Title
Understanding the Role of Optimization Algorithms in Learning Over-parameterized Models

Author
Zou, Difan

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Understanding the Role of Optimization Algorithms in Learning Over-parameterized Models

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Difan Zou

2022

ABSTRACT OF THE DISSERTATION

Understanding the Role of Optimization Algorithms in
Learning Over-parameterized Models

by

Difan Zou
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2022
Professor Quanquan Gu, Chair

Deep learning has witnessed fast growth and wide application in recent years. One of the essential properties of the modern deep learning model is that it is sufficiently over-parameterized, i.e., it has much more learnable parameters than the number of training examples. The over-parameterization, on one hand, is the core of the superior approximation/representation ability of the neural network function, while, on the other hand, could lead to severe over-fitting issues, according to the conventional wisdom in learning theory. However, this is not consistent with the empirical success in deep learning, where the neural network model, trained by standard optimization algorithms (e.g., stochastic gradient descent, Adam, etc.), can not only perfectly fit the training data (i.e., finding the global solution of the training objective), but also generalizes well on the test data. This dissertation seeks to understand and explain this phenomenon by carefully characterizing the role of optimization algorithms in learning over-parameterized models.

We begin the dissertation by studying arguably the simplest model: over-parameterized linear regression problems. In particular, we consider a class of SGD algorithms and prove

problem-dependent generalization error bounds accordingly. Based on the established generalization guarantees, we will further characterize the sufficient conditions on the least square problem itself (e.g., conditions on data distribution and ground-truth model parameters) such that the SGD algorithm can generalize. Moreover, motivated by the recent work on the implicit regularization of SGD, we also provide a complete comparison between the SGD solution and the solution of regularized least square (i.e., ridge regression). We demonstrate the benefit of SGD compared to ridge regression for a large class of the least square problems classes, which partially explains the implicit regularization effect of SGD.

In the second part, we study the effect of optimization algorithms for learning over-parameterized neural network models. Different from linear models that their optimization guarantees can be easily established, studying the optimization in training deep neural networks is challenging since the training objective is nonconvex or even nonsmooth. Therefore, we first study the optimization in training over-parameterized neural network models and establish global convergence guarantees for both GD and SGD under mild conditions on the data distribution. Based on the optimization analysis, we further establish an algorithm-dependent generalization analysis for SGD/GD. We show that if the data distribution admits certain good separation properties, a deep ReLU network with polylogarithmic width can be successfully trained with a global convergence guarantee and good generalization ability. Finally, we compare the generalization ability of two different optimization algorithms in learning over-parameterized neural networks: GD and Adam, and show that Adam and GD exhibit different algorithmic biases, which consequently leads to different solutions that have different generalization performances.

The works covered in this dissertation form exploration in understanding the role of optimization algorithms in learning over-parameterized models, which is an incomplete collection of the recent advances in deep learning theory but shed light on a broader class of future directions in deep learning research.

The dissertation of Difan Zou is approved.

Sham Kakade

Stefano Soatto

Stanley Osher

Yizhou Sun

Quanquan Gu, Committee Chair

University of California, Los Angeles

2022

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I would like to give great thanks to my PhD advisor, Prof. Quanquan Gu, who gave me such an opportunity to join his fantastic research group. When I first joined his group, I nearly have no background knowledge and experience in machine learning research. He has provided enormous help and extensive constructive advice during my early years in my PhD study, without which I can never be able to become a mature Ph.D. student and machine learning researcher. I also want to express my gratitude to Prof. Gu for helping me collaborate with other amazing researchers, which greatly broadens my research focus and benefits my future career development.

I want to thank my doctoral committee members, Prof. Sham Kakade, Prof. Stefano Soatto, Prof. Stanley Osher, and Prof. Yizhou Sun, who have not only provided valuable feedbacks on my dissertation work, but also given innovative ideas and constructive suggestions throughout our collaborations in many research projects. I also want to thank Prof. Yuanzhi Li for his insightful advice in our collaborations. Besides, I am greatly thankful to Bloomberg for awarding me the fellowship during two years of my Ph.D. study. I also want to extend my gratitude to Ni Ma and Saher Esmeir for their guidance, who were my mentors during my internship in Bloomberg.

I am extremely fortunate to work with so many amazing people in Statistical Machine Learning Lab: Yuan Cao, Jinghui Chen, Pan Xu, Lingxiao Wang, Lu Tian, Spencer Frei, Zixiang Chen, Jiafan He, Yue Wu, Weitong Zhang, and Dongruo Zhou, Yihe Deng, and Xuheng Li. Particular thanks to Jinghui, Pan, Lu, and Lingxiao for helping me get adapt to work in the lab during early years of my Ph.D. study. Also thank Pan, Jinghui, Lingxiao, Yuan, Dongruo, Zixiang, Weitong, and Dongruo for their help and efforts in many of our collaborated projects. Besides, I would like to express my special gratitude to Jingfeng Wu, who is an extremely fantastic and responsible collaborator. Two of my dissertation chapters are based on our collaborated projects.

Lastly, I want to express tremendous thanks to my parents. Your encouragement and love are the biggest support for me to pursue this wonderful journey. Also thank to Xiaona and lovely "Benben" Shuming for filling my life with love and happiness.

VITA

2010-2014	B.S. (Applied Physics), School of Gifted Young, University of Science and Technology of China
2014-2017	M.S. (Information and Communication Engineering), Department of Electrical Engineering and Information Science, University of Science and Technology of China
2017-2018	Teaching Assistant, Department of System and Information Engineering, University of Virginia
2018-2022	Research Assistant, Computer Science Department, University of California, Los Angeles
2019-2020	Teaching Assistant, Computer Science Department, University of California, Los Angeles

PUBLICATIONS

*We select publications that are the most relevant to the topic of this dissertation. * indicates equal contribution.*

Difan Zou and Quanquan Gu. An Improved Analysis of Training Over-parameterized Deep Neural Networks. NeurIPS, 2019.

Difan Zou*, Yuan Cao*, Dongruo Zhou, and Quanquan Gu. Gradient Descent Optimizes Over-parameterized Deep ReLU Networks. MLJ, 2019.

Difan Zou, Philip M. Long, and Quanquan Gu, On the Global Convergence of Training Deep Linear ResNets. ICLR, 2020.

Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu, Direction Matters: On the Implicit Regularization Effect of Stochastic Gradient Descent with Moderate Learning Rate. ICLR, 2021.

Zixiang Chen*, Yuan Cao*, Difan Zou*, and Quanquan Gu, How Much Over-parameterization Is Sufficient to Learn Deep ReLU Networks? ICLR, 2021.

Difan Zou*, Jingfeng Wu*, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade, Benign Overfitting of Constant-Stepsize SGD for Linear Regression. COLT, 2021.

Difan Zou*, Jingfeng Wu*, Vladimir Braverman, Quanquan Gu, Dean P. Foster, and Sham M. Kakade, The Benefits of Implicit Regularization from SGD in Least Squares Problems. NeurIPS, 2021.

Difan Zou, Yuan Cao, Yuanzhi Li and Quanquan Gu, Understanding the Generalization of Adam in Learning Neural Networks with Proper Regularization. NeurIPS 2021, OPT Workshop.

CHAPTER 1

Introduction

Deep neural networks have achieved great success in many applications like image processing [KSH12], speech recognition [HDY12] and Go games [SHM16]. However, the success of deep learning has not been well-explained in theory. It remains mysterious why the neural network models can be effectively learned by standard optimization algorithms, such as gradient descent (GD) and stochastic gradient descent (SGD), despite the extremely large amount of learnable parameters (i.e., the model is over-parameterized) and highly non-convex landscape of the training loss function. In particular, over-parameterization implies that the training loss function may have multiple or even infinite global minima, while only a large portion of them may generalize poorly on the test data [ZBH16; LPA20]. Nonconvexity will bring huge difficulties in finding the global solutions to the training loss function since there may exist many spurious local minima or saddle points. It is conjectured that the optimization algorithms have some "magic power" to implicitly regularizes the over-parameterized models [NTS14; ZBH16; KMN16] and find certain good solutions from the numerous candidate solutions, including spurious local minima, saddle points, and the global minima that over-fit the training data. This is typically referred to as *implicit regularization* [NTS14]. However, there still lacks a rigorous theoretical justification of this conjecture, i.e., it remains unclear why and how such implicit regularization can benefit the generalization. Motivated by this, the primary goal of this dissertation is to develop a sharp theoretical understanding of the role of optimization algorithms in learning over-parameterized models.

As the first step to understanding the behavior of optimization algorithms in learning

over-parameterized models, there is reason to believe that characterizing these effects even in conceptually simpler (e.g. linear model) settings will also help our understanding of more complex settings, because many high dimensional effects are also observed even in simple linear models. Recently, there is a growing body of work studying the generalization for certain statistical estimators such as ordinary least square (OLS) or ridge regression estimators [NKB21; BLL20; BHX20; HMR22; TB20; MNS21; CL21; NVK20]. In contrast, the algorithmic aspects of generalization are far less well understood, where we lack a sharp characterization of the generalization error achieved by optimization algorithms, such as SGD. With regards to SGD, existing works on its generalization analysis mainly lies in the classical under-parameterized regime [DB15a; BM13; DFB17; JNK17; JKK18a] (few exceptions [DB15b; BBG20] are discussed in Section 2.3), where the dimension is less than the number of training examples. To this end, our first work (see **Chapter** 2) aims to provide a sharp generalization error bound showing how (unregularized) SGD can generalize in the over-parameterized, or even infinite-dimensional setting.

Additionally, the implicit regularization of SGD/GD has also been extensively studied for linear models. For example, (multi-pass) SGD for linear regression converges to the *minimum-norm interpolator*, which corresponds to the limit of the ridge solution with a vanishing penalty [ZBH16; GLS18a]. Tangential evidence for this also comes from examining gradient descent, where a continuous time (gradient flow) analysis shows how the optimization path of gradient descent is (pointwise) closely connected to an explicit, ℓ_2-regularization [SPR18; AKT19]. Similar results [ADT20] have been further extended to SGD, where an (early-stopped) continuous-time SGD is demonstrated to perform similarly to ridge regression with certain regularization parameters. However, as of yet, a precise comparison between the implicit regularization afforded by SGD and the explicit regularization of ridge regression (in terms of the *generalization performance*) is still lacking, especially when the hyperparameters (e.g., stepsize for SGD and regularization parameter for ridge regression) are allowed to be tuned. Motivated by this, our second work (see **Chapter** 3)

aims to deliver an *instance-based* risk comparison between SGD and ridge regression, based on the sharp generalization bounds established in our first work.

Provided with the understanding gained for linear models, the next step is to consider more complicated and practical models: deep neural network (DNN) models. Unlike the linear setting where the training loss function is convex so that the optimization is never an issue, both the optimization and generalization are challenging topics in learning DNN models since the training objective is highly nonconvex or even nonsmooth (if using ReLU activation functions). It is well known that without any additional assumption, even training a shallow neural network is an NP-hard problem [BR89]. Establishing better convergence guarantees typically requires certain assumptions on the data distribution [BGM18; DLT18; ZSD17] and network structures [CHM15; Kaw16]. More recently, a series of works [ALS19b; DZP18; LL18] have observed that the key to the convergence of GD/SGD lies in two aspects: over-parameterization and random weight initialization. [DZP18; LL18] showed that for a one-hidden-layer network with ReLU activation function using over- parameterization and random initialization, if the training data are non-degenerate, GD and SGD can find the near global-optimal solutions in polynomial time with respect to the accuracy parameter and training sample size. Based on these prior findings, the aim of our third & work (see **Chapter**s 4) is to advance this line of research by establishing a sharp convergence guarantee of gradient based methods for deep ReLU networks.

Moreover, it has been further demonstrated that with the standard random initialization, the training of over-parameterized deep neural networks can be characterized by a kernel function called neural tangent kernel (NTK) [JGH18; ADH19b]. In the neural tangent kernel regime (or lazy training regime [CB18]), the neural network function behaves similarly to its first-order Taylor expansion at initialization [JGH18; LXS19; ADH19b; CG19], which enables feasible optimization and, more importantly, generalization analysis. Accordingly, [ALL19; ADH19a; CG19] established generalization bounds of neural networks trained with (stochastic) gradient descent, and showed that the neural networks can learn target functions

in certain reproducing kernel Hilbert space (RKHS) or the corresponding random feature function class. However, their theoretical analysis requires the neural network function to be extremely close to the corresponding NTK function, which consequently leads to a nearly unrealistic requirement on the neural network width (i.e., a high degree polynomial of the training sample size and the inverse target error). To address this problem and attempt to fill the gap between practice and theory, we (see **Chapter** 5) develop a new theoretical analysis that only requires a constant approximation error to the NTK function and proves sharper learning guarantees for deep ReLU networks trained by GD/SGD. Consequently, we show that if the data can be well separated in the RHKS, a deep ReLU network with polylogarithmic width can be successfully trained by GD and SGD with a global convergence guarantee and good generalization ability.

Finally, we explore the limitation of NTK based analysis in learning over-parameterized neural networks. In particular, our last work (see **Chapter** 6) is motivated by a practical observation that different optimization algorithms (e.g., Adam, SGD) perform differently in many deep learning applications such as image classification, even with a fine-tuned regularization. This cannot be well explained by linear models or neural networks trained in the "almost convex" NTK regime, since they suggest that SGD and Adam will achieve similar generalization performance, in the presence of weight decay regularization. We focus on this particular problem and provide a new theoretical explanation for this phenomenon by explicitly showing that in the nonconvex setting of learning overparameterized two-layer CNNs, SGD and Adam will converge to different global solutions with provably different generalization errors.

1.1 Organization of the paper

In the first part of this dissertation, we study the (over-parameterized) linear regression problem and develop a novel theoretical analysis for sharply characterizing the generaliza-

tion error or excess risk achieved by stochastic gradient descent. In particular, **Chapter** 2 provides an instance-dependent excess risk bounds for SGD, which is stated as a function with respect to the full eigenspectrum of the data covariance, sample size, and ground truth model parameters. We also provide a matching lower bound (up to constant factors) to justify the tightness of the developed theoretical characterizations. Moreover, in order to understand the implicit regularization of SGD, **Chapter** 3 compares the generalization ability of SGD to that of the solution found by adding explicit regularization, i.e., ridge regression in an instance-wise manner. We show that for a large class of statistically interesting problems, SGD can provably generalize no worse than ridge regression with optimally tuned parameters, when provided with logarithmically more samples. Conversely, for some problem instances, optimally tuned ridge regression may require quadratically more samples than SGD to achieve the same generalization performance.

In the second part of this dissertation, we study the optimization and generalization of over-parameterized neural network models. In particular, **Chapter** 4, as one of the first works, proved the global convergence of gradient descent for training over-parameterized neural networks. Compared to the concurrent works [ALS19a; DLL19], we proved a faster convergence rate with milder assumptions on the neural network width and data distribution. Moreover, in **Chapter** 5, we conduct the generalization analysis of training over-parameterized deep ReLU networks in the NTK regime, and established the state-of-the-art optimization and generalization guarantees under certain data separation conditions. Lastly, in **Chapter** 6, we explore the theoretical analysis beyond the NTK regime by studying the generalization gap between GD and Adam in learning over-parameterized CNN models for image classification tasks. We show that GD and Adam exhibit different algorithmic biases, which will consequently converge to different solutions (with different generalization performances), for a class of image-like data distribution.

We summarize this dissertation in **Chapter** 7.

1.2 Notations

We use lower case, lower case bold face, and upper case bold face letters to denote scalars, vectors and matrices respectively. For a positive integer n, we denote $[n] = \{1, \ldots, n\}$. For a vector $\mathbf{x} = (x_1, \ldots, x_d)^\top$, we denote by $\|\mathbf{x}\|_p = \big(\sum_{i=1}^d |x_i|^p\big)^{1/p}$ the ℓ_p norm of $\mathbf{x}$, $\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,d} |x_i|$ the ℓ_∞ norm of $\mathbf{x}$, and $\|\mathbf{x}\|_0 = |\{x_i : x_i \neq 0, i = 1, \ldots, d\}|$ Cthe number of non-zero entries of $\mathbf{x}$. We use $\mathrm{Diag}(\mathbf{x})$ to denote a square diagonal matrix with the elements of vector $\mathbf{x}$ on the main diagonal. For a matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{m\times n}$, we use $\|\mathbf{A}\|_F$ to denote the Frobenius norm of $\mathbf{A}$, $\|\mathbf{A}\|_2$ to denote the spectral norm (maximum singular value), and $\|\mathbf{A}\|_0$ to denote the number of nonzero entries. We denote by $S^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$ the unit sphere in $\mathbb{R}^d$. For a positive semi-definite (PSD) matrix $\mathbf{A}$ and a vector $\mathbf{v}$ of appropriate dimension, we write $\|\mathbf{v}\|_{\mathbf{A}}^2 := \mathbf{v}^\top \mathbf{A}\mathbf{v}$. For two matrices, we use $\langle \mathbf{A}, \mathbf{B}\rangle = \sum A_{ij}B_{ij}$ to denote the inner product between two matrices and use $\mathbf{A} \otimes \mathbf{B}$ to denote their Kronecker/tensor product.

For two sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n = O(b_n)$ to denote that $a_n \leq C_1 b_n$ for some absolute constant $C_1 > 0$, and use $a_n = \Omega(b_n)$ to denote that $a_n \geq C_2 b_n$ for some absolute constant $C_2 > 0$. We use $a_n = \Theta(b_n)$ if $a_n = \Omega(b_n)$ and $a_n = O(b_n)$. In addition, we also use $\widetilde{O}(\cdot)$, $\widetilde{\Omega}(\cdot)$, $\widetilde{\Theta}(\cdot)$ to hide logarithmic terms in Big-O, Big-Omega, and Big-Theta notations.

We also use the following matrix product notation. For indices l_1, l_2 and a collection of matrices $\{\mathbf{A}_r\}_{r\in\mathbb{Z}_+}$, we denote

$$\prod_{r=l_1}^{l_2} \mathbf{A}_r := \begin{cases} \mathbf{A}_{l_2}\mathbf{A}_{l_2-1}\cdots\mathbf{A}_{l_1} & \text{if } l_1 \leq l_2 \\ \mathbf{I} & \text{otherwise.} \end{cases} \tag{1.2.1}$$

Part I

Learning Over-parameterized Linear Models

CHAPTER 2

Generalization of SGD for Linear Regression

2.1 Introduction

In this work, we study the standard classical linear regression problem:

$$\min_{\mathbf{w}} L(\mathbf{w}), \text{ where } L(\mathbf{w}) = \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(y - \langle \mathbf{w}, \mathbf{x}\rangle)^2\right], \tag{2.1.1}$$

where $\mathbf{x} \in \mathcal{H}$, is the feature vector, where, $\mathcal{H}$ is some (finite d-dimensional or countably infinite dimensional) Hilbert space; $y \in \mathbb{R}$ is the response; $\mathcal{D}$ is an unknown distribution over $\mathbf{x}$ and y; and $\mathbf{w} \in \mathcal{H}$ is the weight vector to be optimized. We consider the stochastic approximation approach using constant stepsize SGD, with iterate averaging: at each iteration t, an i.i.d. example $(\mathbf{x}_t, y_t) \sim \mathcal{D}$ is observed, and the weight is updated according to SGD as follows:

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma\left(y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t\rangle\right)\mathbf{x}_t, \qquad t = 1, \ldots, N, \tag{2.1.2}$$

where $\gamma > 0$ is a constant stepsize, N is the number of samples observed, and the weights are initialized at $\mathbf{w}_0 \in \mathcal{H}$. The final output will be the average of the iterates:

$$\overline{\mathbf{w}}_N := \frac{1}{N}\sum_{t=0}^{N-1} \mathbf{w}_t.$$

In the underparameterized setting with finite dimension d ($d \ll N$), a rich body of work has established that $\overline{\mathbf{w}}_N$ enjoys the optimal risk (up to constant factors) of $\mathcal{O}\left(d\sigma^2/N\right)$, for sufficiently large N. The focus of this work is on the over-parameterized regime, where $d \gg N$ (or possibly countably infinite).

Contributions. Our main result can be viewed as a counterpart to the classical analysis of iterate averaged SGD to the overparameterized regime for linear regression: we provide a sharp excess risk bound showing how (unregularized) SGD can generalize even in the infinite-dimensional setting. Our bound is stated in a general manner, in terms of the full eigenspectrum of the data covariance matrix along with a functional dependency on the initial iterate; our lower bound shows our characterization is tight. As a corollary, we see how the benign-overfitting phenomenon can be observed for SGD, provided certain spectrum decay conditions on the data covariance are met. We also extend our results to SGD with tail-averaging [JKK18a; JNK17], where we run SGD for s iterations and then take average over the subsequent N iterates as the output. (see Section 2.2.2 for more details.)

Some additional notable contributions are:

- The sharpness of our bounds permits us to make comparisons to OLS (the minimum-norm interpolator) and ridge regression. Notably, in a contrast to the variance of OLS [BLL20], the variance contribution to SGD is well behaved under substantially weaker assumptions on the spectrum of the data covariance. This shows how inductive bias of SGD, in comparison to the minimum-norm interpolator, can lead to better generalization with no regularization. We also constrast our results to ridge regression based on the recent work by [TB20].

- One notable aspect of our work is a sharp characterization of a "bias process" in SGD. In particular, consider the special case where $y = \mathbf{w}^\star \cdot \mathbf{x}$ (with probability one), for some $\mathbf{w}^\star$. Here, SGD still differs from gradient descent on $L(\mathbf{w})$. Our characterization gives a novel characterization of how the variance in this process contributes to the final excess risk bound.

- From a technical standpoint, our work develops new proof techniques for iterate averaged SGD. Our analysis tools are based on the operator view of averaged SGD [DB15b; JNK17; JKK18a]. A core idea in the proof is in connecting the finite sample (infinite

dimensional) covariance matrices of the variance and bias stochastic processes to those of their corresponding (asymptotic) stationary covariance matrices — an idea that was introduced in [JKK18a] for the finite dimensional, variance analysis.

2.2 Main Results

We now provide matching (upto absolute constants) upper and lower excess risk bounds for iterate averaged SGD. We then compare these rates to those of OLS and ridge regression, where we see striking similarities and notable differences.

2.2.1 Benign Overfitting of SGD

We first introduce relevant notation and our assumptions. Our first assumption is mild regularity conditions on the moments of the data distribution.

Assumption 2.2.1 (Regularity conditions)**.** Assume $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$, $\mathbb{E}[\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x}\otimes\mathbf{x}]$, and $\mathbb{E}[y^2]$ exist and are all finite. Furthermore, denote the second moment of $\mathbf{x}$ by

$$\mathbf{H} := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top],$$

and suppose that $\mathrm{tr}(\mathbf{H})$ is finite. For convenience, we assume that $\mathbf{H}$ is strictly positive definite and that $L(\mathbf{w})$ admits a unique global optimum, which we denote by $\mathbf{w}^* := \mathrm{argmin}_{\mathbf{w}}\, L(\mathbf{w})$. [1]

Our second assumption is on the behavior of the fourth moment, when viewed as a linear operator on PSD matrices:

Assumption 2.2.2 (Fourth moment condition)**.** Assume there exists a positive constant

[1] This is not necessary. In the case where $\mathbf{H}$ has eigenvalues which are 0, we could instead choose $\mathbf{w}^*$ to be the minimum norm vector in the set $\mathrm{argmin}_{\mathbf{w}}\, L(\mathbf{w})$, and our results would hold for this choice of $\mathbf{w}^\star$. For example, see [SSB02] for a rigorous treatment of working in a reproducing kernel Hilbert space.

