, \end{aligned}$$

where the first inequality follows by Lemmas D.2, the second inequality is obtained from (C.1), and the third inequality follows by Lemma D.4. Setting the j-th entry of $\mathbf{e}^{(t)}$ as $I_{1,j,t} + I_{2,j,t} + I_{3,j,t}$ and writing (C.2) into matrix form completes the proof. □

C.3 Proof of Lemma B.4

Proof of Lemma B.4. Denote $\mathbf{u}^{(t)} = \mathbf{y} - \widehat{\mathbf{y}}^{(t)}$, $t \in T$. Then we have

$$\begin{aligned}\|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(t'+1)}\|_2 &\leqslant \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top [\mathbf{I} - (\eta m\theta^2/n)\mathbf{K}^{\infty}]\mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot \|\mathbf{u}^{(t')}\|_2 \\ &\leqslant \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot \sqrt{n} \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m}),\end{aligned}$$

where the first inequality follows by Lemma B.3, and the second inequality follows by Lemma D.5. Therefore we have

$$\|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(t')}\|_2 \leqslant \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(0)}\|_2 + t' \cdot \omega^{1/3}\eta m\theta^2 \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m})$$

for $t' = 0, \ldots, t$. This completes the proof of (B.1). Similarly, we have

$$\begin{aligned}\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t'+1)}\|_2 &\leqslant \|\widehat{\mathbf{V}}_{r_k}^\top [\mathbf{I} - (\eta m\theta^2/n)\mathbf{K}^{\infty}]\mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot \|\mathbf{u}^{(t')}\|_2 \\ &\leqslant (1 - \eta m\theta^2\widehat{\lambda}_{r_k})\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot (\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(t')}\|_2) \\ &\leqslant (1 - \eta m\theta^2\lambda_{r_k}/2)\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(t')}\|_2 \\ &\leqslant (1 - \eta m\theta^2\lambda_{r_k}/2)\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + t' \cdot (\omega^{1/3}\eta m\theta^2)^2 \cdot \sqrt{n} \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m}) \\ &\quad + \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(0)}\|_2\end{aligned}$$

for $t' = 0, \ldots, t-1$, where the third inequality is by Lemma B.1 and the assumption that $\omega \leqslant \widetilde{\mathcal{O}}(\lambda_{r_k}^{-3})$, and the fourth inequality is by (B.1). Therefore we have

$$\begin{aligned}\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 &\leqslant (1 - \eta m\theta^2\lambda_{r_k}/2)^{t'}\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(0)}\|_2 + (\eta m\theta^2\lambda_{r_k}/2)^{-1} \cdot (\omega^{1/3}\eta m\theta^2)^2 \cdot \sqrt{n} \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m}) \\ &\quad + (\eta m\theta^2\lambda_{r_k}/2)^{-1} \cdot \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(0)}\|_2 \\ &= (1 - \eta m\theta^2\lambda_{r_k}/2)^{t'}\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(0)}\|_2 + t'\lambda_{r_k}^{-1} \cdot \omega^{2/3}\eta m\theta^2 \cdot \sqrt{n} \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m}) \\ &\quad + \lambda_{r_k}^{-1} \cdot \widetilde{\mathcal{O}}(\omega^{1/3}) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(0)}\|_2 \\ &\leqslant (1 - \eta m\theta^2\lambda_{r_k}/2)^{t'}\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(0)}\|_2 + t'\lambda_{r_k}^{-1} \cdot \omega^{2/3}\eta m\theta^2 \cdot \sqrt{n} \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m}) \\ &\quad + \lambda_{r_k}^{-1} \cdot \widetilde{\mathcal{O}}(\omega^{1/3}) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(0)}\|_2.\end{aligned}$$