$\alpha > 0$, such that for any PSD matrix $\mathbf{A}$[2], it holds that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha\,\mathrm{tr}(\mathbf{H}\mathbf{A})\mathbf{H}.$$

For Gaussian distributions, it suffices to take $\alpha = 3$. Furthermore, it is worth noting that this assumption is implied if the distribution over $\mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ has sub-Gaussian tails (see Lemma 2.5.1 for a precise claim). Also, it is not difficult to verify that $\alpha \geq 1$.[3]

Assuming sub-Gaussian tails over $\mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ is standard assumption in regression analysis (e.g. [HKZ14; BLL20; TB20]), and, as mentioned above, this assumption is substantially weaker. The assumption is somewhat stronger than what is often assumed for iterate averaged SGD in the underparameterized regime (e.g., [BM13; JNK17]) (see Section 2.3 for further discussion). Additionally, we also remark that Assumption 2.2.2 can be further relaxed to that we only require $\mathbf{A}$ is PSD and commutable with $\mathbf{H}$, rather than all PSD matrix $\mathbf{A}$ (see Section 2.9 for more details).

Our next assumption is a noise condition, where it is helpful to interpret $y - \langle\mathbf{w}^*, \mathbf{x}\rangle$ as the additive noise. Observe that the first order optimality conditions on $\mathbf{w}^*$ imply $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \langle\mathbf{w}^*, \mathbf{x}\rangle)\mathbf{x}] = \nabla L(\mathbf{w}^*) = \mathbf{0}$.

Assumption 2.2.3 (Noise condition)**.** Suppose that:

$$\mathbf{\Sigma} := \mathbb{E}\left[(y - \langle\mathbf{w}^*, \mathbf{x}\rangle)^2\mathbf{x}\mathbf{x}^\top\right], \quad \sigma^2 := ||\mathbf{H}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{H}^{-\frac{1}{2}}||_2$$

exist and are finite. Note that $\mathbf{\Sigma}$ is the covariance matrix of the gradient noise at $\mathbf{w}^\star$.

This assumption places a rather weak requirement on the additive noise (due to that it permits model mis-specification) and is often made in the average SGD literature (e.g.,

[2]This assumption can be relaxed into: for any PSD matrix $\mathbf{A}$ *that commutes with* $\mathbf{H}$, it holds that $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha\,\mathrm{tr}(\mathbf{H}\mathbf{A})\mathbf{H}$. The presented analyzing technique is ready to be modified to cooperate with the relaxed assumption with the observation that the fourth moment operator is linear and self-adjoint. Similar relaxation applies to Assumption 2.2.5 as well.

[3]This is due to that the square of the second moment is less than the fourth moment.

[BM13; DFB17]). Observe that for *well-specified models*, where

$$y = \langle \mathbf{w}^\star, \mathbf{x} \rangle + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2_{\text{noise}}), \tag{2.2.1}$$

we have that $\mathbf{\Sigma} = \sigma^2_{\text{noise}} \mathbf{H}$ and so $\sigma^2 = \sigma^2_{\text{noise}}$.

Before we present our main theorem, a few further definitions are in order: denote the eigendecomposition of the Hessian as $\mathbf{H} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, where $\{\lambda_i\}_{i=1}^\infty$ are the eigenvalues of $\mathbf{H}$ sorted in non-increasing order and $\mathbf{v}_i$'s are the corresponding eigenvectors. We then denote:

$$\mathbf{H}_{0:k} := \textstyle\sum_{i=1}^k \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad \text{and} \quad \mathbf{H}_{k:\infty} := \sum_{i>k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top.$$

Similarly we denote $\mathbf{I}_{0:k} := \sum_{i=1}^k \mathbf{v}_i \mathbf{v}_i^\top$ and $\mathbf{I}_{k:\infty} := \sum_{i>k} \mathbf{v}_i \mathbf{v}_i^\top$. By the above definitions, we know

$$\|\mathbf{w}\|^2_{\mathbf{H}^{-1}_{0:k}} = \sum_{i \le k} \frac{(\mathbf{v}_i^\top \mathbf{w})^2}{\lambda_i}, \quad \|\mathbf{w}\|^2_{\mathbf{H}_{k:\infty}} = \sum_{i>k} \lambda_i (\mathbf{v}_i^\top \mathbf{w})^2,$$

where we have slightly abused notation in that $\mathbf{H}^{-1}_{0:k}$ denotes a pseudo-inverse.

We now present our main theorem:

Theorem 2.2.4 (Benign overfitting of SGD)**.** Suppose Assumptions 2.2.1-2.2.3 hold and that the stepsize is set so that $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. Then the excess risk can be upper bounded as follows,

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \le 2 \cdot \text{EffectiveBias} + 2 \cdot \text{EffectiveVar},$$

where

$$\begin{aligned}
\text{EffectiveBias} &= \frac{1}{\gamma^2 N^2} \cdot ||\mathbf{w}_0 - \mathbf{w}^*||^2_{\mathbf{H}^{-1}_{0:k^*}} + ||\mathbf{w}_0 - \mathbf{w}^*||^2_{\mathbf{H}_{k^*:\infty}}, \\
\text{EffectiveVar} &= \frac{2\alpha\big(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{N\gamma(1 - \gamma\alpha \operatorname{tr}(\mathbf{H}))} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2 \right) \\
&\quad + \frac{\sigma^2}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2 \right)
\end{aligned}$$

with $k^* = \max\{k : \lambda_k \ge \frac{1}{\gamma N}\}$.

The interpretation is as follows: the "effective bias" precisely corresponds to the rate of convergence had we run gradient descent directly on $L(\mathbf{w})$ (i.e., where the latter has no variance due to sampling). The "effective variance" error stems from both the additive noise $y - \langle \mathbf{w}^*, \mathbf{x} \rangle$, i.e., the second term of the EffectiveVariance error, along with that even if there was no additive noise (i.e. $y - \langle \mathbf{w}^*, \mathbf{x} \rangle = 0$ with probability one), i.e., the first term of the EffectiveVariance error, then SGD would still not be equivalent to GD. The cut-off index k^*, which we refer to as the "effective dimension", plays a pivotal role in the excess risk bound, which separates the entire space into a k^*-dimensional "head" subspace where the bias error decays more quickly than that of the bias error in the complement "tail" subspace. To obtain a vanishing bound, the effective dimension k^* must be $\mathrm{o}(N)$ and the tail summation $\sum_{i>k^*} \lambda_i^2$ must be $\mathrm{o}(1/N)$.

In terms of constant factors, the above bound can be improved by a factor of 2 in the effective bias-variance decomposition (see (2.4.6)). We now turn to lower bounds.

A lower bound. We first introduce the following assumption that states a lower bound on the fourth moment.

Assumption 2.2.5 (Fourth moment condition, lower bound)**.** Assume there exists a constant $\beta \geq 0$, such that for any PSD matrix $\mathbf{A}$, it holds that

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\mathbf{x}\mathbf{x}^\top \mathbf{A} \mathbf{x}\mathbf{x}^\top] - \mathbf{H}\mathbf{A}\mathbf{H} \succeq \beta \operatorname{tr}(\mathbf{H}\mathbf{A})\mathbf{H}.$$

For Gaussian distributions, it suffices to take $\beta = 1$.

The following lower bound shows that when the noise is well-specified our upper bound is not improvable except for absolute constants.

Theorem 2.2.6 (Excess risk lower bound)**.** Suppose $N \geq 500$. For any well-specified data distribution $\mathcal{D}$ (see (2.2.1)) that also satisfies Assumptions 2.2.1 and 2.2.5, for any stepsize

such that $\gamma < 1/\lambda_1$, we have that:

$$\begin{aligned}\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \geq & \frac{1}{100\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \frac{1}{100} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\ & + \frac{\beta\left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right)}{16000 N\gamma} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right) \\ & + \frac{\sigma^2_{\text{noise}}}{50} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right)\end{aligned}$$

with $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$.

Similar to the upper bound stated in Theorem 2.2.4, the first two terms represent the EffectiveBias and the last two terms represent the EffectiveVariance, in which the third and last terms are contributed by the model noise and variance in SGD. Our upper bound matches our lower bound up to absolute constants, which indicates the obtained rates are tight, at least for Gaussian data distribution with well-specified noise.

Special cases. It is instructive to consider a few special cases of Theorem 2.2.4. We first show the result for SGD with large stepsizes.

Corollary 2.2.7 (Benign overfitting with large stepsizes)**.** Suppose Assumptions 2.2.1-2.2.3 hold and that the stepsize is set to $\gamma = 1/(2\alpha\sum_i \lambda_i)$. Then

$$\begin{aligned}\text{EffectiveBias} &= \frac{4\alpha^2(\sum_i \lambda_i)^2}{N^2} \cdot ||\mathbf{w}_0 - \mathbf{w}^*||^2_{\mathbf{H}^{-1}_{0:k^*}} + ||\mathbf{w}_0 - \mathbf{w}^*||^2_{\mathbf{H}_{k^*:\infty}} \\ \text{EffectiveVar} &= \left(2\sigma^2 + 4\alpha^2\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}}\right) \cdot \left(\frac{k^*}{N} + \frac{N\sum_{i>k^*}\lambda_i^2}{4\alpha^2(\sum_i \lambda_i)^2}\right),\end{aligned}$$

where $k^* = \max\{k : \lambda_k \geq \frac{2\alpha\sum_i \lambda_i}{N}\}$.

Note that the bias error decays at different rates in different subspaces. Crudely, in the "head" eigenspace (spanned by the eigenvectors corresponding to large eigenvalues) the bias error decays in a faster $\mathcal{O}(1/N^2)$ rate (though there is weighting of λ_i in the head), while in the remaining "tail" eigenspace, the bias error decays at a slower $\mathcal{O}(1/N)$ rate (due to

that all the eigenvalues in the tail are less than $\mathcal{O}(1/N)$). The following corollary provides a crude bias bound, showing that bias never decays more slowly than $\mathcal{O}(1/N)$.

Corollary 2.2.8 (Crude bias-bound)**.** Suppose Assumptions 2.2.1-2.2.3 hold and that the stepsize is set to $\gamma = 1/(2\alpha \sum_i \lambda_i)$. Then

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \leq \frac{8\alpha \|\mathbf{w}_0 - \mathbf{w}^*\|_2^2 \cdot \sum_i \lambda_i}{N} + 4\sigma^2 \cdot \left(\frac{k^*}{N} + \frac{N \sum_{i>k^*} \lambda_i^2}{4\alpha^2 (\sum_i \lambda_i)^2} \right),$$

where $k^* = \max\{k : \lambda_k \geq \frac{2\alpha \sum_i \lambda_i}{N}\}$.

Theorems 2.2.4 and 2.2.6 suggests that the excess risk achieved by SGD depends on the spectrum of the covariance matrix. The following corollary gives examples of data spectrum such that the excess risk is diminishing.

Corollary 2.2.9 (Example data distributions)**.** Under the same conditions as Theorem 2.2.4, suppose $||\mathbf{w}_0 - \mathbf{w}^*||_2$ is bounded.

1. For $\mathbf{H} \in \mathbb{R}^{d \times d}$, let $s = N^r$ and $d = N^q$ for some positive constants $0 < r \leq 1$ and $q \geq 1$. If the spectrum of $\mathbf{H}$ satisfies

$$\lambda_k = \begin{cases} 1/s, & k \leq s, \\ 1/(d-s), & s+1 \leq k \leq d, \end{cases}$$

   then $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}(N^{r-1} + N^{1-q})$.

2. If the spectrum of $\mathbf{H}$ satisfies $\lambda_k = k^{-(1+r)}$ for some $r > 0$, then $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(N^{-r/(1+r)}\right)$.

3. If the spectrum of $\mathbf{H}$ satisfies $\lambda_k = k^{-1} \log^{-\beta}(k+1)$ for some $\beta > 1$, then $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(\log^{-\beta}(N)\right)$.

4. If the spectrum of $\mathbf{H}$ satisfies $\lambda_k = e^{-k}$, then $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}(\log(N)/N)$.

2.2.2 The Effect of Tail-Averaging

We further consider benign overfitting of SGD when *tail-averaging* [JNK17] is applied, i.e.,

$$\overline{\mathbf{w}}_{s:s+N} = \frac{1}{N}\sum_{t=s}^{s+N-1} \mathbf{w}_t.$$

We present the following theorem as a counterpart of Theorem 2.2.4. The proof is deferred to Section 2.8.

Theorem 2.2.10 (Benign overfitting of SGD with tail-averaging)**.** Consider SGD with tail-averaging. Suppose Assumptions 2.2.1-2.2.3 hold and that the stepsize is set so that $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$. Then the excess risk can be upper bounded as follows,

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \le 2 \cdot \text{EffectiveBias} + 2 \cdot \text{EffectiveVar},$$

where

$$\begin{aligned}
\text{EffectiveBias} &= \frac{1}{\gamma^2 N^2} \cdot \left\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\right\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \left\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\right\|^2_{\mathbf{H}_{k^*:\infty}} \\
\text{EffectiveVar} &= \frac{4\alpha\left(\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^\dagger}} + (s+N)\gamma\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}\right)}{N\gamma(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))} \cdot \left(\frac{k^*}{N} + N\gamma^2\sum_{i>k^*}\lambda_i^2\right) \\
&\quad + \frac{\sigma^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \left(\frac{k^*}{N} + \gamma\cdot\sum_{k^*<i\le k^\dagger}\lambda_i + (s+N)\gamma^2\cdot\sum_{i>k^\dagger}\lambda_i^2\right),
\end{aligned}$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{\gamma N}\}$ and $k^\dagger = \max\{k : \lambda_k \ge \frac{1}{\gamma(s+N)}\}$.

Theorem 2.2.10 shows that tail-averaging has improvements over iterate-averaging. This agrees with the results shown in [JNK17]: in the underparameterized regime ($N \gg d$) and for the strongly convex case ($\lambda_d > 0$), one can obtain substantially improved convergence rates on the bias term.

We also provide a lower bound on the excess risk for SGD with tail-averaging as a counterpart of Theorem 2.2.6, which shows that our upper bound is nearly tight. The proof is again deferred to Section 2.8.

Theorem 2.2.11 (Excess risk lower bound, tail-averaging)**.** Consider SGD with tail-averaging. Suppose $N \geq 500$. For any well-specified data distribution $\mathcal{D}$ (see (2.2.1)) that also satisfies Assumptions 2.2.1, 2.2.2 and 2.2.5, for any stepsize such that $\gamma < 1/\lambda_1$, we have that:

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \geq \frac{1}{100\gamma^2 N^2} \cdot \|(\mathbf{I} - \gamma\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \frac{\|(\mathbf{I} - \gamma\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}_{k^*:\infty}}}{100}$$
$$+ \frac{\beta\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}}{10^4}\left(\frac{k^*}{N} + N\gamma^2\sum_{i>k^*}\lambda_i^2\right)$$
$$+ \frac{\sigma^2_{\text{noise}}}{600}\left(\frac{k^*}{N} + \gamma\sum_{k^*<i\leq k^\dagger}\lambda_i + (s+N)\gamma^2\sum_{i>k^\dagger}\lambda_i^2\right),$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$ and $k^\dagger = \max\{k : \lambda_k \geq \frac{1}{(s+N)\gamma}\}$.

Comparing our upper and lower bounds, they are matching (upto absolute constants) for most of the terms, except for the first effective variance term, where a $||\mathbf{w}_0 - \mathbf{w}^*||^2_{\mathbf{I}_{0:k^\dagger}}$ is lost (suppose that $s = \Theta(N)$). Our conjecture is that the upper bound is improvable in this regard. Obtaining matching upper and lower bounds for SGD with tail-averaging is left as a direction for future work.

2.3 Further Related Work

We first discuss the work on iterate averaging in the finite dimensional case before turning to the over-parameterized regime. In the underparameterized regime, where d is assumed to be finite, the behavior of constant stepsize SGD with iterate average or tail average has been well investigated from the perspective of the *bias-variance decomposition* [DB15a; DFB17; LS18; JKK18a; JNK17]. For iterate averaging from the beginning, [DB15a; DFB17] show a $O(1/N^2)$ convergence rate for the bias error and a $O(d/N)$ convergence rate for the variance error, where N is the number of observed samples and d is the number of parameters. The bias error rate can be further improved by considering averaging only the tail iterates [JKK18a; JNK17; JKK18b], provided that the minimal eigenvalue of $\mathbf{H}$ is bounded away

from 0. We note that the work in [JKK18a; JNK17; JKK18b] also give the optimal rates with model misspecification. These results all have dimension factors d and do not apply to the over-parameterized regime, though our results recover the finite dimensional case (and the results for delayed tail averaging from [JKK18a; JNK17] can be applied here for the bias term). We further develop on the proof techniques in [JKK18a], where we use properties of asymptotic stationary distributions for the purposes of finite sample size analysis.

Another notable difference in our work is that Assumption 2.2.2 (which is implied by sub-Gaussianity, see Lemma 2.5.1) is somewhat stronger than what is often assumed for iterate average SGD analysis, where $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preceq R^2\mathbf{H}$, as adopted in [BM13; DB15a; DFB17; JKK18a; JNK17]. Our assumption implies an R^2 bound with $R^2 = \alpha \operatorname{tr}(\mathbf{H})$. In terms of analysis, we note that our variance analysis only relies on an R^2 condition, while our bias analysis relies on our stronger sub-Gaussianity-like assumption.

We now discuss related works in the over-parameterized regime [DB15b; BBG20]. Compared with [DB15b], our bounds apply to least square instances with *any* data covaraince spectrum (under Assumption 2.2.2), while [DB15b] only covered least square instances that have specific data covaraince spectrum (see A3 in [DB15b]). In comparison with [BBG20], their bounds rely on a weaker fourth moment assumption, but rely on a stronger true parameter assumption in that $||\mathbf{H}^{-\alpha}\mathbf{w}^*||_2$ must be finite, where $\alpha > 0$ is a constant (see Theorem 1 condition (a) in [BBG20]).

Our fourth moment assumption (Assumption 2.2.2) is a natural starting point for analyzing the over-parameterized regime because it also allows for direct comparisons to OLS and ridge regression, as discussed above.

Concurrent to this work, [CLT20] provide dimension independent bounds for averaged SGD; their excess risk bounds for linear regression are not as sharp as those provided here.

2.4 Proof Outline

We now provide the high level ideas in the proof. A key idea is relating the finite sample (infinite dimensional) covariance matrices of the variance and bias stochastic processes to those of their corresponding (asymptotic) stationary covariance matrices — an idea developed in [JKK18a] for the finite dimensional, variance analysis.

This section is organized as follows: Section 2.4.1 introduces additional notation and relevant linear operators; Section 2.4.2 presents a refined bound on a now standard bias-variance decomposition; Section 2.4.3 outlines the variance error analysis, followed by Section 2.4.4 outlining the bias error analysis. Complete proofs of the upper and lower bounds are provided in the Section 2.6 and Section 2.7, respectively.

2.4.1 Preliminaries

For two matrices $\mathbf{A}$ and $\mathbf{B}$, their inner product is defined as $\langle \mathbf{A}, \mathbf{B} \rangle := \operatorname{tr}\left(\mathbf{A}^\top \mathbf{B}\right)$. The following properties will be used frequently: if $\mathbf{A}$ is PSD, and $\mathbf{B} \succeq \mathbf{B}'$, then $\langle \mathbf{A}, \mathbf{B} \rangle \geq \langle \mathbf{A}, \mathbf{B}' \rangle$. We use $\otimes$ to denote the kronecker/tensor product. We define the following linear operators:

$$\mathcal{I} = \mathbf{I} \otimes \mathbf{I}, \quad \mathcal{M} = \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}], \quad \widetilde{\mathcal{M}} = \mathbf{H} \otimes \mathbf{H},$$
$$\mathcal{T} = \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H} - \gamma \mathcal{M}, \quad \widetilde{\mathcal{T}} = \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H} - \gamma \mathbf{H} \otimes \mathbf{H}.$$

We use the notation $\mathcal{O} \circ \mathbf{A}$ to denotes the operator $\mathcal{O}$ acting on a symmetric matrix $\mathbf{A}$. For example, with these definitions, we have that for a symmetric matrix $\mathbf{A}$,

$$\mathcal{I} \circ \mathbf{A} = \mathbf{A}, \quad \mathcal{M} \circ \mathbf{A} = \mathbb{E}[(\mathbf{x}^\top \mathbf{A} \mathbf{x}) \mathbf{x}\mathbf{x}^\top], \quad \widetilde{\mathcal{M}} \circ \mathbf{A} = \mathbf{H}\mathbf{A}\mathbf{H},$$
$$(\mathcal{I} - \gamma \mathcal{T}) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \gamma \mathbf{x}\mathbf{x}^\top)\mathbf{A}(\mathbf{I} - \gamma \mathbf{x}\mathbf{x}^\top)], \quad (\mathcal{I} - \gamma \widetilde{\mathcal{T}}) \circ \mathbf{A} = (\mathbf{I} - \gamma \mathbf{H})\mathbf{A}(\mathbf{I} - \gamma \mathbf{H}). \tag{2.4.1}$$

We conclude by summarizing a few technical properties of these operators (see Lemma 2.6.1).

Lemma 2.4.1. An operator $\mathcal{O}$ defined on symmetric matrices is called PSD mapping, if $\mathbf{A} \succeq 0$ implies $\mathcal{O} \circ \mathbf{A} \succeq 0$. Then we have

1. $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ are both PSD mappings.
2. $\mathcal{I} - \gamma\mathcal{T}$ and $\mathcal{I} - \gamma\widetilde{\mathcal{T}}$ are both PSD mappings.
3. $\mathcal{M} - \widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{T}} - \mathcal{T}$ are both PSD mappings.
4. If $0 < \gamma \leq 1/\lambda_1$, then $\widetilde{\mathcal{T}}^{-1}$ exists, and is a PSD mapping.
5. If $0 < \gamma \leq 1/(\alpha \operatorname{tr}(\mathbf{H}))$, then $\mathcal{T}^{-1} \circ \mathbf{A}$ exists for PSD matrix $\mathbf{A}$, and $\mathcal{T}^{-1}$ is a PSD mapping.

2.4.2 The Bias-Variance Decomposition

It is helpful to consider the bias-variance decomposition for averaged SGD, which has been extensively studied before in the underparameterized regime ($N \gg d$) [DB15b; JNK17; JKK18a]. For convenience, we define the centered SGD iterate as $\boldsymbol{\beta}_t := \mathbf{w}_t - \mathbf{w}^*$. Similarly we define $\bar{\boldsymbol{\beta}}_N := \frac{1}{N}\sum_{t=0}^{N-1} \boldsymbol{\beta}_t$.

(1) If the sampled data contains no label noise, i.e., $y_t = \langle \mathbf{w}^*, \mathbf{x}_t \rangle$, then the obtained SGD iterates $\{\boldsymbol{\beta}_t^{\text{bias}}\}$ reveal the *bias error*,

$$\boldsymbol{\beta}_t^{\text{bias}} = \left(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top\right) \boldsymbol{\beta}_{t-1}^{\text{bias}}, \qquad \boldsymbol{\beta}_0^{\text{bias}} = \boldsymbol{\beta}_0. \tag{2.4.2}$$

(2) If the iterates are initialized from the optimal $\mathbf{w}^*$, i.e., $\mathbf{w}_0 = \mathbf{w}^*$, then the obtained SGD iterates $\{\boldsymbol{\beta}_t^{\text{variance}}\}$ reveal the *variance error*,

$$\boldsymbol{\beta}_t^{\text{variance}} = \left(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top\right) \boldsymbol{\beta}_{t-1}^{\text{variance}} + \gamma \xi_t \mathbf{x}_t, \qquad \boldsymbol{\beta}_0^{\text{variance}} = \mathbf{0}, \tag{2.4.3}$$

where $\xi_t := y_t - \langle \mathbf{w}^*, \mathbf{x}_t \rangle$ is the inherent noise. Note the "bias iterates" can be viewed as a stochastic process of SGD on a consistent linear system; similarly, the "variance iterates" should be treated as a stochastic process of SGD initialized from the optimum.

Using the defined operators, the update rule of the iterates (2.4.2) imply the following recursive form of $\mathbf{B}_t := \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{bias}}]$:

$$\mathbf{B}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{B}_{t-1} \qquad \text{and} \qquad \mathbf{B}_0 = \boldsymbol{\beta}_0 \otimes \boldsymbol{\beta}_0, \tag{2.4.4}$$

and the update rule (2.4.3) imply the following recursive form of $\mathbf{C}_t := \mathbb{E}[\boldsymbol{\beta}_t^{\text{variance}} \otimes \boldsymbol{\beta}_t^{\text{variance}}]$:

$$\mathbf{C}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2\boldsymbol{\Sigma}, \qquad \mathbf{C}_0 = \mathbf{0}. \tag{2.4.5}$$

We define the averaged version of $\boldsymbol{\beta}_t^{\text{bias}}$ and $\boldsymbol{\beta}_t^{\text{variance}}$ in the same way as $\overline{\mathbf{w}}_N$, i.e., $\bar{\boldsymbol{\beta}}_N^{\text{bias}} := \frac{1}{N}\sum_{t=0}^{N-1}\boldsymbol{\beta}_t^{\text{bias}}$ and $\bar{\boldsymbol{\beta}}_N^{\text{variance}} := \frac{1}{N}\sum_{t=0}^{N-1}\boldsymbol{\beta}_t^{\text{variance}}$. With a little abuse of probability space, from (2.1.2), (2.4.2) and (2.4.3) we have that

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_t^{\text{bias}} + \boldsymbol{\beta}_t^{\text{variance}},$$

then an application of Cauchy–Schwarz inequality leads to the following *bias-variance decomposition* on the excess risk (see [JNK17], also Lemma 2.6.2):

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N \otimes \bar{\boldsymbol{\beta}}_N]\rangle \le \left(\sqrt{\text{bias}} + \sqrt{\text{variance}}\right)^2, \tag{2.4.6}$$

$$\text{where} \quad \text{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}]\rangle, \quad \text{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{variance}}]\rangle.$$

In the above bound, the two terms are usually referred to as the *bias error* and the *variance error* respectively. Furthermore, expanding the kronecker product between the two averaged iterates, and doubling the squared terms, we have the following upper bounds on the bias error and the variance error (see Lemma 2.6.3 for the proof):

$$\text{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}]\rangle \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_t\big\rangle, \tag{2.4.7}$$

$$\text{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{variance}}] \le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_t\big\rangle. \tag{2.4.8}$$

Note that in the above bounds, we keep both summations in finite steps, and this makes our analysis sharp as $N \ll d$. In comparison, [JKK18a; JNK17] take the inner summation to infinity, which yields looser upper bounds for further analysis in the over-parameterized setting. Next we bound the two error terms (2.4.7) and (2.4.8) separately.

2.4.3 Bounding the Variance Error

We would like to point out that in the analysis of the variance error (2.4.8), Assumption 2.2.2 can be replaced by a weaker assumption: $\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{x}\mathbf{x}^\top] \preceq R^2\mathbf{H}$, where R is a positive

constant [JNK17; JKK18a; DFB17]. A proof under the weaker assumption can be found in Section 2.6.3. Here, for consistency, we sketch the proof under Assumption 2.2.2.

To upper bound (2.4.8), noticing that $(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}$ is PSD, it suffices to upper bound $\mathbf{C}_t$ in PSD sense. In particular, by Lemma 5 in [JKK18a] (restated in Lemma 2.6.5), the sequence $\{\mathbf{C}_t\}_{t=0,\dots}$ has the following property,

$$0=\mathbf{C}_0\preceq\mathbf{C}_1\preceq\cdots\preceq\mathbf{C}_\infty\preceq\frac{\gamma\sigma^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\mathbf{I}. \tag{2.4.9}$$

This gives a uniform but crude upper bound on $\mathbf{C}_t$ for all $t\geq 0$. However, a direct application of this crude bound to (2.4.8) cannot give a sharp rate in the over-parameterized setting. Instead, we seek to refine the bound of $\mathbf{C}_t$ based on its update rule in (2.4.5) (see the proof of Lemma 2.6.6 for details):

$$\begin{aligned}
\mathbf{C}_t&=(\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{C}_{t-1}+\gamma^2\boldsymbol{\Sigma}\\
&=(\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1}+\gamma^2(\mathcal{M}-\widetilde{\mathcal{M}})\circ\mathbf{C}_{t-1}+\gamma^2\boldsymbol{\Sigma}\\
&\preceq(\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1}+\gamma^2\mathcal{M}\circ\mathbf{C}_{t-1}+\gamma^2\boldsymbol{\Sigma}\qquad\text{(since }\widetilde{\mathcal{M}}\text{ is a PSD mapping)}\\
&\preceq(\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1}+\frac{\gamma^3\sigma^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\mathcal{M}\circ\mathbf{I}+\gamma^2\boldsymbol{\Sigma},\quad\text{(by (2.4.9) and }\mathcal{M}\text{ is a PSD mapping)}\\
&\preceq(\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1}+\frac{\gamma^3\sigma^2\alpha\operatorname{tr}(\mathbf{H})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\mathbf{H}+\gamma^2\sigma^2\mathbf{H},\qquad\text{(by Assumptions 2.2.2 and 2.2.3)}\\
&=(\mathcal{I}-\gamma\widetilde{\mathcal{T}})\circ\mathbf{C}_{t-1}+\frac{\gamma^2\sigma^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\mathbf{H}.
\end{aligned}$$

Solving the above recursion, we obtain the following refined upper bound for $\mathbf{C}_t$:

$$\begin{aligned}
\mathbf{C}_t&\preceq\frac{\gamma^2\sigma^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\sum_{k=0}^{t-1}(\mathcal{I}-\gamma\widetilde{\mathcal{T}})^k\circ\mathbf{H}\\
&=\frac{\gamma^2\sigma^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\sum_{k=0}^{t-1}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{H}(\mathbf{I}-\gamma\mathbf{H})^k\qquad\text{(by the property of }\mathcal{I}-\gamma\widetilde{\mathcal{T}}\text{ in (2.4.1))}\\
&\preceq\frac{\gamma^2\sigma^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\sum_{k=0}^{t-1}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{H}=\frac{\gamma\sigma^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\cdot\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^t\big).
\end{aligned}\tag{2.4.10}$$

Now we can plug the above refined upper bound (2.4.10) into (2.4.8), and obtain

$$\begin{aligned}\text{variance} &\leq \frac{\sigma^2}{N^2(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\sum_{t=0}^{N-1}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{t}\right\rangle \\ &= \frac{\sigma^2}{N^2(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\sum_{t=0}^{N-1}\sum_{i}\left(1-(1-\gamma\lambda_i)^{N-t}\right)\left(1-(1-\gamma\lambda_i)^{t}\right) \\ &\leq \frac{\sigma^2}{N^2(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot N\cdot\sum_{i}\left(1-(1-\gamma\lambda_i)^{N}\right)^2. \end{aligned} \tag{2.4.11}$$

The remaining effort is to precisely control the summations in (2.4.11) according to the scale of the eigenvalues: for large eigenvalues $\lambda_i \geq \frac{1}{N\gamma}$, which appear at most k^* times, we use $1-(1-\gamma\lambda_i)^N \leq 1$; and for the remaining small eigenvalues $\lambda_i < \frac{1}{N\gamma}$, we use $1-(1-\gamma\lambda_i)^N \leq \mathcal{O}\left(N\gamma\lambda_i\right)$. Plugging these into (2.4.11) gives us the final full spectrum upper bound on the variance error (see the proof of Lemma 2.6.7 for more details). This bound contributes to part of EffectiveVar in Theorem 2.2.4.

2.4.4 Bounding the Bias Error

Next we discuss how to bound the bias error (2.4.7). A natural idea is to follow the same way in analyzing the variance error, and derive a similar bound on $\mathbf{B}_t$. Yet a fundamental difference between the variance sequence (2.4.5) and the bias sequence (2.4.4) is that: $\mathbf{C}_t$ is increasing, while $\mathbf{B}_t$ is "contracting", hence applying the same procedure in the variance error analysis cannot lead to a tight bound on $\mathbf{B}_t$. Instead, observing that $\mathbf{S}_t := \sum_{k=0}^{t-1}\mathbf{B}_k$, the summation of a contracting sequence, is increasing in the PSD sense. Particularly, we can rewrite $\mathbf{S}_t$ in the following recursive form

$$\mathbf{S}_t = (\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{S}_{t-1}+\mathbf{B}_0, \tag{2.4.12}$$

which resembles that of $\mathbf{C}_t$ in (2.4.5). This motivates us to: (i) express the obtained bias error bound (2.4.7) by $\mathbf{S}_t$, and (ii) derive a tight upper bound on $\mathbf{S}_t$ using similar analysis for the variance error.

For (i), by some linear algebra manipulation (see the derivation of (2.6.13)), we can bound (2.4.7) as follows:

$$\text{bias} \leq \frac{1}{\gamma N^2}\big\langle \mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N, \sum_{t=0}^{N-1}\mathbf{B}_t\big\rangle = \frac{1}{\gamma N^2}\big\langle \mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N, \mathbf{S}_N\big\rangle. \tag{2.4.13}$$

For (ii), we first show that $\{\mathbf{S}_t\}_{t=1,\dots,N}$ is increasing and has a crude upper bound (see Lemmas 2.6.8 and 2.6.10):

$$\mathbf{B}_0 = \mathbf{S}_1 \preceq \mathbf{S}_2 \preceq \cdots \preceq \mathbf{S}_N, \quad \text{and} \quad \mathcal{M}\circ\mathbf{S}_N \preceq \frac{\alpha\cdot\operatorname{tr}\Big(\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{2N}\big)\mathbf{B}_0\Big)}{\gamma(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))}\cdot\mathbf{H}. \tag{2.4.14}$$

Then similar to our previous procedure in bounding $\mathbf{C}_t$, we can tighten the upper bound on $\mathbf{S}_t$ by its recursive form (2.4.12) and the crude bound ($\mathcal{M}\circ\mathbf{S}_{N-1}$ in (2.4.14)), and obtain the following refined bound (see Lemma 2.6.11) for $\mathbf{S}_N$:

$$\mathbf{S}_N \preceq \sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^k + \frac{\gamma\alpha\cdot\operatorname{tr}\Big(\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{2N}\big)\mathbf{B}_0\Big)}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{2k}\mathbf{H}. \tag{2.4.15}$$

The remaining proof will be similar to what we have done for the variance error bound: substituting (2.4.15) into (2.4.13) gives an upper bound on the bias error with respect to the summations over functions of eigenvalues. Then by carefully controlling each summation according to the scale of the corresponding eigenvalues, we will obtain a tight full spectrum upper bound on the bias error (see the proof of Lemma 2.6.12 for more details).

As a final remark, noticing that different from the upper bound of $\mathbf{C}_t$ in (2.4.10), the upper bound for $\mathbf{S}_t$ in (2.4.15) consists of two terms. The first term will contribute to the EffectiveBias term in Theorem 2.2.4, while the second term will be merged to the bound of the variance error and contribute to the EffectiveVar term in Theorem 2.2.4.

2.5 Examples of Assumption 2.2.2

[HKZ14; BLL20; TB20] assume that $\mathbf{z} := \mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ is sub-Gaussian. The following lemma shows that our Assumption 2.2.2 is implied by assuming sub-Gaussianity.

Lemma 2.5.1. Suppose $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{H}$, and $\mathbf{z} := \mathbf{H}^{-\frac{1}{2}}\mathbf{x}$ is σ_z^2-sub-Gaussian random vector, then for any PSD matrix $\mathbf{A}$, we have

$$\mathbb{E}[(\mathbf{x}^\top \mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}^\top] \preceq 16\sigma_z^4 \operatorname{tr}(\mathbf{A}\mathbf{H})\mathbf{H}.$$

Proof. Note that $\mathbf{z}$ is a σ_z^2-sub-Gaussian random vector with identity covariance matrix, implying that for any fixed unit vector $\mathbf{u}$ that $\mathbf{u}^\top\mathbf{z}$ is a σ_z^2-sub-Gaussian random variable. Then we have the following inequality for any unit vectors $\mathbf{u}$ and $\mathbf{v}$

$$\mathbb{E}[(\mathbf{u}^\top\mathbf{z})^2(\mathbf{v}^\top\mathbf{z})^2] \le \sqrt{\mathbb{E}[(\mathbf{u}^\top\mathbf{z})^4]} \cdot \sqrt{\mathbb{E}[(\mathbf{v}^\top\mathbf{z})^4]} \le \max\left\{\mathbb{E}[(\mathbf{u}^\top\mathbf{z})^4], \mathbb{E}[(\mathbf{v}^\top\mathbf{z})^4]\right\} \le 16 \cdot \sigma_z^4,$$

where the first inequality follows from the Cauchy–Schwarz inequality; and the last inequality uses the fact that $\mathbf{u}^\top\mathbf{z}$ is σ_z^2 sub-Gaussian. Here, the factor 16 is due to the sub-Gaussian property (Proposition 2.5.2, [Ver18]). Next, for any PSD matrix $\mathbf{A}$, suppose its eigenvalue decomposition is $\mathbf{A} = \sum_i \mu_i \mathbf{u}_i\mathbf{u}_i^\top$, where $\mu_i \ge 0$ is the eigenvalue and $\mathbf{u}_i$ is the corresponding eigenvector, we have

$$\mathbb{E}[(\mathbf{z}^\top\mathbf{A}\mathbf{z})\mathbf{z}\mathbf{z}^\top] = \sum_i \mu_i \mathbb{E}[(\mathbf{u}_i^\top\mathbf{z})^2\mathbf{z}\mathbf{z}^\top]. \tag{2.5.1}$$

For any unit vector $\mathbf{v}$, we have:

$$\mathbf{v}^\top\mathbb{E}[(\mathbf{z}^\top\mathbf{A}\mathbf{z})\mathbf{z}\mathbf{z}^\top]\mathbf{v} = \sum_i \mu_i \mathbb{E}[(\mathbf{u}_i^\top\mathbf{z})^2(\mathbf{v}^\top\mathbf{z})^2] \le 16 \cdot \sigma_z^4 \cdot \sum_i \mu_i = 16 \cdot \sigma_z^4 \operatorname{tr}(\mathbf{A}).$$

This implies that for any PSD matrix $\mathbf{A}$ we have

$$\mathbb{E}[(\mathbf{z}^\top\mathbf{A}\mathbf{z})\mathbf{z}\mathbf{z}^\top] \le 16 \cdot \sigma_z^4 \operatorname{tr}(\mathbf{A})\mathbf{I}. \tag{2.5.2}$$

Finally considering $\mathbf{x} = \mathbf{H}^{\frac{1}{2}}\mathbf{z}$, we have for any PSD matrix $\mathbf{A}$:

$$\begin{aligned}
\mathbb{E}[(\mathbf{x}^\top\mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}^\top] &= \mathbb{E}[(\mathbf{z}^\top\mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}}\mathbf{z})\mathbf{H}^{\frac{1}{2}}\mathbf{z}\mathbf{z}^\top\mathbf{H}^{\frac{1}{2}}] \\
&= \mathbf{H}^{\frac{1}{2}}\mathbb{E}[(\mathbf{z}^\top\mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}}\mathbf{z})\mathbf{z}\mathbf{z}^\top]\mathbf{H}^{\frac{1}{2}} \\
&\preceq \mathbf{H}^{\frac{1}{2}} \cdot 16\sigma_z^4 \operatorname{tr}(\mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}}) \cdot \mathbf{I} \cdot \mathbf{H}^{\frac{1}{2}} \\
&= 16\sigma_z^4 \operatorname{tr}(\mathbf{A}\mathbf{H})\mathbf{H},
\end{aligned}$$

where the second line holds since $\mathbf{z}^\top\mathbf{H}^{\frac{1}{2}}\mathbf{A}\mathbf{H}^{\frac{1}{2}}\mathbf{z}$ is a scalar and the third line of the above equation is due to (2.5.2). This concludes the proof. □

2.6 Proofs of the Upper Bounds

2.6.1 Technical Lemma

Lemma 2.6.1 (Restatement of Lemma 2.4.1)**.** An operator $\mathcal{O}$ defined on symmetric matrices is called PSD mapping, if $\mathbf{A} \succeq 0$ implies $\mathcal{O} \circ \mathbf{A} \succeq 0$. Then we have

1. $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ are both PSD mappings.
2. $\mathcal{I} - \gamma\mathcal{T}$ and $\mathcal{I} - \gamma\widetilde{\mathcal{T}}$ are both PSD mappings.
3. $\mathcal{M} - \widetilde{\mathcal{M}}$ and $\widetilde{\mathcal{T}} - \mathcal{T}$ are both PSD mappings.
4. If $0 < \gamma < 1/\lambda_1$, then $\widetilde{\mathcal{T}}^{-1}$ exists, and is a PSD mapping.
5. If $0 < \gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$, then $\mathcal{T}^{-1} \circ \mathbf{A}$ exists for PSD matrix $\mathbf{A}$, and $\mathcal{T}^{-1}$ is a PSD mapping.

Proof. The following proofs are summarized from [JKK18a; JNK17], and we include them here for completeness.

1. For any PSD matrix $\mathbf{A} \succeq 0$, by definition, we have

$$\mathcal{M} \circ \mathbf{A} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A} \mathbf{x}\mathbf{x}^\top] \succeq 0,$$
$$\widetilde{\mathcal{M}} \circ \mathbf{A} = \mathbf{H}\mathbf{A}\mathbf{H} \succeq 0.$$

   Therefore, both $\mathcal{M}$ and $\widetilde{\mathcal{M}}$ are PSD mappings.

2. For any PSD matrix $\mathbf{A} \succeq 0$, we have

$$(\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \gamma\mathbf{x}\mathbf{x}^\top)\mathbf{A}(\mathbf{I} - \gamma\mathbf{x}\mathbf{x}^\top)] \succeq 0,$$
$$(\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{A} = (\mathbf{I} - \gamma\mathbf{H})\mathbf{A}(\mathbf{I} - \gamma\mathbf{H}) \succeq 0.$$

   Hence, $\mathcal{I} - \gamma\mathcal{T}$ and $\mathcal{I} - \gamma\widetilde{\mathcal{T}}$ are both PSD mapping.

3. For any PSD matrix $\mathbf{A} \succeq 0$,

$$(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{A} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A} \mathbf{x}\mathbf{x}^\top] - \mathbf{H}\mathbf{A}\mathbf{H} = \mathbb{E}[(\mathbf{x}\mathbf{x}^\top - \mathbf{H})\mathbf{A}(\mathbf{x}\mathbf{x}^\top - \mathbf{H})] \succeq 0.$$

Thus, $\widetilde{\mathcal{T}} - \mathcal{T} = \mathcal{M} - \widetilde{\mathcal{M}}$ is PSD.

4. According to (2.4.1), if $0 < \gamma < 1/\lambda_1$, $\mathbf{I} - \gamma\mathbf{H}$ is a contraction map, thus for any symmetric matrix $\mathbf{A}$, the following exists:

$$\sum_{t=0}^{\infty} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{A} = \sum_{t=0}^{\infty} (\mathbf{I} - \gamma\mathbf{H})^t \mathbf{A} (\mathbf{I} - \gamma\mathbf{H})^t.$$

Therefore, $\sum_{t=0}^{\infty} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t$ exists and $\widetilde{\mathcal{T}}^{-1} = \gamma \sum_{t=0}^{\infty} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t$ exists. Furthermore, for any PSD matrix $\mathbf{A} \succeq 0$, we have

$$\widetilde{\mathcal{T}}^{-1} \circ \mathbf{A} = \gamma \sum_{t=0}^{\infty} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{A} = \gamma \sum_{t=0}^{\infty} (\mathbf{I} - \gamma\mathbf{H})^t \mathbf{A} (\mathbf{I} - \gamma\mathbf{H})^t \succeq 0,$$

which implies $\widetilde{\mathcal{T}}^{-1}$ is a PSD mapping.

5. For any finite PSD matrix $\mathbf{A}$, consider the following identity

$$\mathcal{T}^{-1} \circ \mathbf{A} = \gamma \sum_{t=0}^{\infty} (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A}.$$

Clearly, if the right hand side exists, it must be PSD since $\mathcal{I} - \gamma\mathcal{T}$ is a PSD mapping. It remains to show that $\sum_{t=0}^{\infty} (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A}$ is finite, and it suffices to show that

$$\operatorname{tr}\left(\sum_{t=0}^{\infty} (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A} \right) = \sum_{t=0}^{\infty} \operatorname{tr}\left((\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A} \right) < \infty.$$

Based on the definition of $\mathcal{T}$, let $\mathbf{A}_t = (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A}$, we have

$$\begin{aligned} \operatorname{tr}(\mathbf{A}_t) &= \operatorname{tr}(\mathbf{A}_{t-1}) - \gamma \operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1}) - \gamma \operatorname{tr}(\mathbf{A}_{t-1}\mathbf{H}) + \gamma^2 \operatorname{tr}\left(\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A} \mathbf{x}\mathbf{x}^\top] \right) \\ &= \operatorname{tr}(\mathbf{A}_{t-1}) - 2\gamma \operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1}) + \gamma^2 \operatorname{tr}\left(\mathbf{A}_{t-1} \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \right). \end{aligned} \tag{2.6.1}$$

By Assumption 2.2.2, we have $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preceq \alpha \operatorname{tr}(\mathbf{H})\mathbf{H}$. Therefore, it follows that

$$\begin{aligned} \operatorname{tr}(\mathbf{A}_t) &\leq \operatorname{tr}(\mathbf{A}_{t-1}) - (2\gamma - \gamma^2 \alpha \operatorname{tr}(\mathbf{H})) \operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1}) \\ &\leq \operatorname{tr}\left((\mathbf{I} - \gamma\mathbf{H})\mathbf{A}_{t-1} \right) \\ &\leq (1 - \gamma\lambda_d) \operatorname{tr}(\mathbf{A}_{t-1}), \end{aligned} \tag{2.6.2}$$

where we use the assumption $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$ in the first inequality. This further implies that

$$\sum_{t=0}^{\infty} \operatorname{tr}\left((\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{A}\right) = \sum_{t=0}^{\infty} \operatorname{tr}(\mathbf{A}_t) \leq \frac{\operatorname{tr}(\mathbf{A})}{\gamma\lambda_d} < \infty.$$

Therefore, $\mathcal{T}^{-1} \circ \mathbf{A}$ exists, and is PSD. So $\mathcal{T}^{-1}$ is a PSD mapping.