This completes the proof of (B.2). Finally, for (B.3), by assumption we have $\omega^{1/3}\eta m\theta^2 \leqslant \widetilde{\mathcal{O}}(1)$. Therefore

$$\begin{aligned}\|\mathbf{u}^{(t'+1)}\|_2 &\leqslant \|[\mathbf{I} - (\eta m\theta^2/n)\mathbf{K}^{\infty}]\widehat{\mathbf{V}}_{r_k}\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + \|[\mathbf{I} - (\eta m\theta^2/n)\mathbf{K}^{\infty}]\widehat{\mathbf{V}}_{r_k}^{\perp}(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(t')}\|_2 \\ &\quad + \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot \|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(t')}\|_2 \\ &\leqslant (1 - \eta m\theta^2\widehat{\lambda}_{r_k})\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(\omega^{1/3}\eta m\theta^2) \cdot \|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(1) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(t')}\|_2 \\ &\leqslant (1 - \eta m\theta^2\lambda_{r_k}/2)\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(1) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(t')}\|_2 \\ &\leqslant (1 - \eta m\theta^2\lambda_{r_k}/2)\|\widehat{\mathbf{V}}_{r_k}^\top \mathbf{u}^{(t')}\|_2 + \widetilde{\mathcal{O}}(1) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(0)}\|_2 + t'\omega^{1/3}\eta m\theta^2\sqrt{n} \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m})\end{aligned}$$

for $t' = 0, \ldots, t-1$, where the third inequality is by Lemma B.1 and the assumption that $\omega \leqslant \widetilde{\mathcal{O}}(\lambda_{r_k}^{-3})$, and the fourth inequality follows by (B.1). Therefore we have

$$\|\mathbf{u}^{(t')}\|_2 \leqslant \sqrt{n} \cdot (1 - \eta m\theta^2\lambda_{r_k}/2)^{t'} + \widetilde{\mathcal{O}}((\eta m\theta^2\lambda_{r_k})^{-1}) \cdot \|(\widehat{\mathbf{V}}_{r_k}^{\perp})^\top \mathbf{u}^{(0)}\|_2 + \lambda_{r_k}^{-1}t'\omega^{1/3}\sqrt{n} \cdot \widetilde{\mathcal{O}}(1 + \omega\sqrt{m}).$$

This finishes the proof. □

D Auxiliary Lemmas

In this section we list several auxiliary lemmas on the properties of over-parameterized neural networks we need in our proof of Theorem 3.2. These results are mostly summarized from Allen-Zhu et al. (2018) and Cao and Gu (2019a).

D.1 Auxiliary Lemmas

Denote

$$\begin{aligned}\mathbf{D}_i &= \mathrm{diag}\big(\mathbb{1}\{(\mathbf{W}_1\mathbf{x}_i)_1 > 0\}, \ldots, \mathbb{1}\{(\mathbf{W}_1\mathbf{x}_i)_m > 0\}\big),\\ \mathbf{D}_i^{(0)} &= \mathrm{diag}\big(\mathbb{1}\{(\mathbf{W}_1^{(0)}\mathbf{x}_i)_1 > 0\}, \ldots, \mathbb{1}\{(\mathbf{W}_1^{(0)}\mathbf{x}_i)_m > 0\}\big).\end{aligned}$$

Lemma D.1 (Allen-Zhu et al. (2018))**.** If $\omega \leqslant \mathcal{O}([\log(m)]^{-3/2})$, then with probability at least $1 - \mathcal{O}(n)\cdot\exp[-\Omega(m\omega^{2/3})]$,

$$\|\mathbf{D}_i - \mathbf{D}_i^{(0)}\|_0 \leqslant \mathcal{O}(\omega^{2/3}m)$$

for all $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \omega)$, $i \in [n]$.

Lemma D.2 (Cao and Gu (2019a))**.** There exists an absolute constant κ such that, with probability at least $1 - \mathcal{O}(n)\cdot\exp[-\Omega(m\omega^{2/3})]$ over the randomness of $\mathbf{W}^{(1)}$, for all $i \in [n]$ and $\mathbf{W}, \mathbf{W}' \in \mathcal{B}(\mathbf{W}^{(1)}, \omega)$ with $\omega \leqslant \kappa[\log(m)]^{-3/2}$, it holds uniformly that

$$|f_{\mathbf{W}'}(\mathbf{x}_i) - f_{\mathbf{W}}(\mathbf{x}_i) - \langle\nabla_{\mathbf{W}}f_{\mathbf{W}}(\mathbf{x}_i), \mathbf{W}' - \mathbf{W}\rangle| \leqslant \mathcal{O}\Big(\omega^{1/3}\sqrt{m\log(m)}\Big)\cdot\|\mathbf{W}_1' - \mathbf{W}_1\|_2.$$

Lemma D.3 (Cao and Gu (2019a))**.** For any $\delta > 0$, if $m \geqslant C\log(n/\delta)$ for a large enough absolute constant C, then with probability at least $1-\delta$, $|f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)| \leqslant \mathcal{O}(\sqrt{\log(n/\delta)})$ for all $i \in [n]$.