□

2.6.2 Bias-Variance Decomposition

Lemma 2.6.2 (Bias-variance decomposition)**.**

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N \otimes \bar{\boldsymbol{\beta}}_N]\rangle \leq \left(\sqrt{\text{bias}} + \sqrt{\text{variance}}\right)^2,$$

where

$$\text{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}]\rangle, \qquad \text{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{variance}}]\rangle.$$

Proof. This proof comes from [JKK18a]. For completeness we included it here.

With a slight abuse of notations (or probability spaces), we have $\boldsymbol{\beta}_t = \boldsymbol{\beta}_t^{\text{bias}} + \boldsymbol{\beta}_t^{\text{variance}}$, where the randomness of $\boldsymbol{\beta}_t^{\text{bias}}$ and $\boldsymbol{\beta}_t^{\text{variance}}$ is understood as coming from the same probability space as $\boldsymbol{\beta}_t$. This implies $\bar{\boldsymbol{\beta}}_t = \bar{\boldsymbol{\beta}}_t^{\text{bias}} + \bar{\boldsymbol{\beta}}_t^{\text{variance}}$. Then we have

$$\begin{aligned}
&\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \\
&= \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N \otimes \bar{\boldsymbol{\beta}}_N]\rangle \\
&= \mathbb{E}\left[\frac{1}{\sqrt{2}}\bar{\boldsymbol{\beta}}_N^\top \cdot \mathbf{H} \cdot \frac{1}{\sqrt{2}}\bar{\boldsymbol{\beta}}_N\right] \\
&\leq \left(\sqrt{\mathbb{E}\left[\left(\frac{1}{\sqrt{2}}\bar{\boldsymbol{\beta}}_N^{\text{bias}}\right)^\top \cdot \mathbf{H} \cdot \frac{1}{\sqrt{2}}\bar{\boldsymbol{\beta}}_N^{\text{bias}}\right]} + \sqrt{\mathbb{E}\left[\left(\frac{1}{\sqrt{2}}\bar{\boldsymbol{\beta}}_N^{\text{variance}}\right)^\top \cdot \mathbf{H} \cdot \frac{1}{\sqrt{2}}\bar{\boldsymbol{\beta}}_N^{\text{variance}}\right]}\right)^2 \\
&= \left(\sqrt{\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}]\rangle} + \sqrt{\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{variance}}]\rangle}\right)^2,
\end{aligned}$$

where we use Cauchy–Schwarz inequality in the inequality such that for any vector $\mathbf{u}$ and $\mathbf{v}$, $\mathbb{E}||\mathbf{u}+\mathbf{v}||_{\mathbf{H}}^2 \leq \left(\sqrt{\mathbb{E}||\mathbf{u}||_{\mathbf{H}}^2}+\sqrt{\mathbb{E}||\mathbf{v}||_{\mathbf{H}}^2}\right)^2$. □

Lemma 2.6.3. Recall iterates (2.4.4) and (2.4.5). If the stepsize satisfies $\gamma \leq 1/\lambda_1$, the bias error and variance error are upper bounded respectively as follows:

$$\text{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}]\rangle \leq \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_t\big\rangle,$$
$$\text{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{variance}}] \leq \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_t\big\rangle.$$

Proof. The proof will largely rely on the calculation in [JNK17]. Firstly, based on the definitions of $\boldsymbol{\beta}_t^{\text{bias}}$ and $\boldsymbol{\beta}_t^{\text{bias}}$ provided in (2.4.2) and (2.4.3), we have

$$\mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}}|\boldsymbol{\beta}_{t-1}^{\text{bias}}] = \mathbb{E}[\mathbf{P}_t\boldsymbol{\beta}_{t-1}^{\text{bias}}|\boldsymbol{\beta}_{t-1}^{\text{bias}}] = (\mathbf{I}-\gamma\mathbf{H})\boldsymbol{\beta}_{t-1}^{\text{bias}}. \tag{2.6.3}$$
$$\mathbb{E}[\boldsymbol{\beta}_t^{\text{variance}}|\boldsymbol{\beta}_{t-1}^{\text{variance}}] = \mathbb{E}[\mathbf{P}_t\boldsymbol{\beta}_{t-1}^{\text{variance}} + \gamma\xi_t\mathbf{x}_t|\boldsymbol{\beta}_{t-1}^{\text{variance}}] = (\mathbf{I}-\gamma\mathbf{H})\boldsymbol{\beta}_{t-1}^{\text{variance}}. \tag{2.6.4}$$

Then regarding the quantity $\mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}]$, we have

$$\begin{aligned}
&\mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}] \\
&= \frac{1}{N^2}\cdot\left(\sum_{0\leq k\leq t\leq N-1}\mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}}\otimes\boldsymbol{\beta}_k^{\text{bias}}] + \sum_{0\leq t<k\leq N-1}\mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}}\otimes\boldsymbol{\beta}_k^{\text{bias}}]\right) \\
&\preceq \frac{1}{N^2}\cdot\left(\sum_{0\leq k\leq t\leq N-1}\mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}}\otimes\boldsymbol{\beta}_k^{\text{bias}}] + \sum_{0\leq t\leq k\leq N-1}\mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}}\otimes\boldsymbol{\beta}_k^{\text{bias}}]\right) \\
&= \frac{1}{N^2}\cdot\left(\sum_{0\leq k\leq t\leq N-1}(\mathbf{I}-\gamma\mathbf{H})^{t-k}\mathbb{E}[\boldsymbol{\beta}_k^{\text{bias}}\otimes\boldsymbol{\beta}_k^{\text{bias}}] + \sum_{0\leq t\leq k\leq N-1}\mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}}\otimes\boldsymbol{\beta}_t^{\text{bias}}](\mathbf{I}-\gamma\mathbf{H})^{k-t}\right) \\
&= \frac{1}{N^2}\cdot\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\Big((\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}}\otimes\boldsymbol{\beta}_t^{\text{bias}}] + \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}}\otimes\boldsymbol{\beta}_t^{\text{bias}}](\mathbf{I}-\gamma\mathbf{H})^{k-t}\Big),
\end{aligned} \tag{2.6.5}$$

where we use (2.6.3) for $k-t$ (or $t-k$) times in the second equality. Therefore, plugging

(2.6.5) into the inner product $\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}] \rangle$ and noticing $\mathbf{H}$ is PSD, we have

$$\begin{aligned}
&\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}] \rangle \\
&\leq \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle \mathbf{H}, (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{bias}}] + \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{bias}}](\mathbf{I} - \gamma \mathbf{H})^{k-t} \Big\rangle \\
&= \frac{1}{N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{bias}}] \Big\rangle
\end{aligned}$$

where the last equality holds since $\mathbf{H}$ and $(\mathbf{I} - \gamma \mathbf{H})^{k-t}$ commute.

By (2.6.4), we can similarly obtain the following for $\mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{variance}}]$,

$$\begin{aligned}
&\mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{variance}}] \\
&\preceq \frac{1}{N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big((\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbb{E}[\boldsymbol{\beta}_t^{\text{variance}} \otimes \boldsymbol{\beta}_t^{\text{variance}}] + \mathbb{E}[\boldsymbol{\beta}_t^{\text{variance}} \otimes \boldsymbol{\beta}_t^{\text{variance}}](\mathbf{I} - \gamma \mathbf{H})^{k-t} \Big),
\end{aligned}$$

which further leads to

$$\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{variance}}] \rangle \leq \frac{1}{N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbb{E}[\boldsymbol{\beta}_t^{\text{variance}} \otimes \boldsymbol{\beta}_t^{\text{variance}}] \Big\rangle.$$

This completes the proof.

□

2.6.3 Bounding the Variance Error

We first introduce a weaker assumption (compared with Assumption 2.2.2) on the data distribution, which is sufficient to get our desired results on the variance error.

Assumption 2.6.4. There exists a constant $R > 0$ such that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preceq R^2 \mathbf{H}$.

We make this assumption to emphasize that our variance analysis does not rely on stronger assumptions than those in a number of prior works for iterate averaged SGD [BM13; JNK17; BBG20]. Moreover, note that this assumption is implied by Assumption 2.2.2 by setting $\mathbf{A} = \mathbf{I}$, which gives $R^2 = \alpha \operatorname{tr}(\mathbf{H})$.

Recall the variance error upper bound in Lemma 2.6.3:

$$\text{variance} \leq \frac{1}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{C}_t \right\rangle.$$

We first have the following crude bound on $\mathbf{C}_t$.

Lemma 2.6.5. ([JKK18a] Lemma 5) Under Assumptions 2.2.1, 2.2.3 and 2.6.4, if the step-size satisfies $\gamma < 1/R^2$, it holds that

$$0 = \mathbf{C}_0 \preceq \mathbf{C}_1 \preceq \cdots \preceq \mathbf{C}_\infty \preceq \frac{\gamma \sigma^2}{1 - \gamma R^2} \mathbf{I}.$$

Proof. This lemma directly comes from Lemmas 3 and 5 in [JKK18a]. For completeness, a proof is included as follows.

We first show that $\mathbf{C}_t$ is increasing:

$$\begin{aligned}
\mathbf{C}_t &= (\mathcal{I} - \gamma \mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2 \mathbf{\Sigma} \\
&= \gamma^2 \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \mathcal{T})^k \circ \mathbf{\Sigma} \qquad \text{(solving the recursion)} \\
&= \mathbf{C}_{t-1} + \gamma^2 (\mathcal{I} - \gamma \mathcal{T})^{t-1} \circ \mathbf{\Sigma} \\
&\succeq \mathbf{C}_{t-1}. \qquad \text{(since } \mathcal{I} - \gamma \mathcal{T} \text{ is a PSD mapping by Lemma 2.4.1)}
\end{aligned}$$

Next we show that $\mathbf{C}_\infty$ exists. Since $\mathbf{C}_t$ is PSD and increasing, it suffices to show that $\operatorname{tr}(\mathbf{C}_t)$ can be bounded uniformly. For any $t \geq 1$, we have

$$\mathbf{C}_t = \gamma^2 \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \mathcal{T})^k \circ \mathbf{\Sigma} \preceq \gamma^2 \sum_{t=0}^{\infty} (\mathcal{I} - \gamma \mathcal{T})^t \circ \mathbf{\Sigma}. \tag{2.6.6}$$

Let $\mathbf{A}_t := (\mathcal{I} - \gamma \mathcal{T})^t \circ \mathbf{\Sigma}$, then $\mathbf{A}_t = (\mathcal{I} - \gamma \mathcal{T}) \circ \mathbf{A}_{t-1}$. By Assumption 2.6.4 we have $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preceq R^2 \mathbf{H}$. Then, by (2.6.1), we can get

$$\begin{aligned}
\operatorname{tr}(\mathbf{A}_t) &= \operatorname{tr}(\mathbf{A}_{t-1}) - 2\gamma \operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1}) + \gamma^2 \operatorname{tr}\left(\mathbf{A}_{t-1} \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top]\right) \\
&\leq \operatorname{tr}(\mathbf{A}_{t-1}) - (2\gamma - \gamma^2 R^2) \operatorname{tr}(\mathbf{H}\mathbf{A}_{t-1}) \\
&\leq \operatorname{tr}\left((\mathbf{I} - \gamma \mathbf{H}) \mathbf{A}_{t-1}\right) \\
&\leq (1 - \gamma \lambda_d) \operatorname{tr}(\mathbf{A}_{t-1}),
\end{aligned} \tag{2.6.7}$$

where we use the assumption $\gamma \leq 1/R^2$ in the second inequality. Combining (2.6.6) and (2.6.7), we have for any $t \geq 1$ that

$$\operatorname{tr}(\mathbf{C}_t) \leq \gamma^2 \sum_{t=0}^{\infty} \operatorname{tr}\left((\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{\Sigma}\right) = \gamma^2 \sum_{t=0}^{\infty} \operatorname{tr}(\mathbf{A}_t) \leq \frac{\gamma \operatorname{tr}(\mathbf{\Sigma})}{\lambda_d} < \infty.$$

Therefore, $\operatorname{tr}(\mathbf{C}_t)$ is uniformly upper bounded, hence $\mathbf{C}_\infty$ exists.

Finally we upper bound $\mathbf{C}_\infty$. Taking limits in (2.4.4), we have

$$\mathbf{C}_\infty = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{C}_\infty + \gamma^2\mathbf{\Sigma},$$

which immediately implies

$$\mathbf{C}_\infty = \gamma\mathcal{T}^{-1} \circ \mathbf{\Sigma}.$$

Recalling $\widetilde{\mathcal{T}} = \mathcal{T} + \gamma\mathcal{M} - \gamma\widetilde{\mathcal{M}}$ and the definitions and properties of the operators, we have

$$\begin{aligned}
\widetilde{\mathcal{T}} \circ \mathbf{C}_\infty &= \mathcal{T} \circ \mathbf{C}_\infty + \gamma\mathcal{M} \circ \mathbf{C}_\infty - \gamma\widetilde{\mathcal{M}} \circ \mathbf{C}_\infty \\
&= \gamma\mathbf{\Sigma} + \gamma\mathcal{M} \circ \mathbf{C}_\infty - \gamma\widetilde{\mathcal{M}} \circ \mathbf{C}_\infty \qquad (\text{since } \mathbf{C}_\infty = \gamma\mathcal{T}^{-1} \circ \mathbf{\Sigma}) \\
&\preceq \gamma\mathbf{\Sigma} + \gamma\mathcal{M} \circ \mathbf{C}_\infty \qquad (\text{since } \widetilde{\mathcal{M}} \text{ is a PSD mapping by Lemma 2.4.1}) \\
&\preceq \gamma\sigma^2\mathbf{H} + \gamma\mathcal{M} \circ \mathbf{C}_\infty. \qquad (\text{since } \mathbf{\Sigma} \preceq \sigma^2\mathbf{H} \text{ by Assumption 2.2.3})
\end{aligned}$$

Recall that $\widetilde{\mathcal{T}}^{-1}$ exists and is a PSD mapping by Lemma 2.4.1, we then have

$$\begin{aligned}
\mathbf{C}_\infty &\preceq \gamma\sigma^2 \cdot \widetilde{\mathcal{T}}^{-1} \circ \mathbf{H} + \gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{C}_\infty \\
&\preceq \gamma\sigma^2 \cdot \sum_{t=0}^{\infty} (\gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M})^t \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{H}. \qquad (\text{solving the recursion})
\end{aligned} \tag{2.6.8}$$

In addition, we have

$$\begin{aligned}
\widetilde{\mathcal{T}}^{-1} \circ \mathbf{H} &= \gamma \sum_{t=0}^{\infty} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{H} \\
&= \gamma \sum_{t=0}^{\infty} (\mathbf{I} - \gamma\mathbf{H})^t \mathbf{H} (\mathbf{I} - \gamma\mathbf{H})^t \qquad (\text{by the property of } \mathcal{I} - \widetilde{\mathcal{T}} \text{ in (2.4.1)}) \\
&\preceq \gamma \sum_{t=0}^{\infty} (\mathbf{I} - \gamma\mathbf{H})^t \mathbf{H} \\
&= \mathbf{I}.
\end{aligned} \tag{2.6.9}$$

Substituting (2.6.9) into (2.6.8), we obtain

$$\begin{aligned}
\mathbf{C}_\infty &\preceq \gamma\sigma^2 \cdot \sum_{t=0}^{\infty} (\gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M})^t \circ \mathbf{I} \\
&= \gamma\sigma^2 \cdot \sum_{t=0}^{\infty} (\gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M})^{t-1} \circ \gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{I} \\
&\preceq \gamma\sigma^2 \cdot \sum_{t=0}^{\infty} (\gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M})^{t-1} \circ \gamma R^2 \mathbf{I} \\
&\preceq \gamma\sigma^2 \cdot \sum_{t=0}^{\infty} (\gamma R^2)^t \mathbf{I} \\
&= \frac{\gamma\sigma^2}{1-\gamma R^2}\mathbf{I},
\end{aligned}$$

where the second inequality is due to $\mathcal{M} \circ \mathbf{I} \preceq R^2\mathbf{H}$ by Assumption 2.6.4 and $\widetilde{\mathcal{T}}^{-1} \circ \mathbf{H} \preceq \mathbf{I}$ in (2.6.9), and the third inequality is by recursion. This completes the proof. □

The following lemma refines the bound on $\mathbf{C}_t$ by its update rule and its crude bound shown in previous lemma.

Lemma 2.6.6. Under Assumptions 2.2.1, 2.2.3 and 2.6.4, if the stepsize satisfies $\gamma < 1/R^2$, it holds that

$$\mathbf{C}_t \preceq \frac{\gamma\sigma^2}{1-\gamma R^2} \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t\right).$$

Proof. By (2.4.5) and the definitions of $\mathcal{T}$ and $\widetilde{\mathcal{T}}$, we have

$$\begin{aligned}
\mathbf{C}_t &= (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2\mathbf{\Sigma} \\
&= (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{C}_{t-1} + \gamma^2\mathbf{\Sigma} \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2\mathcal{M} \circ \mathbf{C}_{t-1} + \gamma^2\mathbf{\Sigma},
\end{aligned} \tag{2.6.10}$$

where the last inequality is due to the fact that $\widetilde{\mathcal{M}}$ is a PSD mapping. Then by Lemma 2.6.5, we have for all $t \geq 0$,

$$\mathcal{M} \circ \mathbf{C}_t \preceq \mathcal{M} \circ \mathbf{C}_\infty \preceq \mathcal{M} \circ \frac{\gamma\sigma^2}{1-\gamma R^2}\mathbf{I} = \frac{\gamma\sigma^2}{1-\gamma R^2} \cdot \mathbb{E}[\|\mathbf{x}\|_2^2 \mathbf{x}\mathbf{x}^\top] \preceq \frac{\gamma R^2\sigma^2}{1-\gamma R^2} \cdot \mathbf{H}. \tag{2.6.11}$$

Substituting (2.6.11) and $\mathbf{\Sigma} \preceq \left\|\mathbf{H}^{-1/2}\mathbf{\Sigma}\mathbf{H}^{-1/2}\right\|_2 \cdot \mathbf{H}$ into (2.6.10), we obtain

$$
\begin{aligned}
\mathbf{C}_t &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2 \cdot \frac{\gamma R^2 \sigma^2}{1-\gamma R^2} \cdot \mathbf{H} + \gamma^2 \cdot \|\mathbf{H}^{-1/2}\mathbf{\Sigma}\mathbf{H}^{-1/2}\|_2 \cdot \mathbf{H} \\
&= (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2 \cdot \frac{\gamma R^2 \sigma^2}{1-\gamma R^2} \cdot \mathbf{H} + \gamma^2\sigma^2 \cdot \mathbf{H} \\
&= (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \frac{\gamma^2 \sigma^2}{1-\gamma R^2} \cdot \mathbf{H} \\
&\preceq \frac{\gamma^2\sigma^2}{1-\gamma R^2} \cdot \sum_{k=0}^{t-1} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{H}. \qquad \text{(solving the recursion)} \\
&= \frac{\gamma^2\sigma^2}{1-\gamma R^2} \cdot \sum_{k=0}^{t-1} (\mathbf{I} - \gamma\mathbf{H})^k \mathbf{H} (\mathbf{I} - \gamma\mathbf{H})^k \qquad \text{(by the property of } \mathcal{I} - \gamma\widetilde{\mathcal{T}} \text{ in (2.4.1))} \\
&\preceq \frac{\gamma^2\sigma^2}{1-\gamma R^2} \cdot \sum_{k=0}^{t-1} (\mathbf{I} - \gamma\mathbf{H})^k \mathbf{H} \\
&= \frac{\gamma\sigma^2}{1-\gamma R^2} \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^t\right),
\end{aligned}
$$

where in the last inequality we use $\gamma \leq 1/R^2 \leq 1/\operatorname{tr}(\mathbf{H}) \leq 1/\lambda_1$. This completes the proof. □

We are ready to provide the variance error upper bound.

Lemma 2.6.7. Under Assumptions 2.2.1, 2.2.3 and 2.6.4, if the stepsize satisfies $\gamma < 1/R^2$, then it holds that

$$
\text{variance} \leq \frac{\sigma^2}{1-\gamma R^2} \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2 \right),
$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$.

Proof. By Lemma 2.6.2, we can bound the variance error as follows

$$\begin{aligned}
\text{variance} &\leq \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_t\right\rangle \\
&= \frac{1}{\gamma N^2}\sum_{t=0}^{N-1}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{C}_t\right\rangle \\
&\leq \frac{\sigma^2}{N^2(1-\gamma R^2)}\sum_{t=0}^{N-1}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{t}\right)\right\rangle \\
&= \frac{\sigma^2}{N^2(1-\gamma R^2)}\sum_{i}\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N-t}\right)\left(1-(1-\gamma\lambda_i)^{t}\right) \\
&\leq \frac{\sigma^2}{N^2(1-\gamma R^2)}\sum_{i}\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N}\right)\left(1-(1-\gamma\lambda_i)^{N}\right) \\
&= \frac{\sigma^2}{N(1-\gamma R^2)}\left(1-(1-\gamma\lambda_i)^{N}\right)^2,
\end{aligned}$$

where the second inequality is due to Lemma 2.6.6, $\{\lambda_i\}_{i\geq 1}$ are the eigenvalues of $\mathbf{H}$ and are sorted in decreasing order. Since $\gamma \leq 1/\lambda_1$, we have for all $i \geq 1$ that

$$1-(1-\gamma\lambda_i)^N \leq \min\left\{1, \gamma N\lambda_i\right\}. \tag{2.6.12}$$

Set $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma N}\}$, then

$$\begin{aligned}
\text{variance} &\leq \frac{\sigma^2}{N(1-\gamma R^2)}\sum_{i}\min\left\{1, \gamma^2 N^2\lambda_i^2\right\} \\
&\leq \frac{\sigma^2}{N(1-\gamma R^2)}\left(k^* + N^2\gamma^2\cdot\sum_{i>k^*}\lambda_i^2\right) \\
&= \frac{\sigma^2}{1-\gamma R^2}\left(\frac{k^*}{N} + \gamma^2 N\cdot\sum_{i>k^*}\lambda_i^2\right).
\end{aligned}$$

□

2.6.4 Bounding the Bias Error

In this part we will focus on bounding the bias error. Recall the bias error bound in Lemma 2.6.3:

$$\begin{aligned}\text{bias} &\le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H},\mathbf{B}_t\right\rangle \\ &= \frac{1}{\gamma N^2}\sum_{t=0}^{N-1}\left\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t},\mathbf{B}_t\right\rangle \\ &\le \frac{1}{\gamma N^2}\Big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N},\sum_{t=0}^{N-1}\mathbf{B}_t\Big\rangle. \end{aligned} \tag{2.6.13}$$

Let $\mathbf{S}_n = \sum_{t=0}^{n-1}\mathbf{B}_t$, then we only need to bound $\mathbf{S}_N$.