Lemma D.4 (Cao and Gu (2019a))**.** There exists an absolute constant C such that, with probability at least $1 - \mathcal{O}(n)\cdot\exp[-\Omega(m\omega^{2/3})]$, for all $i \in [n]$, $l \in [L]$ and $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(1)}, \omega)$ with $\omega \leqslant C[\log(m)]^{-3}$, it holds uniformly that

$$\|\nabla_{\mathbf{W}_l}f_{\mathbf{W}}(\mathbf{x}_i)\|_F \leqslant \mathcal{O}(\sqrt{m}).$$

The following lemma provides a uniform bound of the neural network function value over $\mathcal{B}(\mathbf{W}^{(0)}, \omega)$.

Lemma D.5. Suppose that $m \geqslant \Omega(\omega^{-2/3}\log(n/\delta))$ and $\omega \leqslant \mathcal{O}([\log((m))]^{-3})$. Then with probability at least $1-\delta$, $|f_{\mathbf{W}}(\mathbf{x}_i)| \leqslant \mathcal{O}(\sqrt{\log(n/\delta)} + \omega\sqrt{m})$ for all $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \omega)$ $i \in [n]$.

Lemma D.6. If $\omega \leqslant \mathcal{O}([\log(m)]^{-3/2})$, then with probability at least $1 - \mathcal{O}(n)\cdot\exp[-\Omega(m\omega^{2/3})]$,

$$\begin{aligned}&\|\nabla_{\mathbf{W}}f_{\mathbf{W}}(\mathbf{x}_i) - \nabla_{\mathbf{W}}f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\|_F \leqslant \mathcal{O}(\omega^{1/3}\sqrt{m}),\\ &|\langle\nabla_{\mathbf{W}}f_{\mathbf{W}}(\mathbf{x}_i), \nabla_{\mathbf{W}}f_{\mathbf{W}}(\mathbf{x}_j)\rangle - \langle\nabla_{\mathbf{W}}f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \nabla_{\mathbf{W}}f_{\mathbf{W}^{(0)}}(\mathbf{x}_j)\rangle| \leqslant \mathcal{O}(\omega^{1/3}m)\end{aligned}$$

for all $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \omega)$ and $i \in [n]$.

D.2 Proofs of Lemmas D.5 and D.6

Proof of Lemma D.5. By Lemmas D.2 and D.4, we have

$$\begin{aligned}|f_{\mathbf{W}}(\mathbf{x}_i) - f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)| &\leqslant \|\nabla_{\mathbf{W}_1}f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\|_F\|\mathbf{W}_1 - \mathbf{W}_1^{(0)}\|_F + \|\nabla_{\mathbf{W}_2}f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\|_F\|\mathbf{W}_2 - \mathbf{W}_2^{(0)}\|_F\\ &\quad + \mathcal{O}(\omega^{1/3}\sqrt{m\log(m)})\cdot\|\mathbf{W}_1 - \mathbf{W}_1^{(0)}\|_2\\ &\leqslant \mathcal{O}(\omega\sqrt{m}),\end{aligned}$$

where the last inequality is by the assumption $\omega \leqslant [\log(m)]^{-3}$. Applying triangle inequality and Lemma D.3 then gives

$$\begin{aligned}|f_{\mathbf{W}}(\mathbf{x}_i)| &\leqslant |f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)| + |f_{\mathbf{W}}(\mathbf{x}_i) - f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)| \leqslant \mathcal{O}(\sqrt{\log(n/\delta)}) + \widetilde{\mathcal{O}}(\omega\sqrt{m})\\ &= \mathcal{O}(\sqrt{\log(n/\delta)} + \omega\sqrt{m}),\end{aligned}$$

This completes the proof. □

Proof of Lemma D.6. By direct calculation, we have

$$\nabla_{\mathbf{W}_1}f_{\mathbf{W}^{(0)}}(\mathbf{x}_i) = \sqrt{m}\cdot\mathbf{D}_i^{(0)}\mathbf{W}_2^{(0)\top}\mathbf{x}_i^\top, \nabla_{\mathbf{W}_1}f_{\mathbf{W}}(\mathbf{x}_i) = \sqrt{m}\cdot\mathbf{D}_i\mathbf{W}_2^\top\mathbf{x}_i^\top.$$

Therefore we have

$$\begin{aligned}\|\nabla_{\mathbf{W}_1} f_{\mathbf{W}}(\mathbf{x}_i) - \nabla_{\mathbf{W}_1} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\|_F &= \sqrt{m}\cdot\|\mathbf{D}_i\mathbf{W}_2^\top\mathbf{x}_i^\top - \mathbf{D}_i^{(0)}\mathbf{W}_2^{(0)\top}\mathbf{x}_i^\top\|_F \\ &= \sqrt{m}\cdot\|\mathbf{x}_i\mathbf{W}_2\mathbf{D}_i - \mathbf{x}_i\mathbf{W}_2^{(0)}\mathbf{D}_i^{(0)}\|_F \\ &= \sqrt{m}\cdot\|\mathbf{W}_2\mathbf{D}_i - \mathbf{W}_2^{(0)}\mathbf{D}_i^{(0)}\|_F \\ &\leqslant \sqrt{m}\cdot\|\mathbf{W}_2^{(0)}(\mathbf{D}_i^{(0)} - \mathbf{D}_i)\|_F + \sqrt{m}\cdot\|(\mathbf{W}_2^{(0)} - \mathbf{W}_2)\mathbf{D}_i\|_F\end{aligned}$$

By Lemma 7.4 in Allen-Zhu et al. (2018) and Lemma D.1, with probability at least $1 - n\cdot\exp[-\Omega(m)]$, $\sqrt{m}\cdot\|\mathbf{W}_2^{(0)}(\mathbf{D}_i^{(0)} - \mathbf{D}_i)\|_F \leqslant \mathcal{O}(\omega^{1/3}\sqrt{m})$ for all $i\in[n]$. Moreover, clearly $\|(\mathbf{W}_2^{(0)} - \mathbf{W}_2)\mathbf{D}_i\|_F \leqslant \|\mathbf{W}_2^{(0)} - \mathbf{W}_2\|_F \leqslant \omega$. Therefore

$$\|\nabla_{\mathbf{W}_1} f_{\mathbf{W}}(\mathbf{x}_i) - \nabla_{\mathbf{W}_1} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\|_F \leqslant \mathcal{O}(\omega^{1/3}\sqrt{m})$$

for all $i\in[n]$. This proves the bound for the first layer gradients. For the second layer gradients, we have

$$\nabla_{\mathbf{W}_2} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i) = \sqrt{m}\cdot[\sigma(\mathbf{W}_1^{(0)}\mathbf{x}_i)]^\top, \nabla_{\mathbf{W}_2} f_{\mathbf{W}}(\mathbf{x}_i) = \sqrt{m}\cdot[\sigma(\mathbf{W}_1\mathbf{x}_i)]^\top$$

It therefore follows by the 1-Lipschitz continuity of $\sigma(\cdot)$ that

$$\|\nabla_{\mathbf{W}_2} f_{\mathbf{W}}(\mathbf{x}_i) - \nabla_{\mathbf{W}_2} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\|_F \leqslant \sqrt{m}\cdot\|\mathbf{W}_1\mathbf{x}_i - \mathbf{W}_1^{(0)}\mathbf{x}_i\|_F \leqslant \omega\sqrt{m} \leqslant \omega^{1/3}\sqrt{m}.$$

This completes the proof of the first inequality.

The second inequality directly follows by triangle inequality and Lemma D.4:

$$\begin{aligned}|\langle\nabla_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{x}_i), \nabla_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{x}_j)\rangle - m\mathbf{K}^{(0)}| &\leqslant |\langle\nabla_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{x}_i) - \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \nabla_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{x}_j)\rangle| \\ &\quad + |\langle\nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \nabla_{\mathbf{W}} f_{\mathbf{W}}(\mathbf{x}_j) - \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_j)\rangle| \\ &\leqslant \mathcal{O}(\omega^{1/3}m).\end{aligned}$$

This finishes the proof. □