Lemma 2.6.8. Let $\mathbf{S}_t = \sum_{k=0}^{t-1}\mathbf{B}_k$, if $\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{A}))$, we have

$$\mathbf{S}_t = (\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{S}_{t-1}+\mathbf{B}_0.$$

Moreover, it holds that

$$\mathbf{B}_0 = \mathbf{S}_0 \preceq \mathbf{S}_1 \preceq \cdots \preceq \mathbf{S}_\infty.$$

Proof. By (2.4.4), we have

$$\mathbf{B}_t = (\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{B}_{t-1} = (\mathcal{I}-\gamma\mathcal{T})^t\circ\mathbf{B}_0, \tag{2.6.14}$$

where we used recursion. Then we have

$$\mathbf{S}_t = \sum_{k=0}^{t-1}(\mathcal{I}-\gamma\mathcal{T})^k\circ\mathbf{B}_0 = (\mathcal{I}-\gamma\mathcal{T})\circ\left(\sum_{k=0}^{t-1}(\mathcal{I}-\gamma\mathcal{T})^k\circ\mathbf{B}_0\right)+\mathbf{B}_0 = (\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{S}_{t-1}+\mathbf{B}_0.$$

Moreover, since $\mathbf{B}_t$ is PSD for all $t\ge 0$, it is clear that $\mathbf{S}_t = \mathbf{S}_{t-1}+\mathbf{B}_t \succeq \mathbf{S}_{t-1}$. Besides, by Lemma 2.4.1, we know that

$$\mathbf{S}_\infty := \sum_{k=0}^{\infty}(\mathcal{I}-\gamma\mathcal{T})^t\circ\mathbf{B}_0 = \gamma^{-1}\mathcal{T}^{-1}\circ\mathbf{B}_0$$

exists. Thus it can be readily shown that

$$\mathbf{B}_0 = \mathbf{S}_1 \preceq \cdots \preceq \mathbf{S}_t \preceq \mathbf{S}_{t+1} \preceq \cdots \preceq \mathbf{S}_\infty,$$

which completes the proof. □

Lemma 2.6.9. Under Assumptions 2.2.2, for any symmetric matrix $\mathbf{A}$, if $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$, it holds that

$$\mathcal{M} \circ \mathcal{T}^{-1} \circ \mathbf{A} \preceq \frac{\alpha \operatorname{tr}(\mathbf{A})}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \cdot \mathbf{H}.$$

Proof. We first tackle $\mathcal{T}^{-1} \circ \mathbf{A}$. In particular, by Lemma 2.4.1 we have the operator $\mathcal{T}^{-1}$ exists and thus $\mathcal{T}^{-1} \circ \mathbf{A}$ also exists, which can be obtained by solving for the PSD matrix $\mathbf{D}$ satisfying the following equation,

$$\mathcal{T} \circ \mathbf{D} = \mathbf{A}.$$

Using the definition of $\widetilde{\mathcal{T}}$, we have:

$$\widetilde{\mathcal{T}} \circ \mathbf{D} = \gamma\mathcal{M} \circ \mathbf{D} + \mathbf{A} - \gamma\mathbf{HDH}, \tag{2.6.15}$$

where $\mathcal{M} \circ \mathbf{D} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{D} \mathbf{x}\mathbf{x}^\top]$. Further by Lemma 2.4.1 we know that $\widetilde{\mathcal{T}}^{-1}$ and $\mathcal{M}$ are both PSD mapping. This implies that for any PSD matrices $\mathbf{U}$ and $\mathbf{U}'$ satisfying $\mathbf{0} \preceq \mathbf{U} \preceq \mathbf{U}'$, it holds that

$$\mathbf{0} \preceq \mathcal{M} \circ \mathbf{U} \preceq \mathcal{M} \circ \mathbf{U}', \qquad \mathbf{0} \preceq \widetilde{\mathcal{T}}^{-1} \circ \mathbf{U} \preceq \widetilde{\mathcal{T}}^{-1} \circ \mathbf{U}'.$$

Combining the above two results we also have

$$\mathbf{0} \preceq \mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{U} \preceq \mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{U}'. \tag{2.6.16}$$

Therefore, applying the operator $\mathcal{T}^{-1}$ to both sides of (2.6.15) yields

$$\begin{aligned} \mathbf{D} &= \gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{D} + \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A} - \gamma\widetilde{\mathcal{T}}^{-1} \circ (\mathbf{HDH}) \\ &\preceq \gamma\widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{D} + \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A}. \end{aligned} \tag{2.6.17}$$

Then we can apply the operator $\mathcal{M}$ to both sides of (2.6.17), by the monotonicity property in (2.6.16), we have

$$\begin{aligned}\mathcal{M} \circ \mathbf{D} &\preceq \gamma \mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathcal{M} \circ \mathbf{D} + \mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A} \\ &\preceq \sum_{t=0}^{\infty} (\gamma \mathcal{M} \circ \widetilde{\mathcal{T}}^{-1})^t \circ (\mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A}). \end{aligned} \tag{2.6.18}$$

By Assumption 2.2.2 we have

$$\mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A} \preceq \alpha \operatorname{tr}(\mathbf{H} \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A}) \mathbf{H}. \tag{2.6.19}$$

Additionally, based on the definition of $\widetilde{\mathcal{T}}$, we have

$$\widetilde{\mathcal{T}}^{-1} \circ \mathbf{A} = \gamma \sum_{t=0}^{\infty} (\mathcal{I} - \gamma \widetilde{\mathcal{T}})^t \circ \mathbf{A} = \gamma \sum_{t=0}^{\infty} (\mathbf{I} - \gamma \mathbf{H})^t \mathbf{A} (\mathbf{I} - \gamma \mathbf{H})^t.$$

Therefore, it follows that

$$\begin{aligned}\operatorname{tr}(\mathbf{H} \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A}) &= \gamma \operatorname{tr} \left(\sum_{t=0}^{\infty} \mathbf{H} (\mathbf{I} - \gamma \mathbf{H})^t \mathbf{A} (\mathbf{I} - \gamma \mathbf{H})^t \right) \\ &= \gamma \operatorname{tr} \left(\sum_{t=0}^{\infty} \mathbf{H} (\mathbf{I} - \gamma \mathbf{H})^{2t} \mathbf{A} \right) \\ &= \operatorname{tr} \left(\mathbf{H} (2\mathbf{H} - \gamma \mathbf{H}^2)^{-1} \mathbf{A} \right) \\ &\leq \operatorname{tr}(\mathbf{A}), \end{aligned} \tag{2.6.20}$$

where the last inequality is because we have $\gamma \leq 1/\lambda_1$ and thus $\mathbf{H}(2\mathbf{H} - \gamma \mathbf{H}^2)^{-1} \preceq \mathbf{I}$. Substituting (2.6.20) into (2.6.19) yields

$$\mathcal{M} \circ \widetilde{\mathcal{T}}^{-1} \circ \mathbf{A} \preceq \alpha \operatorname{tr}(\mathbf{A}) \mathbf{H}.$$

Note that we have $\widetilde{\mathcal{T}}^{-1} \mathbf{H} \preceq \mathbf{I}$ and $\mathcal{M} \circ \mathbf{I} \preceq \alpha \operatorname{tr}(\mathbf{H}) \mathbf{H}$, plugging the above inequality into (2.6.18) gives

$$\mathcal{M} \circ \mathcal{T}^{-1} \circ \mathbf{A} = \mathcal{M} \circ \mathbf{D} \preceq \alpha \operatorname{tr}(\mathbf{A}) \sum_{t=0}^{\infty} (\gamma \alpha \operatorname{tr}(\mathbf{H}))^t \mathbf{H} \preceq \frac{\alpha \operatorname{tr}(\mathbf{A})}{1 - \gamma \alpha \operatorname{tr}(\mathbf{H})} \cdot \mathbf{H}.$$

This completes the proof. □

Lemma 2.6.10. Under Assumptions 2.2.1, and 2.2.2, if the stepsize satisfies $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$, then

$$\mathcal{M} \circ \mathbf{S}_t \preceq \frac{\alpha \cdot \operatorname{tr}\left(\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t\right] \circ \mathbf{B}_0\right)}{\gamma(1 - \gamma\alpha \operatorname{tr}(\mathbf{H}))} \cdot \mathbf{H}.$$

Proof. Note that $\mathbf{S}_t$ takes the following form

$$\mathbf{S}_t := \sum_{k=0}^{t-1} (\mathcal{I} - \gamma\mathcal{T})^k \circ \mathbf{B}_0 = \gamma^{-1}\mathcal{T}^{-1} \circ \left[\mathcal{I} - (\mathcal{I} - \gamma\mathcal{T})^t\right]\mathbf{B}_0.$$

Note that by Lemma 2.4.1, we have $\mathcal{I} - \gamma\widetilde{\mathcal{T}} \leq \mathcal{I} - \gamma\mathcal{T}$ so that $\mathcal{I} - (\mathcal{I} - \gamma\mathcal{T})^t \preceq \mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t$. Therefore, further note that $\mathcal{T}^{-1}$ is a PSD mapping, we have the following bound on $\mathbf{S}_t$,

$$\mathbf{S}_t \preceq \gamma^{-1}\mathcal{T}^{-1} \circ \left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t\right] \circ \mathbf{B}_0.$$

Then note that $\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t\right] \circ \mathbf{B}_0$ is a PSD matrix, applying Lemma 2.6.9, we get

$$\mathcal{M} \circ \mathbf{S}_t \preceq \gamma^{-1}\mathcal{M} \circ \mathcal{T}^{-1} \circ \left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t\right] \circ \mathbf{B}_0 \preceq \frac{\alpha \cdot \operatorname{tr}\left(\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t\right] \circ \mathbf{B}_0\right)}{\gamma(1 - \gamma\alpha \operatorname{tr}(\mathbf{H}))} \cdot \mathbf{H}.$$

This completes the proof.

□

The following lemma shows that using this crude bound on $\mathcal{M} \circ \mathbf{S}_t$ we are able to get a tighter upper bound on $\mathbf{S}_t$.

Lemma 2.6.11. Under Assumptions 2.2.1 and 2.2.2, let $\mathbf{B}_{a,b} = \mathbf{B}_a - (\mathbf{I} - \gamma\mathbf{H})^{b-a}\mathbf{B}_a(\mathbf{I} - \gamma\mathbf{H})^{b-a}$, if the stepsize satisfies $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$, then for any $t \leq N$, it holds that

$$\mathbf{S}_t \preceq \sum_{k=0}^{t-1} (\mathbf{I} - \gamma\mathbf{H})^k \left(\frac{\gamma\alpha \operatorname{tr}(\mathbf{B}_{0,N})}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0\right)(\mathbf{I} - \gamma\mathbf{H})^k.$$

Proof. Recall the recursive form of $\mathbf{S}_t$ given in Lemma 2.6.8, we have

$$\mathbf{S}_t = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{S}_{t-1} + \mathbf{B}_0.$$

Note that this is similar to the recursive form of $\mathbf{C}_t$ provided in (2.4.5) but replacing $\gamma^2\boldsymbol{\Sigma}$ with $\mathbf{B}_0$. Then we can use the similar proof of Lemma 2.6.6 to get the upper bound of $\mathbf{S}_t$. In particular, note that we will run SGD with N steps, then $\mathbf{S}_N$ can be used as a uniform upper bound on $\mathbf{S}_1, \ldots, \mathbf{S}_N$, we can upper bound $\mathbf{S}_t$ by

$$\begin{aligned}
\mathbf{S}_t &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{S}_{t-1} + \gamma^2 \mathcal{M} \circ \mathbf{S}_N + \mathbf{B}_0 \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{S}_{t-1} + \frac{\gamma\alpha \cdot \operatorname{tr}\left(\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^N\right] \circ \mathbf{B}_0\right)}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0 \\
&= \sum_{k=0}^{t-1} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \left(\frac{\gamma\alpha \cdot \operatorname{tr}\left(\left[\mathcal{I} - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^N\right] \circ \mathbf{B}_0\right)}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0 \right) \\
&= \sum_{k=0}^{t-1} (\mathbf{I} - \gamma\mathbf{H})^k \left(\frac{\gamma\alpha \operatorname{tr}\left(\mathbf{B}_0 - (\mathbf{I} - \gamma\mathbf{H})^N \mathbf{B}_0 (\mathbf{I} - \gamma\mathbf{H})^N\right)}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0 \right) (\mathbf{I} - \gamma\mathbf{H})^k.
\end{aligned}$$

where we use Lemma 2.6.10 in the second inequality, the first equality is by recursion, and the last equality is by the definition of $\widetilde{\mathcal{T}}$. □

We now put these lemmas together and provide our upper bound on the bias error:

Lemma 2.6.12. Under Assumptions 2.2.1 and 2.2.2, if the stepsize satisfies $\gamma < 1/(\alpha\operatorname{tr}(\mathbf{H}))$, it holds that

$$\begin{aligned}
\text{bias} \leq \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\
+ \frac{2\alpha\left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right)}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2 \right),
\end{aligned}$$

where $k^* = \max\{k : \lambda_k \geq \gamma^{-1}/N\}$.

Proof. We can plug the upper bound of $\mathbf{S}_t$ derived in Lemma 2.6.11 into (2.6.13) and get

$$\begin{aligned}
\text{bias} &\leq \frac{1}{\gamma N^2} \sum_{k=0}^{N-1} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N, (\mathbf{I} - \gamma\mathbf{H})^k \left(\frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0 \right) (\mathbf{I} - \gamma\mathbf{H})^k \right\rangle \\
&= \frac{1}{\gamma N^2} \sum_{k=0}^{N-1} \left\langle (\mathbf{I} - \gamma\mathbf{H})^{2k} - (\mathbf{I} - \gamma\mathbf{H})^{N+2k}, \frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{0,N})}{1 - \gamma\alpha\operatorname{tr}(\mathbf{H})} \cdot \mathbf{H} + \mathbf{B}_0 \right\rangle.
\end{aligned}$$

Note that

$$(\mathbf{I}-\gamma\mathbf{H})^{2k}-(\mathbf{I}-\gamma\mathbf{H})^{N+2k}=(\mathbf{I}-\gamma\mathbf{H})^{k}\big((\mathbf{I}-\gamma\mathbf{H})^{k}-(\mathbf{I}-\gamma\mathbf{H})^{N+k}\big)$$
$$\preceq(\mathbf{I}-\gamma\mathbf{H})^{k}-(\mathbf{I}-\gamma\mathbf{H})^{N+k}.$$

We obtain

$$\text{bias}\leq\frac{1}{\gamma N^{2}}\sum_{k=0}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H})^{k}-(\mathbf{I}-\gamma\mathbf{H})^{N+k},\frac{\gamma\alpha\,\mathrm{tr}(\mathbf{B}_{0,N})}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\cdot\mathbf{H}+\mathbf{B}_{0}\right\rangle,$$

Therefore, it suffices to upper bound the following two terms:

$$I_{1}=\frac{\alpha\,\mathrm{tr}(\mathbf{B}_{0,N})}{N^{2}(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\sum_{k=0}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k}-(\mathbf{I}-\gamma\mathbf{H})^{N+k},\mathbf{H}\big\rangle$$
$$I_{2}=\frac{1}{\gamma N^{2}}\sum_{k=0}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k}-(\mathbf{I}-\gamma\mathbf{H})^{N+k},\mathbf{B}_{0}\big\rangle.$$

Regarding I_1, since $\mathbf{H}$ and $\mathbf{I}-\gamma\mathbf{H}$ can be diagonalized simultaneously, we have

$$\begin{aligned}I_{1}&=\frac{\alpha\,\mathrm{tr}(\mathbf{B}_{0,N})}{N^{2}(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\sum_{k=0}^{N-1}\sum_{i}\big[(1-\gamma\lambda_{i})^{k}-(1-\gamma\lambda_{i})^{N+k}\big]\lambda_{i}\\&=\frac{\alpha\,\mathrm{tr}(\mathbf{B}_{0,N})}{\gamma N^{2}(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\sum_{i}\big[1-(1-\gamma\lambda_{i})^{N}\big]^{2}\\&\leq\frac{\alpha\,\mathrm{tr}(\mathbf{B}_{0,N})}{\gamma N^{2}(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\sum_{i}\min\big\{1,\gamma^{2}N^{2}\lambda_{i}^{2}\big\}\\&\leq\frac{\alpha\,\mathrm{tr}(\mathbf{B}_{0,N})}{\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot\bigg(\frac{k^{*}}{N^{2}}+\gamma^{2}\sum_{i>k^{*}}\lambda_{i}^{2}\bigg),\end{aligned}\tag{2.6.21}$$

where k^* is the index of the smallest eigenvalue of $\mathbf{H}$ satisfying $\lambda_{k^*}\geq\gamma^{-1}/N$. Moreover, recall that $\widetilde{\mathbf{B}}=\mathbf{B}_0-(\mathbf{I}-\gamma\mathbf{H})^N\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^N\big)$ and $\mathbf{B}_0=(\mathbf{w}_0-\mathbf{w}^*)\otimes(\mathbf{w}_0-\mathbf{w}^*)$, we have

$$\mathrm{tr}(\mathbf{B}_{0,N})=\mathrm{tr}\left(\mathbf{B}_{0}-(\mathbf{I}-\gamma\mathbf{H})^{N}\mathbf{B}_{0}(\mathbf{I}-\gamma\mathbf{H})^{N}\big)\right)=\sum_{i}\big(1-(1-\gamma\lambda_{i})^{2N}\big)\cdot\big(\langle\mathbf{w}_{0}-\mathbf{w}^{*},\mathbf{v}_{i}\rangle\big)^{2}.$$

Note that

$$\big(1-(1-\gamma\lambda_{i})^{2N}\big)\leq\min\{2,2N\gamma\lambda_{i}\},$$

thus it follows that,

$$\operatorname{tr}(\mathbf{B}_{0,N}) \le 2\sum_i \min\{1, N\gamma\lambda_i\}\big(\langle \mathbf{w}_0 - \mathbf{w}^*, \mathbf{v}_i\rangle\big)^2 \le 2\big(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big). \tag{2.6.22}$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{N\gamma}\}$. Then plug this bound into (2.6.21), we have

$$I_1 \le \frac{2\alpha\big(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{N\gamma(1 - \gamma\alpha\operatorname{tr}(\mathbf{H}))} \cdot \bigg(\frac{k^*}{N} + N\gamma^2\sum_{i>k^*}\lambda_i^2\bigg), \tag{2.6.23}$$

In the sequel we will upper bound I_2. Let $\mathbf{H} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ be the orthogonal decomposition of $\mathbf{H}$, where $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots)$ and $\mathbf{\Lambda}$ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \dots$. Then we have

$$I_2 = \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\big\langle (\mathbf{I} - \gamma\mathbf{\Lambda})^k - (\mathbf{I} - \gamma\mathbf{\Lambda})^{N+k}, \mathbf{V}^\top\mathbf{B}_0\mathbf{V}\big\rangle.$$

Note that $(\mathbf{I} - \gamma\mathbf{\Lambda})^k - (\mathbf{I} - \gamma\mathbf{\Lambda})^{N+k}$ is a diagonal matrix, thus the above inner product only operates on the diagonal entries of $\mathbf{V}^\top\mathbf{B}_0\mathbf{V}$. Note that $\mathbf{B}_0 = \boldsymbol{\beta}_0\boldsymbol{\beta}_0^\top$, it can be shown that the diagonal entries of $\mathbf{V}^\top\mathbf{B}_0\mathbf{V}$ are $\omega_1^2, \omega_2^2, \dots$, where $\omega_i = \mathbf{v}_i^\top\boldsymbol{\beta}_0 = \mathbf{v}_i^\top(\mathbf{w}_0 - \mathbf{w}^*)$.

$$\begin{aligned}
I_2 &= \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\big\langle (\mathbf{I} - \gamma\mathbf{H})^k - (\mathbf{I} - \gamma\mathbf{H})^{N+k}, \mathbf{B}_0\big\rangle \\
&= \frac{1}{\gamma N^2}\sum_{k=0}^{N-1}\sum_i \big[(1 - \gamma\lambda_i)^k - (1 - \gamma\lambda_i)^{N+k}\big]\omega_i^2 \\
&= \frac{1}{\gamma^2 N^2}\sum_i \frac{\omega_i^2}{\lambda_i}\big[1 - (1 - \gamma\lambda_i)^N\big]^2 \\
&\le \frac{1}{\gamma^2 N^2}\sum_i \frac{\omega_i^2}{\lambda_i}\min\big\{1, \gamma^2 N^2\lambda_i^2\big\} \\
&\le \frac{1}{\gamma^2 N^2}\cdot\sum_{i\le k^*}\frac{\omega_i^2}{\lambda_i} + \sum_{i>k^*}\lambda_i\omega_i^2 \\
&= \frac{1}{\gamma^2 N^2}\cdot\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}},
\end{aligned}$$

where the first inequality is by (2.6.12) and $k^* = \max\{k : \lambda_k \ge \gamma^{-1}/N\}$. Combining the upper bounds on I_1 and I_2 directly completes the proof. □

2.6.5 Proof of Theorem 2.2.4

Proof. By Lemma 2.6.2, it suffices to substitute into the upper bounds on the bias and variance errors. In particular, by Young's inequality we have

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \leq \left(\sqrt{\text{bias}} + \sqrt{\text{variance}}\right)^2 \leq 2 \cdot \text{bias} + 2 \cdot \text{variance}.$$

Then we can directly substitute the bounds of variance and bias we proved in Lemmas 2.6.7 and 2.6.12. In particular, by Assumptions 2.2.2 we can directly get $R^2 = \alpha \operatorname{tr}(\mathbf{H})$. Therefore, it holds that

$$\begin{aligned}
&\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \\
&\leq 2\Bigg[\frac{\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))} \cdot \left(\frac{k^*}{N^2} + \gamma^2 \sum_{i>k^*} \lambda_i^2\right) + \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\
&\quad + \frac{\sigma_z^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right)\Bigg] \\
&= 2 \cdot \text{EffectiveBias} + 2 \cdot \text{EffectiveVar},
\end{aligned}$$

where

$$\begin{aligned}
\text{EffectiveBias} &= \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\
\text{EffectiveVar} &= \left(\frac{\sigma_z^2}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})} + \frac{\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{N\gamma(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))}\right)\left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right).
\end{aligned}$$

□

2.6.6 Proof of Corollary 2.2.8

Proof. We will show that the corollary can be directly implied by Theorem 2.2.4. In terms of the effective bias term, it is clear that

$$\begin{aligned}
\text{EffectiveBias} &\leq \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\
&= \frac{1}{\gamma^2 N^2} \cdot \lambda_{k^*}^{-1} \sum_{i \leq k^*} \left(\mathbf{v}_i^\top \mathbf{w}_0 - \mathbf{v}_i^\top \mathbf{w}^*\right)^2 + \lambda_{k^*+1} \sum_{i>k^*} \left(\mathbf{v}_i^\top \mathbf{w}_0 - \mathbf{v}_i^\top \mathbf{w}^*\right)^2.
\end{aligned}$$

where $\mathbf{v}_i$ is the eigenvector of $\mathbf{H}$ corresponding to the eigenvalue λ_i. Based on our definition of k^*, we have $\lambda_{k^*}^{-1} \leq N\gamma$ and $\lambda_{k^*+1} \leq 1/(N\gamma)$. Therefore, it follows that

$$\text{EffectiveBias} \leq \frac{1}{\gamma N} \cdot \sum_i \left(\mathbf{v}_i^\top \mathbf{w}_0 - \mathbf{v}_i^\top \mathbf{w}^*\right)^2 = \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N}. \tag{2.6.24}$$

Then regarding the effective variance, given the choice of stepsize that $\gamma = 1/(2\alpha \operatorname{tr}(\mathbf{H}))$, we have

$$\begin{aligned}\text{EffectiveVar} &\leq 2\left(\sigma^2 + \frac{\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{N\gamma}\right)\left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right) \\ &= 2\sigma^2 \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right) + \frac{2\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N} \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right).\end{aligned}$$

Based on the definition of k^*, we have $\lambda_i \leq 1/(N\gamma)$ for $i > k^*$, thus

$$\gamma^2 N \sum_{i>k^*} \lambda_i^2 \leq \gamma \sum_{i>k^*} \lambda_i.$$

Besides, we also have $k^*/N \leq \gamma \sum_{i=1}^{k^*} \lambda_i$. Therefore, we have

$$\text{EffectiveVar} \leq 2\sigma^2 \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right) + \frac{2\gamma\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N} \cdot \sum_i \lambda_i.$$

According to our choice of stepsize that $\gamma = 1/(2\alpha \operatorname{tr}(\mathbf{H}))$, we can get

$$\frac{2\gamma\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N} \cdot \sum_i \lambda_i = \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N}.$$

This further implies that

$$\text{EffectiveVar} \leq 2\sigma^2 \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right) + \frac{\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N}. \tag{2.6.25}$$

Combining (2.6.24) and (2.6.25), we have

$$\begin{aligned}\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) &\leq 2 \cdot \text{EffectiveBias} + 2 \cdot \text{EffectiveVar} \\ &\leq \frac{4\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\gamma N} + 4\sigma^2 \cdot \left(\frac{k^*}{N} + \gamma^2 N \cdot \sum_{i>k^*} \lambda_i^2\right).\end{aligned}$$

Further using the assumption that $\gamma = 1/(2\alpha \operatorname{tr}(\mathbf{H}))$ completes the proof. □

2.6.7 Proof of Corollary 2.2.9

Proof. For the bias error term, recall the definition of k^*, we have

$$\begin{aligned}\text{EffectiveBias} &\leq \mathcal{O}\left(\frac{1}{N^2}\cdot||\mathbf{w}_0-\mathbf{w}^*||^2_{\mathbf{H}^{-1}_{0:k^*}}+||\mathbf{w}_0-\mathbf{w}^*||^2_{\mathbf{H}_{k^*:\infty}}\right)\\ &\leq \mathcal{O}\left(\frac{1}{N^2}\cdot\frac{1}{\lambda_{k^*}}\cdot||\mathbf{w}_0-\mathbf{w}^*||_2^2+\lambda_{k^*}\cdot||\mathbf{w}_0-\mathbf{w}^*||_2^2\right)\\ &\leq \mathcal{O}\left(\frac{1}{N}\right).\end{aligned}$$

For the variance error term, it can be verified that all these examples satisfies $\sum_i \lambda_i < \infty$, thus we have

$$\text{EffectiveVar} = \mathcal{O}\left(\frac{k^*}{N}+N\sum_{i>k^*}\lambda_i^2\right).$$

1. By the definition of k^* we have $k^* = s = N^r$, therefore

$$\text{EffectiveVar} = \mathcal{O}\left(N^{-1}\cdot N^r + N\cdot N^{-q}\right) = \mathcal{O}\left(N^{r-1}+N^{1-q}\right).$$

2. By the definition of k^* we have $k^* = \Theta\big(N^{1/(1+r)}\big)$, therefore

$$\text{EffectiveVar} = \mathcal{O}\left(N^{-1}\cdot N^{1/(1+r)} + N\cdot\left(N^{1/(1+r)}\right)^{-1-2r}\right) = \mathcal{O}\left(N^{-r/(1+r)}\right).$$

3. By the definition of k^* it can be shown that $k^* = \Omega\big(N/\log^\beta(N)\big)$ since otherwise

$$\lambda_{k^*+1} = \omega\left(\frac{\log^\beta(N)}{N}\cdot\frac{1}{\big[\log(N)-\beta\log(\log(N))\big]^\beta}\right) = \omega(1/N),$$

which contradicts to the fact that $\lambda_{k^*+1} = \mathcal{O}\left(1/N\right)$. Besides, we have

$$\sum_{i\geq k^*}\lambda_i^2 = \mathcal{O}\left(\int_{k^*}^\infty \frac{1}{x^2\log^{2\beta}(x+1)}\mathrm{d}x\right).$$

Then note that

$$\frac{1}{x^2\log^{2\beta}(x+1)} \leq \frac{\log^{2\beta}(x+1)+2\beta x\log^{2\beta-1}(x)/(x+1)}{x^2\log^{4\beta}(x)}.$$

This implies that

$$\begin{aligned}\int_{k^*}^{\infty} \frac{1}{x^2 \log^{2\beta}(x)} \mathrm{d}x &\leq \int_{k^*}^{\infty} \frac{\log^{2\beta}(x+1) + 2\beta x \log^{2\beta-1}(x)/(x+1)}{x^2 \log^{4\beta}(x)} \mathrm{d}x \\ &= \frac{1}{k^* \log^{2\beta}(k^*+1)} \\ &= \mathcal{O}\left(N^{-1} \log^{-\beta}(k^*)\right),\end{aligned}$$

where the last equality is due to the fact that $1/(k^* \log^{\beta}(k^*+1)) = \Theta(1/N)$. As a result, we can get

$$\text{EffectiveVar} = \mathcal{O}\left(k^* \cdot N^{-1} + N \sum_{i \geq k^*} \lambda_i^2\right) = \mathcal{O}\left(\log^{-\beta}(k^*)\right) = \mathcal{O}\left(\log^{-\beta}(N)\right),$$

where the second equality is due to the fact that $k*/N = \mathcal{O}\left(\log^{-\beta}(k^*)\right)$ and the last equality is due to $k^* = \Omega\big(N/\log^{\beta}(N)\big)$.

4. By definition of k^* we have $k^* = \Theta\left(\log N\right)$, therefore

$$\text{EffectiveVar} = \mathcal{O}\left(N^{-1} \cdot \log N + N \cdot e^{-2\log N}\right) = \mathcal{O}\left(N^{-1} \log N\right).$$

Summing up the bias error and variance error concludes the proof. □

2.7 Proofs of the Lower Bounds

2.7.1 Lower Bound for Bias-Variance Decomposition

We first introduce the following lemma to lower bound the excess risk when the noise is well-specified as in (2.2.1).

Lemma 2.7.1. Suppose the model noise ξ_t is well-specified, i.e., ξ_t and $\mathbf{x}_t$ are independent and $\mathbb{E}[\xi_t] = 0$. Then

$$\begin{aligned}\mathbb{E}[L(\overline{\mathbf{w}}_N) - L(\mathbf{w}^*)] \geq \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{B}_t \Big\rangle \\ + \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \Big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{C}_t \Big\rangle.\end{aligned}$$

Proof. Let $\mathbf{P}_t = \mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top$, then the definitions of $\boldsymbol{\beta}_t^{\text{bias}}$ in (2.4.3) and $\boldsymbol{\beta}_t^{\text{variance}}$ (2.4.2) imply

$$\boldsymbol{\beta}_t^{\text{bias}} = \prod_{k=1}^{t} \mathbf{P}_k \boldsymbol{\beta}_0, \qquad \boldsymbol{\beta}_t^{\text{variance}} = \gamma \sum_{i=1}^{t} \prod_{j=i+1}^{t} \xi_i \mathbf{P}_j \mathbf{x}_i.$$

Note that in the well specified case, the noise $\xi_t := y_t - \langle \mathbf{w}^*, \mathbf{x}_t \rangle$ is independent of the data $\mathbf{x}_t$, and is of zero mean, hence

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{variance}}] &= \gamma \mathbb{E}\Big[\prod_{k=1}^{t} \mathbf{P}_k \boldsymbol{\beta}_0 \otimes \sum_{i=1}^{t} \prod_{j=i+1}^{t} \xi_i \mathbf{P}_j \mathbf{x}_i \Big] \\
&= \gamma \sum_{i=1}^{t} \mathbb{E}\big[\prod_{k=1}^{t} \mathbf{P}_k \boldsymbol{\beta}_0 \otimes \prod_{j=i+1}^{t} \mathbf{P}_j \mathbf{x}_i \big] \cdot \mathbb{E}[\xi_i] = \mathbf{0}.
\end{aligned}$$

This implies that

$$\mathbb{E}[\bar{\boldsymbol{\beta}}_t \otimes \bar{\boldsymbol{\beta}}_t] = \mathbb{E}[\bar{\boldsymbol{\beta}}_t^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_t^{\text{bias}}] + \mathbb{E}[\bar{\boldsymbol{\beta}}_t^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_t^{\text{variance}}],$$

and furthermore,

$$\begin{aligned}
\mathbb{E}[L(\overline{\mathbf{w}}_N) - L(\mathbf{w}^*)] &= \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_t \otimes \bar{\boldsymbol{\beta}}_t] \rangle \\
&= \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_t^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_t^{\text{bias}}] \rangle + \frac{1}{2} \langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_t^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_t^{\text{variance}}] \rangle. \qquad (2.7.1)
\end{aligned}$$

Next, we lower bound each term on the R.H.S. of (2.7.1) separately. By (2.6.5), we have

$$\mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}] = \frac{1}{N^2} \cdot \Bigg(\sum_{0 \le k < t \le N-1} \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_k^{\text{bias}}] + \sum_{0 \le t \le k \le N-1} \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_k^{\text{bias}}] \Bigg).$$

Additionally, by (2.6.3) we can get

$$\begin{aligned}
\Big\langle \mathbf{H}, \sum_{0 \le k < t \le N-1} \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_k^{\text{bias}}] \Big\rangle &= \Big\langle \mathbf{H}, \sum_{k=0}^{N-1} \sum_{t=k+1}^{N-1} (\mathbf{I} - \gamma \mathbf{H})^{t-k} \mathbb{E}[\boldsymbol{\beta}_k^{\text{bias}} \otimes \boldsymbol{\beta}_k^{\text{bias}}] \Big\rangle \\
&= \sum_{k=0}^{N-1} \sum_{t=k+1}^{N-1} \big\langle (\mathbf{I} - \gamma \mathbf{H})^{t-k} \mathbf{H}, \mathbb{E}[\boldsymbol{\beta}_k^{\text{bias}} \otimes \boldsymbol{\beta}_k^{\text{bias}}] \big\rangle \ge 0,
\end{aligned}$$

where the inequality is due to the fact that $(\mathbf{I} - \gamma \mathbf{H})^{t-k} \mathbf{H}$ and $\mathbb{E}[\boldsymbol{\beta}_k^{\text{bias}} \otimes \boldsymbol{\beta}_k^{\text{bias}}]$ are both PSD.

Therefore, it follows that

$$\begin{aligned}
\text{bias} &:= \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{bias}}]\rangle \\
&\geq \frac{1}{2N^2} \cdot \left\langle \mathbf{H}, \sum_{0\leq t\leq k\leq N-1} \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_k^{\text{bias}}] \right\rangle \\
&= \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1}\sum_{k=t}^{N-1} \left\langle \mathbf{H}, \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{bias}}] \cdot (\mathbf{I}-\gamma\mathbf{H})^{k-t} \right\rangle \\
&= \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1}\sum_{k=t}^{N-1} \left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{bias}}] \right\rangle, \qquad (2.7.2)
\end{aligned}$$

where the last equality holds since $\mathbf{H}$ and $(\mathbf{I}-\gamma\mathbf{H})^{k-t}$ commute. Repeating the computation for the variance terms, we can similarly obtain

$$\begin{aligned}
\text{variance} &:= \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\beta}}_N^{\text{variance}}]\rangle \\
&\geq \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1}\sum_{k=t}^{N-1} \left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbb{E}[\boldsymbol{\beta}_t^{\text{variance}} \otimes \boldsymbol{\beta}_t^{\text{variance}}] \right\rangle. \qquad (2.7.3)
\end{aligned}$$

Plugging (2.7.2) and (2.7.3) into (2.7.1) gives

$$\begin{aligned}
\mathbb{E}[L(\overline{\mathbf{w}}_N) - L(\mathbf{w}^*)] = \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_t \otimes \bar{\boldsymbol{\beta}}_t]\rangle &= \text{bias} + \text{variance} \\
&\geq \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1}\sum_{k=t}^{N-1} \left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{bias}}] \right\rangle \\
&\quad + \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1}\sum_{k=t}^{N-1} \left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbb{E}[\boldsymbol{\beta}_t^{\text{variance}} \otimes \boldsymbol{\beta}_t^{\text{variance}}] \right\rangle.
\end{aligned}$$

□

2.7.2 Lower Bounding the Variance Error

Lemma 2.7.2. Suppose Assumptions 2.2.1 hold. Suppose the noise is well-specified as in (2.2.1). If the stepsize satisfies $\gamma < 1/\lambda_1$, it holds that

$$\mathbf{C}_t \succeq \frac{\gamma\sigma_{\text{noise}}^2}{2}\left(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^{2t}\right).$$

Proof. Recall that $\mathcal{M} - \widetilde{\mathcal{M}}$ is a PSD mapping by Lemma 2.4.1 and $\mathbf{C}_{t-1}$ is PSD, then from (2.4.5) we have

$$\begin{aligned}
\mathbf{C}_t &= (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2\mathbf{\Sigma} \\
&= (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + (\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{C}_{t-1} + \gamma^2\mathbf{\Sigma} \\
&\succeq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{C}_{t-1} + \gamma^2\sigma_{\text{noise}}^2\mathbf{H} \qquad \text{(since in the well-specified case } \mathbf{\Sigma} = \sigma_{\text{noise}}^2\mathbf{H}) \\
&\succeq \gamma^2\sigma_{\text{noise}}^2 \cdot \sum_{k=0}^{t-1} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{H} \qquad \text{(solving the recursion)} \\
&= \gamma^2\sigma_{\text{noise}}^2 \cdot \sum_{k=0}^{t-1} (\mathbf{I} - \gamma\mathbf{H})^k\mathbf{H}(\mathbf{I} - \gamma\mathbf{H})^k \qquad \text{(by the property of } \mathcal{I} - \gamma\widetilde{\mathcal{T}} \text{ in (2.4.1))} \\
&= \gamma^2\sigma_{\text{noise}}^2 \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{2t}\right) \cdot \left(2\gamma\mathbf{I} - \gamma^2\mathbf{H}\right)^{-1} \\
&\succeq \frac{\gamma\sigma_{\text{noise}}^2}{2} \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{2t}\right),
\end{aligned}$$

where in the last inequality we use $2\gamma\mathbf{I} - \gamma^2\mathbf{H} \preceq 2\gamma\mathbf{I}$. This completes the proof. □

Lemma 2.7.3. Suppose Assumptions 2.2.1 hold. Suppose the noise is well-specified as in (2.2.1) and $N \geq 500$. Denote

$$\text{variance} = \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_t \right\rangle.$$

If the stepsize satisfies $\gamma < 1/\lambda_1$, then

$$\text{variance} \geq \frac{\sigma_{\text{noise}}^2}{50} \left(\frac{k^*}{N} + N\gamma^2 \cdot \sum_{i>k^*} \lambda_i^2 \right),$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$.

Proof. We can lower bound the variance error as follows

$$\begin{aligned}
\text{variance} &= \frac{1}{2N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_t\right\rangle \\
&= \frac{1}{2\gamma N^2}\sum_{t=0}^{N-1}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{C}_t\right\rangle \\
&\geq \frac{\sigma_{\text{noise}}^2}{4N^2}\sum_{t=0}^{N-1}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{2t}\right\rangle \qquad \text{(use Lemma 2.7.2)} \\
&= \frac{\sigma_{\text{noise}}^2}{4N^2}\sum_{i}\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N-t}\right)\left(1-(1-\gamma\lambda_i)^{2t}\right) \\
&\geq \frac{\sigma_{\text{noise}}^2}{4N^2}\sum_{i}\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N-t-1}\right)\left(1-(1-\gamma\lambda_i)^{t}\right),
\end{aligned}$$

where $\{\lambda_i\}_{i\geq 1}$ are the eigenvalues of $\mathbf{H}$ and are sorted in decreasing order. Define

$$f(x) := \sum_{t=0}^{N-1}\left(1-(1-x)^{N-t-1}\right)\left(1-(1-x)^{t}\right), \qquad 0<x<1,$$

then

$$\text{variance} \geq \frac{\sigma_{\text{noise}}^2}{4N^2}\sum_{i\geq 1} f(\gamma\lambda_i).$$

Clearly $f(x)$ is increasing for $0<x<1$. Moreover:

$$\begin{aligned}
f(x) &= \sum_{t=0}^{N-1}\left(1-(1-x)^{N-1-t}-(1-x)^t+(1-x)^{N-1}\right) \\
&= N-2\frac{1-(1-x)^N}{x}+N(1-x)^{N-1}.
\end{aligned}$$

Next we lower bound $f(x)$ within the range $\frac{1}{N}<x<1$ and $0<x<\frac{1}{N}$, respectively.

First consider $\frac{1}{N}\leq x<1$. Notice that $f(x)$ is increasing and $\left(1-\frac{1}{N}\right)^N \geq \left(1-\frac{1}{500}\right)^{500} > 1.1/3$ if $N\geq 500$, thus for $\frac{1}{N}\leq x<1$, we have

$$f(x)\geq N-2N+3N\cdot(1-1/N)^N \geq 0.1N.$$

On the other hand, note that we have the fourth-order derivative of $f(x)$ is positive when $x\in(0,1/N)$, thus for $0\leq x\leq 1/N$, we can perform third-order Taylor expansion on $f(x)$

at $x = 0$, which gives

$$\begin{aligned} f(x) &\geq \frac{N(N-1)(N-2)x^2}{6} - \frac{N(N-1)(N-2)(N-3)x^3}{12} \\ &\geq \frac{N(N-1)(N-2)x^2}{12} \qquad (\text{since } x \leq 1/N) \\ &\geq \frac{2N^3x^2}{25}. \qquad (\text{since } N \geq 500) \end{aligned}$$

In sum,

$$f(x) \geq \begin{cases} \frac{N}{10}, & \frac{1}{N} \leq x < 1, \\ \frac{2N^3}{25}x^2, & 0 < x < \frac{1}{N}. \end{cases}$$

Set $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$, then

$$\begin{aligned} \text{variance} &\geq \frac{\sigma_{\text{noise}}^2}{4N^2} \sum_i f(\gamma\lambda_i) \\ &\geq \frac{\sigma_{\text{noise}}^2}{4N^2} \left(\frac{Nk^*}{10} + \frac{2N^3}{25}\gamma^2 \cdot \sum_{i>k^*} \lambda_i^2 \right) \\ &\geq \frac{\sigma_{\text{noise}}^2}{50} \left(\frac{k^*}{N} + N\gamma^2 \cdot \sum_{i>k^*} \lambda_i^2 \right). \end{aligned}$$

This completes the proof. □

2.7.3 Lower Bounding the Bias Error

Recall that we have the following lower bound on the bias error

$$\text{bias} \geq \frac{1}{2N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_t \right\rangle,$$

from which we notice that

$$\begin{aligned} \text{bias} &\geq \frac{1}{2N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_t \right\rangle = \frac{1}{2\gamma N^2} \sum_{t=0}^{N-1} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N-t}, \mathbf{B}_t \right\rangle \\ &\geq \frac{1}{2\gamma N^2} \sum_{t=0}^{N/2} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N-t}, \mathbf{B}_t \right\rangle \\ &\geq \frac{1}{2\gamma N^2} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/2}, \sum_{t=0}^{N/2} \mathbf{B}_t \right\rangle. \end{aligned} \tag{2.7.4}$$

Let $\mathbf{S}_n := \sum_{t=0}^{n-1} \mathbf{B}_t$. Then the reminding challenge is to lower bound $\mathbf{S}_{N/2+1} = \sum_{t=0}^{N/2} \mathbf{B}_t$. Similarly to the idea of proving the upper bound, we first establish a crude lower bound on $\mathbf{S}_n$ then improve it to a fine lower bound.

Lemma 2.7.4. Suppose Assumptions 2.2.1 and 2.2.5 hold. If the stepsize satisfies $\gamma < 1/\lambda_1$, then for any $n \geq 2$, it holds that

$$\mathbf{S}_n \succeq \frac{\beta}{4} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n/2}\right)\mathbf{B}_0\right) \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n/2}\right) + \sum_{t=0}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^t \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \gamma\mathbf{H})^t.$$

Proof. We first build a crude bound for $\mathbf{S}_n$. Recall that $\widetilde{\mathcal{T}} - \mathcal{T}$ is a PSD mapping by Lemma 2.4.1, then

$$\mathbf{S}_n = \sum_{t=0}^{n-1} \mathbf{B}_t = \sum_{t=0}^{n-1} (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{B}_0 \succeq \sum_{t=0}^{n-1} (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^t \circ \mathbf{B}_0 = \sum_{t=0}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^t \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \gamma\mathbf{H})^t.$$

Now we apply Assumption 2.2.5 with the above crude bound to obtain that

$$\begin{aligned}
(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{S}_n &\succeq \beta \operatorname{tr}(\mathbf{H}\mathbf{S}_n)\,\mathbf{H} \\
&\succeq \beta \operatorname{tr}\left(\sum_{t=0}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^{2t}\mathbf{H} \cdot \mathbf{B}_0\right)\mathbf{H} \\
&\succeq \beta \operatorname{tr}\left(\sum_{t=0}^{n-1} (\mathbf{I} - 2\gamma\mathbf{H})^{t}\mathbf{H} \cdot \mathbf{B}_0\right)\mathbf{H} \\
&= \frac{\beta}{2\gamma} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - 2\gamma\mathbf{H})^n\right)\mathbf{B}_0\right)\mathbf{H} \\
&\succeq \frac{\beta}{2\gamma} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^n\right)\mathbf{B}_0\right)\mathbf{H}.
\end{aligned}$$

Next we use the above inequality to build a fine lower bound for $\mathbf{S}_n$:

$$\begin{aligned}
\mathbf{S}_n &= (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{S}_{n-1} + \mathbf{B}_0 = (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{S}_{n-1} + \gamma^2(\mathcal{M} - \widetilde{\mathcal{M}}) \circ \mathbf{S}_{n-1} + \mathbf{B}_0 \\
&\succeq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{S}_{n-1} + \frac{\beta\gamma}{2} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n-1}\right)\mathbf{B}_0\right)\mathbf{H} + \mathbf{B}_0.
\end{aligned}$$

Solving the recursion we obtain

$$\begin{aligned}\mathbf{S}_n &\succeq \sum_{t=0}^{n-1}(\mathcal{I}-\gamma\widetilde{\mathcal{T}})^t \circ \left\{\frac{\beta\gamma}{2}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right)\mathbf{H}+\mathbf{B}_0\right\}\\ &= \frac{\beta\gamma}{2}\sum_{t=0}^{n-1}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right)\cdot(\mathbf{I}-\gamma\mathbf{H})^{2t}\mathbf{H}\\ &\quad+\sum_{t=0}^{n-1}(\mathbf{I}-\gamma\mathbf{H})^t\cdot\mathbf{B}_0\cdot(\mathbf{I}-\gamma\mathbf{H})^t.\end{aligned}$$

For the first term, noticing the following:

$$\begin{aligned}&\sum_{t=0}^{n-1}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right)\cdot(\mathbf{I}-\gamma\mathbf{H})^{2t}\mathbf{H}\\ &\succeq \sum_{t=0}^{n-1}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right)\cdot(\mathbf{I}-2\gamma\mathbf{H})^{t}\mathbf{H}\\ &\succeq \sum_{t=0}^{n/2-1}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n-1-t}\right)\mathbf{B}_0\right)\cdot(\mathbf{I}-2\gamma\mathbf{H})^{t}\mathbf{H}\\ &\succeq \operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\right)\mathbf{B}_0\right)\cdot\sum_{t=0}^{n/2-1}(\mathbf{I}-2\gamma\mathbf{H})^{t}\mathbf{H}\\ &= \frac{1}{2\gamma}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\right)\mathbf{B}_0\right)\cdot\left(\mathbf{I}-(\mathbf{I}-2\gamma\mathbf{H})^{n/2}\right)\\ &\succeq \frac{1}{2\gamma}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\right)\mathbf{B}_0\right)\cdot\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{n/2}\right),\end{aligned}$$

inserting which back to the lower bound for $\mathbf{S}_n$, we complete the proof.

□

Lemma 2.7.5. Suppose Assumptions 2.2.1 and 2.2.5 hold and $N \geq 2$. If the stepsize satisfies $\gamma < 1/\gamma_1$, then

$$\begin{aligned}\text{bias} \geq\ & \frac{1}{100\gamma^2N^2}\cdot\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}}+\frac{1}{100}\cdot\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\\ &+\frac{\beta\left(\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}}+\gamma N\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right)}{1000\gamma N^2}\cdot\left(k^*+\gamma^2N^2\sum_{i>k^*}\lambda_i^2\right),\end{aligned}$$

where $k^*=\max\{k:\lambda_k\geq\frac{1}{N\gamma}\}$.

Proof. According to (2.7.4) and Lemma 2.7.4, we have that

$$\begin{aligned}\text{bias} &\geq \frac{1}{2\gamma N^2}\big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \mathbf{S}_{N/2+1}\big\rangle \geq \frac{1}{2\gamma N^2}\big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \mathbf{S}_{N/2}\big\rangle \\ &\geq \underbrace{\frac{\beta}{8\gamma N^2}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right)\mathbf{B}_0\right)\cdot\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right\rangle}_{I_1} \\ &\quad + \underbrace{\frac{1}{2\gamma N^2}\Big\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \sum_{t=0}^{N/2-1}(\mathbf{I}-\gamma\mathbf{H})^t\cdot\mathbf{B}_0\cdot(\mathbf{I}-\gamma\mathbf{H})^t\Big\rangle}_{I_2}.\end{aligned}$$

The first term is lower bounded by

$$\begin{aligned}I_1 &\geq \frac{\beta}{8\gamma N^2}\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right)\mathbf{B}_0\right)\cdot\operatorname{tr}\left(\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right)^2\right) \\ &= \frac{\beta}{8\gamma N^2}\left(\sum_i\left(1-(1-\gamma\lambda_i)^{N/4}\right)\omega_i^2\right)\cdot\left(\sum_i\left(1-(1-\gamma\lambda_i)^{N/4}\right)^2\right),\end{aligned}$$

where $\omega_i = \mathbf{v}_i^\top(\mathbf{w}_0-\mathbf{w}^*)$ for $\mathbf{v}_1,\ldots,\mathbf{v}_d$ being the eigenvectors of $\mathbf{H}$; and the second term is lower bounded by

$$\begin{aligned}I_2 &= \frac{1}{2\gamma N^2}\Big\langle\sum_{t=0}^{N/2-1}(\mathbf{I}-\gamma\mathbf{H})^{2t}\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}\right),\mathbf{B}_0\Big\rangle \\ &\geq \frac{1}{2\gamma N^2}\Big\langle\sum_{t=0}^{N/2-1}(\mathbf{I}-2\gamma\mathbf{H})^{t}\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}\right),\mathbf{B}_0\Big\rangle \\ &\geq \frac{1}{4\gamma^2 N^2}\big\langle\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}\right)^2\mathbf{H}^{-1},\mathbf{B}_0\big\rangle \\ &\geq \frac{1}{4\gamma^2 N^2}\big\langle\left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/4}\right)^2\mathbf{H}^{-1},\mathbf{B}_0\big\rangle \\ &= \frac{1}{4\gamma^2 N^2}\sum_i\left(1-(1-\gamma\lambda_i)^{N/4}\right)^2\lambda_i^{-1}\omega_i^2.\end{aligned}$$

To further lower bound the two terms, noticing the following inequality:

$$1-(1-\gamma\lambda_i)^{\frac{N}{4}} \geq \begin{cases}1-(1-\frac{1}{N})^{\frac{N}{4}} \geq 1-e^{-\frac{1}{4}} \geq \frac{1}{5}, & \lambda_i \geq \frac{1}{\gamma N},\\ \frac{N}{4}\cdot\gamma\lambda_i - \frac{N(N-4)}{32}\cdot\gamma^2\lambda_i^2 \geq \frac{N}{5}\cdot\gamma\lambda_i, & \lambda_i < \frac{1}{\gamma N}.\end{cases}$$

Plugging this into the bounds for I_1 and I_2, and setting $k^* := \max\{k : \lambda_k \geq 1/(\gamma N)\}$, we then obtain that

$$\begin{aligned} I_1 &\geq \frac{\beta}{8\gamma N^2} \cdot \left(\frac{1}{5} \cdot \sum_{i \leq k^*} \omega_i^2 + \frac{\gamma N}{5} \sum_{i > k^*} \lambda_i \omega_i^2 \right) \cdot \left(\frac{1}{25} \cdot k^* + \frac{\gamma^2 N^2}{25} \cdot \sum_{i > k^*} \lambda_i^2 \right) \\ &= \frac{\beta}{1000\gamma N^2} \cdot \left(\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{0:k^*}}^2 + \gamma N \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right) \cdot \left(k^* + \gamma^2 N^2 \sum_{i > k^*} \lambda_i^2 \right), \end{aligned}$$

and that

$$\begin{aligned} I_2 &\geq \frac{1}{4\gamma^2 N^2} \left(\frac{1}{25} \cdot \sum_{i \leq k^*} \lambda_i^{-1} \omega_i^2 + \frac{\gamma^2 N^2}{25} \cdot \sum_{i > k^*} \lambda_i \omega_i^2 \right) \\ &= \frac{1}{100\gamma^2 N^2} \left(\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \gamma^2 N^2 \|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \right). \end{aligned}$$

Summing up the two terms completes the proof.

□

2.7.4 Proof of Theorem 2.2.6

Proof. Plugging the bounds of the bias error and variance error in Lemmas 2.7.5 and 2.7.3 into Lemma 2.7.1 immediately completes the proof. □

2.8 Proofs for Tail-Averaging

In this section, we provide the proofs for SGD with tail-averaging. Recall that in tail-averaging, we take average from the s-th iterate, i.e., the output of the tail-average SGD is

$$\overline{\mathbf{w}}_{s:s+N} = \frac{1}{N} \sum_{t=s}^{s+N-1} \mathbf{w}_t.$$

2.8.1 Upper Bounds for Tail-Averaging

The following two lemmas are straightforward extensions of Lemmas 2.6.2 and 2.6.3.

Lemma 2.8.1 (Variant of Lemma 2.6.2)**.**

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) = \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}_{s:s+N} \otimes \bar{\boldsymbol{\beta}}_{s:s+N}]\rangle \leq \left(\sqrt{\text{bias}} + \sqrt{\text{variance}}\right)^2,$$

where

$$\text{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}^{\text{bias}}_{s:s+N} \otimes \bar{\boldsymbol{\beta}}^{\text{bias}}_{s:s+N}]\rangle, \qquad \text{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}^{\text{variance}}_{s:s+N} \otimes \bar{\boldsymbol{\beta}}^{\text{variance}}_{s:s+N}]\rangle.$$

Lemma 2.8.2 (Variant of Lemma 2.6.3)**.** Recall iterates (2.4.4) and (2.4.5). If the stepsize satisfies $\gamma < 1/\lambda_1$, the bias error and variance error are upper bounded respectively as follows:

$$\text{bias} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}^{\text{bias}}_{s:s+N} \otimes \bar{\boldsymbol{\beta}}^{\text{bias}}_{s:s+N}]\rangle \leq \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_{s+t}\right\rangle,$$

$$\text{variance} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\beta}}^{\text{variance}}_{s:s+N} \otimes \bar{\boldsymbol{\beta}}^{\text{variance}}_{s:s+N}] \leq \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_{s+t}\right\rangle.$$

Proof. By replacing $\mathbf{B}_0$ and $\mathbf{C}_0$ by $\mathbf{B}_s$ and $\mathbf{C}_s$ in the proof of Lemma 2.6.3, and repeating the remaining arguments, we can easily complete the proof. □

2.8.1.1 Bounding the Variance Error

Lemma 2.8.3 (Variant of Lemma 2.6.7)**.** Under Assumptions 2.2.1, 2.2.3 and 2.6.4, if the stepsize satisfies $\gamma < 1/R^2$, then it holds that

$$\text{variance} \leq \frac{\sigma^2}{1-\gamma R^2} \cdot \left(\frac{k^*}{N} + \gamma \cdot \sum_{k^* < i \leq k^\dagger} \lambda_i + \gamma^2(s+N) \cdot \sum_{i > k^\dagger} \lambda_i^2\right),$$

where $k^* = \min\{k : \lambda_i < \frac{1}{\gamma N}\}$ and $k^\dagger = \min\{k : \lambda_i < \frac{1}{\gamma(s+N)}\}$.

Proof. By Lemma 2.8.2, we can bound the variance error as follows

$$\begin{aligned}
\text{variance} &\leq \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle (\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{C}_{s+t}\right\rangle \\
&= \frac{1}{\gamma N^2}\sum_{t=0}^{N-1}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{C}_{s+t}\right\rangle \\
&\leq \frac{\sigma^2}{N^2(1-\gamma R^2)}\sum_{t=0}^{N-1}\left\langle \mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \left(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{s+t}\right)\right\rangle \\
&= \frac{\sigma^2}{N^2(1-\gamma R^2)}\sum_{i}\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N-t}\right)\left(1-(1-\gamma\lambda_i)^{s+t}\right) \\
&\leq \frac{\sigma^2}{N^2(1-\gamma R^2)}\sum_{i}\sum_{t=0}^{N-1}\left(1-(1-\gamma\lambda_i)^{N}\right)\left(1-(1-\gamma\lambda_i)^{s+N}\right) \\
&= \frac{\sigma^2}{N(1-\gamma R^2)}\sum_{i}\left(1-(1-\gamma\lambda_i)^{N}\right)\left(1-(1-\gamma\lambda_i)^{s+N}\right),
\end{aligned}$$

where the second inequality is due to Lemma 2.6.6, $\{\lambda_i\}_{i\geq 1}$ are the eigenvalues of $\mathbf{H}$ and are sorted in decreasing order. Now we will move to upper bound the quantity $\left(1-(1-\gamma\lambda_i)^N\right)\left(1-(1-\gamma\lambda_i)^{s+N}\right)$, which will be separately discussed according to the following three cases: (1) $\gamma\lambda_i \geq 1/N$, (2) $1/(s+N) \leq \gamma\lambda_i < 1/N$, and (3) $\gamma\lambda < 1/(s+N)$. In case (1), we can crudely bound this quantity as follows,

$$\left(1-(1-\gamma\lambda_i)^{N}\right)\left(1-(1-\gamma\lambda_i)^{s+N}\right) \leq 1.$$

In case (2), we can use $(1-\gamma\lambda_i)^N \geq 1-\gamma N\lambda_i$ and get

$$\left(1-(1-\gamma\lambda_i)^{N}\right)\left(1-(1-\gamma\lambda_i)^{s+N}\right) \leq \gamma N\lambda_i \cdot 1 = \gamma N\lambda_i.$$

In case (3), we can use $(1-\gamma\lambda_i)^N \geq 1-\gamma N\lambda_i$ and $(1-\gamma\lambda_i)^{s+N} \geq 1-\gamma(s+N)\lambda_i$, and get

$$\left(1-(1-\gamma\lambda_i)^{N}\right)\left(1-(1-\gamma\lambda_i)^{s+N}\right) \leq \gamma N\lambda_i \cdot \gamma(s+N)\lambda_i = \gamma^2 N(s+N)\lambda_i^2.$$

Therefore, set $k^* = \min\{k : \lambda_i < \frac{1}{N\gamma}\}$ and $k^\dagger = \min\{k : \lambda_i < \frac{1}{(s+N)\gamma}\}$, we have

$$\begin{aligned}
\text{variance} &\leq \frac{\sigma^2}{N(1-\gamma R^2)}\cdot\left(k^* + \gamma N\sum_{k^*<i\leq k^\dagger}\lambda_i + \gamma^2 N(s+N)\sum_{i>k^\dagger}\lambda_i^2\right) \\
&= \frac{\sigma^2}{1-\gamma R^2}\cdot\left(\frac{k^*}{N} + \gamma\cdot\sum_{k^*<i\leq k^\dagger}\lambda_i + \gamma^2(s+N)\cdot\sum_{i>k^\dagger}\lambda_i^2\right).
\end{aligned}$$

This completes the proof.

□

2.8.1.2 Bounding the Bias Error

Similarly to (2.6.13) and using Lemma 2.8.2, we have the following upper bound for the bias error:

$$\begin{aligned}\text{bias} &\leq \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H},\mathbf{B}_{s+t}\right\rangle \\ &= \frac{1}{\gamma N^2}\sum_{t=0}^{N-1}\left\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t},\mathbf{B}_{s+t}\right\rangle \\ &\leq \frac{1}{\gamma N^2}\left\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N},\sum_{t=0}^{N-1}\mathbf{B}_{s+t}\right\rangle. \end{aligned} \tag{2.8.1}$$

Let $\mathbf{S}_{s:s+t}=\sum_{k=s}^{s+t-1}\mathbf{B}_k$, then we only need to establish an upper bound for $\mathbf{S}_{s:s+N}$.

Lemma 2.8.4 (Variant of Lemma 2.6.11)**.** Let $\mathbf{S}_{s:s+t}=\sum_{k=s}^{s+t-1}\mathbf{B}_k$ for any $t\geq s$ and $\mathbf{B}_{a,b}=\mathbf{B}_a-(\mathbf{I}-\gamma\mathbf{H})^{b-a}\mathbf{B}_a(\mathbf{I}-\gamma\mathbf{H})^{b-a}$. Under Assumptions 2.2.1 and 2.2.2, if the stepsize satisfies $\gamma<1/\big(\alpha\operatorname{tr}(\mathbf{H})\big)$, it holds that

$$\mathbf{S}_{s:s+N}\preceq\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{k+s}\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^{k+s}+\frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{s,s+N}+\mathbf{B}_{0,s})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{2k}\mathbf{H}.$$

Proof. Based on the definition of $\mathbf{S}_{s:s+t}$, we have

$$\mathbf{S}_{s:s+t}=\sum_{k=s}^{s+t-1}\mathbf{B}_k=\sum_{k=0}^{t-1}(\mathcal{I}-\gamma\mathcal{T})^k\circ\mathbf{B}_s=(\mathcal{I}-\gamma\mathcal{T})\circ\mathbf{S}_{s:s+t-1}+\mathbf{B}_s.$$

Therefore, following the similar proof technique of Lemma 2.6.11, we can get

$$\mathbf{S}_{s:s+N}\preceq\underbrace{\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{k}\mathbf{B}_s(\mathbf{I}-\gamma\mathbf{H})^{k}}_{I_1}+\underbrace{\frac{\gamma\alpha\operatorname{tr}(\mathbf{B}_{s,s+N})}{1-\gamma\alpha\operatorname{tr}(\mathbf{H})}\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{2k}\mathbf{H}}_{I_2}. \tag{2.8.2}$$

Now we will upper bound I_1, which requires a carefully characterization on $\mathbf{B}_s$. Particularly, the update form of $\mathbf{B}_k$ in (2.4.2) implies

$$\mathbf{B}_k = (\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{B}_{k-1} \preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{B}_{k-1} + \gamma^2 \mathcal{M} \circ \mathbf{B}_{k-1}.$$

By Assumption 2.2.2, we have $\mathcal{M} \circ \mathbf{B}_k \preceq \alpha \operatorname{tr}(\mathbf{H}\mathbf{B}_k) \cdot \mathbf{H}$. Thus,

$$\begin{aligned}
\mathbf{B}_k &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{B}_{k-1} + \gamma^2 \mathcal{M} \circ \mathbf{B}_{k-1} \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{B}_{k-1} + \alpha\gamma^2 \operatorname{tr}(\mathbf{H}\mathbf{B}_{k-1}) \cdot \mathbf{H} \\
&= (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{B}_0 + \alpha\gamma^2 \sum_{t=0}^{k-1} \operatorname{tr}(\mathbf{H}\mathbf{B}_t) \cdot (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^{k-1-t} \circ \mathbf{H} \\
&\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{B}_0 + \alpha\gamma^2 \sum_{t=0}^{k-1} \operatorname{tr}(\mathbf{H}\mathbf{B}_t) \cdot \mathbf{H} \qquad (2.8.3)
\end{aligned}$$

where in the third inequality we use the fact that $\mathcal{I} - \gamma\widetilde{\mathcal{T}}$ is a PSD mapping and the last inequality is due to $(\mathcal{I} - \gamma\widetilde{\mathcal{T}})^{k-1-t}\mathbf{H} = (\mathbf{I} - \gamma\mathbf{H})^{2(k-1-t)}\mathbf{H} \preceq \mathbf{H}$. Next we will upper bound $\sum_{t=0}^{k-1} \operatorname{tr}(\mathbf{H}\mathbf{B}_t)$. Recall the definition of $\boldsymbol{\beta}_k^{\text{bias}}$ and its update rule, we have

$$\begin{aligned}
&\mathbb{E}[\|\boldsymbol{\beta}_k^{\text{bias}}\|_2^2 | \boldsymbol{\beta}_{k-1}^{\text{bias}}] \\
&= \mathbb{E}[\|(\mathbf{I} - \gamma\mathbf{x}_k\mathbf{x}_k^\top)\boldsymbol{\beta}_{k-1}^{\text{bias}}\|_2^2 | \boldsymbol{\beta}_{k-1}^{\text{bias}}] \\
&= \|\boldsymbol{\beta}_{k-1}^{\text{bias}}\|_2^2 - 2\gamma\mathbb{E}[\langle \mathbf{x}_k\mathbf{x}_k^\top, \boldsymbol{\beta}_{k-1}^{\text{bias}} \otimes \boldsymbol{\beta}_{k-1}^{\text{bias}} \rangle | \boldsymbol{\beta}_{k-1}^{\text{bias}}] + \gamma^2\mathbb{E}[\langle \mathbf{x}_k\mathbf{x}_k^\top\mathbf{x}_k\mathbf{x}_k^\top, \boldsymbol{\beta}_{k-1}^{\text{bias}} \otimes \boldsymbol{\beta}_{k-1}^{\text{bias}} | \boldsymbol{\beta}_{k-1}^{\text{bias}}] \\
&= \|\boldsymbol{\beta}_{k-1}^{\text{bias}}\|_2^2 - 2\gamma\langle \mathbf{H}, \boldsymbol{\beta}_{k-1}^{\text{bias}} \otimes \boldsymbol{\beta}_{k-1}^{\text{bias}} \rangle + \gamma^2\langle \mathcal{M} \circ \mathbf{I}, \boldsymbol{\beta}_{k-1}^{\text{bias}} \otimes \boldsymbol{\beta}_{k-1}^{\text{bias}} \rangle \\
&\leq \|\boldsymbol{\beta}_{k-1}^{\text{bias}}\|_2^2 - \big(2\gamma - \gamma^2\alpha \operatorname{tr}(\mathbf{H})\big) \cdot \langle \mathbf{H}, \boldsymbol{\beta}_{k-1}^{\text{bias}} \otimes \boldsymbol{\beta}_{k-1}^{\text{bias}} \rangle,
\end{aligned}$$

where the inequality is due to the fact that $\mathcal{M} \circ \mathbf{I} \preceq \alpha \operatorname{tr}(\mathbf{H})\mathbf{H}$. Note that $\mathbf{B}_k = \mathbb{E}[\boldsymbol{\beta}_k^{\text{bias}} \otimes \boldsymbol{\beta}_k^{\text{bias}}]$, taking total expectation further gives

$$\operatorname{tr}(\mathbf{B}_k) \leq \operatorname{tr}(\mathbf{B}_{k-1}) - \big(2\gamma - \gamma^2\alpha \operatorname{tr}(\mathbf{H})\big) \cdot \operatorname{tr}(\mathbf{H}\mathbf{B}_{k-1}),$$

which implies that

$$\sum_{t=0}^{k-1} \operatorname{tr}(\mathbf{H}\mathbf{B}_t) \leq \frac{\operatorname{tr}(\mathbf{B}_0) - \operatorname{tr}(\mathbf{B}_k)}{2\gamma - \gamma^2\alpha \operatorname{tr}(\mathbf{H})}. \qquad (2.8.4)$$

Substituting (2.8.4) into (2.8.3) gives

$$\begin{aligned}\mathbf{B}_k &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{B}_0 + \alpha\gamma^2 \sum_{t=0}^{k-1} \operatorname{tr}(\mathbf{H}\mathbf{B}_t) \cdot \mathbf{H} \\ &\preceq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^k \circ \mathbf{B}_0 + \frac{\gamma\alpha \operatorname{tr}(\mathbf{B}_0 - \mathbf{B}_k)}{2 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \cdot \mathbf{H}.\end{aligned}$$

Therefore, we further have

$$I_1 \preceq \sum_{k=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^{k+s} \mathbf{B}_0 (\mathbf{I} - \gamma\mathbf{H})^{k+s} + \frac{\gamma\alpha \operatorname{tr}(\mathbf{B}_0 - \mathbf{B}_s)}{2 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \sum_{k=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^{2k} \mathbf{H}. \tag{2.8.5}$$

Further note that $\mathbf{B}_s = (\mathcal{I} - \gamma\mathcal{T})^s \mathbf{B}_0$ and $\mathcal{T} \succeq \widetilde{\mathcal{T}}$, we have

$$\begin{aligned}\operatorname{tr}(\mathbf{B}_0 - \mathbf{B}_s) &= \operatorname{tr}\left(\mathbf{B}_0 - (\mathcal{I} - \gamma\mathcal{T})^s \mathbf{B}_0\right) \\ &\leq \operatorname{tr}\left(\mathbf{B}_0 - (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^s \mathbf{B}_0\right) \\ &\leq \operatorname{tr}\left(\mathbf{B}_0 - (\mathbf{I} - \gamma\mathbf{H})^s \mathbf{B}_0 (\mathbf{I} - \gamma\mathbf{H})^s\right) \\ &= \operatorname{tr}(\mathbf{B}_{0,s}).\end{aligned}$$

Now, we can substitute the above inequality and (2.8.5) into (2.8.2) and obtain the following upper bound on $\mathbf{S}_{s:s+N}$,

$$\mathbf{S}_{s:s+N} \preceq I_1 + I_2 \preceq \sum_{k=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^{k+s} \mathbf{B}_0 (\mathbf{I} - \gamma\mathbf{H})^{k+s} + \frac{\gamma\alpha \operatorname{tr}(\mathbf{B}_{s,s+N} + \mathbf{B}_{0,s})}{1 - \gamma\alpha \operatorname{tr}(\mathbf{H})} \sum_{k=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^{2k} \mathbf{H},$$

where we use the fact that $0 \leq 1 - \gamma\alpha \operatorname{tr}(\mathbf{H}) \leq 2 - \gamma\alpha \operatorname{tr}(\mathbf{H})$. This completes the proof. □

Lemma 2.8.5 (Variant of Lemma 2.6.12)**.** Under Assumptions 2.2.1 and 2.2.2, if the stepsize satisfies $\gamma < 1/(\alpha \operatorname{tr}(\mathbf{H}))$, it holds that

$$\begin{aligned}\text{bias} \leq \frac{1}{\gamma^2 N^2} \cdot \left\|(\mathbf{I} - \gamma\mathbf{H})^s (\mathbf{w}_0 - \mathbf{w}^*)\right\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \left\|(\mathbf{I} - \gamma\mathbf{H})^s (\mathbf{w}_0 - \mathbf{w}^*)\right\|^2_{\mathbf{H}_{k^*:\infty}} \\ + \frac{4\alpha\big(\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{0:k^*}} + (s+N)\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{\gamma(1 - \gamma\alpha \operatorname{tr}(\mathbf{H}))} \cdot \left(\frac{k^*}{N^2} + \gamma^2 \sum_{i>k^*} \lambda_i^2\right),\end{aligned}$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma N}\}$.

Proof. Substituting the upper bound of $\mathbf{S}_{s:s+N}$ into (2.8.1), we can get

$$\text{bias} \le \underbrace{\frac{\alpha \operatorname{tr}(\mathbf{B}_{s,s+N} + \mathbf{B}_{0,s})}{N^2(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))} \sum_{k=0}^{N-1} \big\langle \mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N, (\mathbf{I}-\gamma\mathbf{H})^{2k}\mathbf{H} \big\rangle}_{I_1}$$
$$+ \underbrace{\frac{1}{\gamma N^2} \sum_{k=0}^{N-1} \big\langle \mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N, (\mathbf{I}-\gamma\mathbf{H})^{k+s}\mathbf{B}_0(\mathbf{I}-\gamma\mathbf{H})^{k+s} \big\rangle}_{I_2}. \tag{2.8.6}$$

By (2.6.21), we can get the following bound on I_1,

$$I_1 \le \frac{\alpha \operatorname{tr}(\mathbf{B}_{s,s+N} + \mathbf{B}_{0,s})}{\gamma(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))} \cdot \bigg(\frac{k^*}{N^2} + \gamma^2 \sum_{i>k^*} \lambda_i^2 \bigg). \tag{2.8.7}$$

Then following the same procedure in (2.6.22), we have

$$\operatorname{tr}(\mathbf{B}_{s,s+N} + \mathbf{B}_{0,s}) \le 2\operatorname{tr}(\mathbf{B}_{0,s+N}) \le 4\big(\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{0:k^*}} + (s+N)\gamma \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \big)$$

where $k^* = \max\big\{k : \lambda_k \ge \frac{1}{\gamma N}\big\}$ (in fact k^* can be arbitrary choosen). Plugging this into (2.8.6) gives

$$I_1 \le \frac{4\alpha\big(\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{I}_{0:k^*}} + (s+N)\gamma \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \big)}{\gamma(1-\gamma\alpha\operatorname{tr}(\mathbf{H}))} \cdot \bigg(\frac{k^*}{N^2} + \gamma^2 \sum_{i>k^*} \lambda_i^2 \bigg).$$

Additionally, we have the following upper bound on I_2,

$$\begin{aligned} I_2 &= \frac{1}{\gamma N^2} \sum_{k=0}^{N-1} \big\langle (\mathbf{I}-\gamma\mathbf{H})^{2(k+s)} \big(\mathbf{I} - (\mathbf{I}-\gamma\mathbf{H})^N \big), \mathbf{B}_0 \big\rangle \\ &\le \frac{1}{\gamma N^2} \sum_{k=0}^{N-1} \big\langle (\mathbf{I}-\gamma\mathbf{H})^{k+2s} - (\mathbf{I}-\gamma\mathbf{H})^{N+k+2s}, \mathbf{B}_0 \big\rangle. \end{aligned}$$

Similar to the proof of Lemma 2.6.12, let $\mathbf{v}_1, \mathbf{v}_2, \dots$ be the eigenvectors of $\mathbf{H}$ corresponding

to its eigenvalues $\lambda_1, \lambda_2, \dots$ and $\omega_i = \mathbf{v}_i^\top (\mathbf{I} - \gamma\mathbf{H})^s (\mathbf{w}_0 - \mathbf{w}^*)$, we have

$$\begin{aligned}
I_2 &\le \frac{1}{\gamma N^2} \sum_{k=0}^{N-1} \left\langle (\mathbf{I} - \gamma\mathbf{H})^k - (\mathbf{I} - \gamma\mathbf{H})^{N+k}, (\mathbf{I} - \gamma\mathbf{H})^{2s} \mathbf{B}_0 \right\rangle \\
&= \frac{1}{\gamma N^2} \sum_{k=0}^{N-1} \sum_i \left[(1 - \gamma\lambda_i)^k - (1 - \gamma\lambda_i)^{N+k} \right] \omega_i^2 \\
&= \frac{1}{\gamma^2 N^2} \sum_i \frac{\omega_i^2}{\lambda_i} \left[1 - (1 - \gamma\lambda_i)^N \right]^2 \\
&\le \frac{1}{\gamma^2 N^2} \sum_i \frac{\omega_i^2}{\lambda_i} \cdot \min\{1, \gamma^2 N^2 \lambda_i^2\} \\
&\le \frac{1}{\gamma^2 N^2} \cdot \sum_{i \le k^*} \frac{\omega_i^2}{\lambda_i} + \sum_{i > k^*} \lambda_i \omega_i^2 \\
&= \frac{1}{\gamma^2 N^2} \cdot \left\| (\mathbf{I} - \gamma\mathbf{H})^s (\mathbf{w}_0 - \mathbf{w}^*) \right\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \left\| (\mathbf{I} - \gamma\mathbf{H})^s (\mathbf{w}_0 - \mathbf{w}^*) \right\|_{\mathbf{H}_{k^*:\infty}}^2, \qquad (2.8.8)
\end{aligned}$$

where $k^* = \max\{k : \lambda_k \ge \frac{1}{\gamma N}\}$. Combining (2.8.7) and (2.8.8) immediately completes the proof.

□

2.8.1.3 Proof of Theorem 2.2.10

Proof. By Lemma 2.8.2, it suffices to substitute into the upper bounds on the bias and variance errors. In particular, by Young's inequality we have

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \le \left(\sqrt{\text{bias}} + \sqrt{\text{variance}} \right)^2 \le 2 \cdot \text{bias} + 2 \cdot \text{variance}.$$

Then we can directly substitute the bounds of variance and bias we proved in Lemmas 2.8.3 and 2.8.5. In particular, by Assumptions 2.2.2 we can directly get $R^2 = \alpha \operatorname{tr}(\mathbf{H})$. Therefore,

it holds that

$$
\begin{aligned}
&\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \\
&\le 2\Bigg[\frac{1}{\gamma^2 N^2}\cdot\big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}_{k^*:\infty}} \\
&\quad + \frac{2\alpha\|\mathbf{w}_0-\mathbf{w}^*\|_2^2}{\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot\left(\frac{k^*}{N^2}+\gamma^2\sum_{i>k^*}\lambda_i^2\right) \\
&\quad + \frac{\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\cdot\left(\frac{k^*}{N}+\gamma\cdot\sum_{k^*<i\le k^\dagger}\lambda_i+\gamma^2(s+N)\cdot\sum_{i>k^\dagger}\lambda_i^2\right)\Bigg] \\
&= 2\cdot\mathrm{EffectiveBias} + 2\cdot\mathrm{EffectiveVar},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathrm{EffectiveBias} &= \frac{1}{\gamma^2 N^2}\cdot\big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \big\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\big\|^2_{\mathbf{H}_{k^*:\infty}} \\
\mathrm{EffectiveVar} &= \frac{\sigma^2}{1-\gamma\alpha\,\mathrm{tr}(\mathbf{H})}\cdot\left(\frac{k^*}{N}+\gamma\cdot\sum_{k^*<i\le k^\dagger}\lambda_i+\gamma^2(s+N)\cdot\sum_{i>k^\dagger}\lambda_i^2\right) \\
&\quad + \frac{4\alpha\big(\|\mathbf{w}_0-\mathbf{w}^*\|_{\mathbf{I}_{0:k^*}}+(s+N)\gamma\|\mathbf{w}_0-\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{N\gamma(1-\gamma\alpha\,\mathrm{tr}(\mathbf{H}))}\cdot\left(\frac{k^*}{N}+\gamma^2 N\sum_{i>k^*}\lambda_i^2\right).
\end{aligned}
$$

□

2.8.2 Lower Bounds for Tail-Averaging

In this part we assume the noise is well-specified as in (2.2.1), and consider the SGD with tail-averaging

$$
\overline{\mathbf{w}}_{s:s+N} = \frac{1}{N}\sum_{t=s}^{s+N}\mathbf{w}_t.
$$

The following lemma is a variant of Lemma 2.7.1, and lowers bound the excess risk.

Lemma 2.8.6 (Variant of Lemma 2.7.1)**.** Suppose the model noise ξ_t is well-specified, i.e.,

ξ_t and $\mathbf{x}_t$ are independent and $\mathbb{E}[\xi_t] = 0$. Then

$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N}) - L(\mathbf{w}^*)] \geq \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{B}_{s+t} \right\rangle + \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{C}_{s+t} \right\rangle.$$

We then present the lower bound for the variance error.

Lemma 2.8.7 (Variant of Lemma 2.7.3)**.** Suppose Assumptions 2.2.1 hold. Suppose the noise is well-specified (as in (2.2.1)). Suppose $N \geq 500$. Denote

$$\text{variance} = \frac{1}{2N^2} \cdot \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{C}_{s+t} \right\rangle.$$

If the stepsize satisfies $\gamma < 1/\lambda_1$, then

$$\text{variance} \geq \frac{\sigma_{\text{noise}}^2}{600} \left(\frac{k^*}{N} + \gamma \cdot \sum_{k^* < i \leq k^\dagger} \lambda_i + (s+N)\gamma^2 \cdot \sum_{i > k^\dagger} \lambda_i^2 \right),$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$ and $k^\dagger = \max\{k : \lambda_k \geq \frac{1}{(s+N)\gamma}\}$.

Proof. We can lower bound the variance error as follows

$$\begin{aligned}
\text{variance} &= \frac{1}{2N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{C}_{s+t} \right\rangle \\
&= \frac{1}{2\gamma N^2} \sum_{t=0}^{N-1} \left\langle \mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{N-t}, \mathbf{C}_{s+t} \right\rangle \\
&\geq \frac{\sigma_{\text{noise}}^2}{4N^2} \sum_{t=0}^{N-1} \left\langle \mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{N-t}, \mathbf{I} - (\mathbf{I} - \gamma \mathbf{H})^{2(s+t)} \right\rangle \qquad \text{(use Lemma 2.7.2)} \\
&= \frac{\sigma_{\text{noise}}^2}{4N^2} \sum_{i} \sum_{t=0}^{N-1} \left(1 - (1 - \gamma\lambda_i)^{N-t}\right) \left(1 - (1 - \gamma\lambda_i)^{2(s+t)}\right) \\
&\geq \frac{\sigma_{\text{noise}}^2}{4N^2} \sum_{i} \sum_{t=0}^{N-1} \left(1 - (1 - \gamma\lambda_i)^{N-t-1}\right) \left(1 - (1 - \gamma\lambda_i)^{s+t}\right),
\end{aligned}$$

where $\{\lambda_i\}_{i\geq 1}$ are the eigenvalues of $\mathbf{H}$ and are sorted in decreasing order. Define

$$f(x) := \sum_{t=0}^{N-1} \left(1-(1-x)^{N-t-1}\right)\left(1-(1-x)^{s+t}\right), \qquad 0<x<1,$$

then

$$\text{variance} \geq \frac{\sigma^2_{\text{noise}}}{4N^2} \sum_i f(\gamma\lambda_i).$$

We have the following lower bound for $f(x)$.

$$\begin{aligned} f(x) &= \sum_{t=0}^{N-1} \left(1-(1-x)^{N-t-1}\right)\left(1-(1-x)^{s+t}\right) \\ &\geq \sum_{t=\frac{N}{4}}^{\frac{3N}{4}-1} \left(1-(1-x)^{N-t-1}\right)\left(1-(1-x)^{s+t}\right) \\ &\geq \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-(1-x)^{s+\frac{N}{4}}\right) \end{aligned}$$

We then bound $f(x)$ by the range of x.

1. For $x > 1/N$, we have that

$$\begin{aligned} f(x) &\geq \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-(1-x)^{\frac{N}{4}}\right) \\ &\geq \frac{N}{2}\left(1-\left(1-\frac{1}{N}\right)^{\frac{N}{4}}\right)\left(1-\left(1-\frac{1}{N}\right)^{\frac{N}{4}}\right) \\ &\geq \frac{N}{2}\left(1-\frac{1}{e^{1/4}}\right)\left(1-\frac{1}{e^{1/4}}\right) \geq \frac{N}{50}. \end{aligned}$$

2. For $1/N > x > 1/(s+N)$, we have that

$$\begin{aligned} f(x) &\geq \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-(1-x)^{\frac{s+N}{4}}\right) \\ &\geq \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-\left(1-\frac{1}{s+N}\right)^{\frac{s+N}{4}}\right) \\ &\geq \frac{N}{2}\left(1-\left(1-\frac{N}{8}x\right)\right)\left(1-\frac{1}{e^{1/4}}\right) \geq \frac{N^2 x}{100}. \end{aligned}$$

3. For $x < 1/(s+N) < 1/N$, we have that

$$\begin{aligned} f(x) &\geq \frac{N}{2}\left(1-(1-x)^{\frac{N}{4}}\right)\left(1-(1-x)^{s+\frac{N}{4}}\right) \\ &\geq \frac{N}{2}\left(1-\left(1-\frac{N}{8}x\right)\right)\left(1-\left(1-\frac{s+N/4}{2}x\right)\right) \\ &\geq \frac{(s+N)N^2}{128}x^2. \end{aligned}$$

In sum, we have that

$$f(x) \geq \begin{cases} \frac{N}{50}, & \frac{1}{N} \leq x < 1, \\ \frac{N^2}{100}x, & \frac{1}{s+N} \leq x < \frac{1}{N}, \\ \frac{(s+N)N^2}{128}x^2, & 0 < x < \frac{1}{s+N}. \end{cases}$$

Set $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$ and $k^\dagger = \max\{k : \lambda_k \geq \frac{1}{(s+N)\gamma}\}$, then

$$\begin{aligned} \text{variance} &\geq \frac{\sigma_{\text{noise}}^2}{4N^2}\sum_i f(\gamma\lambda_i) \\ &\geq \frac{\sigma_{\text{noise}}^2}{4N^2}\left(\frac{Nk^*}{50} + \frac{N^2}{100}\gamma\cdot\sum_{k^*<i\leq k^\dagger}\lambda_i + \frac{(s+N)N^2}{128}\gamma^2\cdot\sum_{i>k^\dagger}\lambda_i^2\right) \\ &\geq \frac{\sigma_{\text{noise}}^2}{600}\left(\frac{k^*}{N} + \gamma\cdot\sum_{k^*<i\leq k^\dagger}\lambda_i + (s+N)\gamma^2\cdot\sum_{i>k^\dagger}\lambda_i^2\right). \end{aligned}$$

This completes the proof.

□

Next we discuss the lower bound for the bias error. Similarly to (2.7.4) and using Lemma 2.8.6, we have that

$$\begin{aligned} \text{bias} &\geq \frac{1}{2N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_{s+t}\right\rangle = \frac{1}{2\gamma N^2}\sum_{t=0}^{N-1}\left\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{B}_{s+t}\right\rangle \\ &\geq \frac{1}{2\gamma N^2}\sum_{t=0}^{N/2}\left\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N-t}, \mathbf{B}_{s+t}\right\rangle \\ &\geq \frac{1}{2\gamma N^2}\left\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^{N/2}, \sum_{t=0}^{N/2}\mathbf{B}_{s+t}\right\rangle. \end{aligned} \tag{2.8.9}$$

Let $\mathbf{S}_{s:s+n} := \sum_{t=0}^{n-1} \mathbf{B}_{s+t} = \sum_{t=0}^{n-1} (\mathcal{I} - \gamma\mathcal{T})^t \circ \mathbf{B}_s$. We remain to build lower bound for $\mathbf{S}_{s:s+N/2+1}$. Comparing the definitions of $\mathbf{S}_{s:s+n}$ with $\mathbf{S}_n$, the only difference is that $\mathbf{B}_0$ is replaced by $\mathbf{B}_s$. Therefore we directly have the following lemma.

Lemma 2.8.8 (Variant of Lemma 2.7.4)**.** Suppose Assumptions 2.2.1 and 2.2.5 hold. If the stepsize satisfies $\gamma < 1/\lambda_1$, then for any $n \geq 2$, it holds that

$$\mathbf{S}_{s:s+n} \succeq \frac{\beta}{4} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n/2}\right)\mathbf{B}_s\right) \cdot \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{n/2}\right) + \sum_{t=0}^{n-1} (\mathbf{I} - \gamma\mathbf{H})^t \cdot \mathbf{B}_s \cdot (\mathbf{I} - \gamma\mathbf{H})^t.$$

Lemma 2.8.9 (Variant of Lemma 2.7.5)**.** Suppose Assumptions 2.2.1 and 2.2.5 hold and $N \geq 2$. Denote

$$\text{bias} = \frac{1}{2N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \left\langle (\mathbf{I} - \gamma\mathbf{H})^{k-t}\mathbf{H}, \mathbf{B}_{s+t} \right\rangle,$$

then if the stepsize satisfies $\gamma < 1/\gamma_1$, it holds that

$$\begin{aligned}\text{bias} \geq \frac{1}{100\gamma^2 N^2} &\left(\left\| (\mathbf{I} - \gamma\mathbf{H})^s (\mathbf{w}_0 - \mathbf{w}^*) \right\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \gamma^2 N^2 \left\| (\mathbf{I} - \gamma\mathbf{H})^s (\mathbf{w}_0 - \mathbf{w}^*) \right\|^2_{\mathbf{H}_{k^*:\infty}} \right) \\ &+ \frac{\beta \left\| \mathbf{w}_0 - \mathbf{w}^* \right\|^2_{\mathbf{H}_{k^\dagger:\infty}}}{16000N} \cdot \left(k^* + \gamma^2 N^2 \sum_{i>k^*} \lambda_i^2 \right),\end{aligned}$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}$ and $k^\dagger = \max\{k : \lambda_k \geq \frac{1}{(s+N)\gamma}\}$.

Proof. According to (2.8.9) and Lemma 2.8.8, we have that

$$\begin{aligned}\text{bias} &\geq \frac{1}{2\gamma N^2} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/2}, \mathbf{S}_{s:s+N/2+1} \right\rangle \geq \frac{1}{2\gamma N^2} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/2}, \mathbf{S}_{s:s+N/2} \right\rangle \\ &\geq \underbrace{\frac{\beta}{8\gamma N^2} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/4}\right)\mathbf{B}_s\right) \cdot \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/2}, \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/4} \right\rangle}_{I_1} \\ &\quad + \underbrace{\frac{1}{2\gamma N^2} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/2}, \sum_{t=0}^{N/2-1} (\mathbf{I} - \gamma\mathbf{H})^t \cdot \mathbf{B}_s \cdot (\mathbf{I} - \gamma\mathbf{H})^t \right\rangle}_{I_2}.\end{aligned}$$

Also noticing a lower bound for $\mathbf{B}_s$:

$$\mathbf{B}_s = (\mathcal{I} - \gamma\mathcal{T})^s \circ \mathbf{B}_0 \geq (\mathcal{I} - \gamma\widetilde{\mathcal{T}})^s \circ \mathbf{B}_0 = (\mathbf{I} - \gamma\mathbf{H})^s \cdot \mathbf{B}_0 \cdot (\mathbf{I} - \gamma\mathbf{H})^s.$$

Then the first term is lower bounded by

$$I_1 \geq \frac{\beta}{8\gamma N^2} \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/4}\right)(\mathbf{I} - \gamma\mathbf{H})^{2s}\mathbf{B}_0\right) \cdot \operatorname{tr}\left(\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/4}\right)^2\right)$$
$$= \frac{\beta}{8\gamma N^2}\left(\sum_i \left(1 - (1-\gamma\lambda_i)^{N/4}\right)(1-\gamma\lambda_i)^{2s}\omega_i^2\right) \cdot \left(\sum_i \left(1 - (1-\gamma\lambda_i)^{N/4}\right)^2\right),$$

where $\omega_i = \mathbf{v}_i^\top(\mathbf{w}_0 - \mathbf{w}^*)$ for $\mathbf{v}_1, \ldots, \mathbf{v}_d$ being the eigenvectors of $\mathbf{H}$; and the second term is lower bounded by

$$\begin{aligned}
I_2 &= \frac{1}{2\gamma N^2}\langle \sum_{t=0}^{N/2-1} (\mathbf{I} - \gamma\mathbf{H})^{2t}\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/2}\right), \mathbf{B}_s\rangle \\
&\geq \frac{1}{2\gamma N^2}\langle \sum_{t=0}^{N/2-1} (\mathbf{I} - 2\gamma\mathbf{H})^{t}\left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/2}\right), \mathbf{B}_s\rangle \\
&\geq \frac{1}{4\gamma^2 N^2}\langle \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/2}\right)^2 \mathbf{H}^{-1}, \mathbf{B}_s\rangle \\
&\geq \frac{1}{4\gamma^2 N^2}\langle \left(\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N/4}\right)^2 \mathbf{H}^{-1}, (\mathbf{I} - \gamma\mathbf{H})^s\mathbf{B}_0(\mathbf{I} - \gamma\mathbf{H})^s\rangle \\
&= \frac{1}{4\gamma^2 N^2}\sum_i \left(1 - (1-\gamma\lambda_i)^{N/4}\right)^2 \lambda_i^{-1}\left((1-\gamma\lambda_i)^s\omega_i\right)^2.
\end{aligned}$$

To further lower bound the two terms, noticing the following inequalities:

$$1 - (1-\gamma\lambda_i)^{\frac{N}{4}} \geq \begin{cases} 1 - (1 - \frac{1}{N})^{\frac{N}{4}} \geq 1 - e^{-\frac{1}{4}} \geq \frac{1}{5}, & \lambda_i \geq \frac{1}{\gamma N}, \\ \frac{N}{4}\cdot\gamma\lambda_i - \frac{N(N-4)}{32}\cdot\gamma^2\lambda_i^2 \geq \frac{N}{5}\cdot\gamma\lambda_i, & \lambda_i < \frac{1}{\gamma N}, \end{cases}$$

and

$$(1-\gamma\lambda_i)^{2s} \geq \begin{cases} 0, & \lambda_i \geq \frac{1}{\gamma s}, \\ (1 - \frac{1}{s})^{2s} \geq \frac{1}{16}, & \lambda_i < \frac{1}{\gamma s}. \end{cases}$$

Plugging these into the bounds for I_1 and I_2, and setting $k^* := \max\{k : \lambda_k \geq 1/(\gamma N)\}$ and $k^\dagger := \max\{k : \lambda_k \geq 1/(\gamma(s+N))\}$, we then obtain that

$$I_1 \geq \frac{\beta}{8\gamma N^2}\cdot\left(\frac{\gamma N}{80}\sum_{i>k^\dagger}\lambda_i\omega_i^2\right)\cdot\left(\frac{1}{25}\cdot k^* + \frac{\gamma^2 N^2}{25}\cdot\sum_{i>k^*}\lambda_i^2\right)$$
$$= \frac{\beta\left\|\mathbf{w}_0 - \mathbf{w}^*\right\|_{\mathbf{H}_{k^\dagger:\infty}}^2}{16000N}\cdot\left(k^* + \gamma^2 N^2\sum_{i>k^*}\lambda_i^2\right),$$

and that

$$I_2 \geq \frac{1}{4\gamma^2 N^2}\left(\frac{1}{25}\cdot\sum_{i\leq k^*}\lambda_i^{-1}\left((1-\gamma\lambda_i)^s\omega_i\right)^2+\frac{\gamma^2N^2}{25}\cdot\sum_{i>K^*}\lambda_i\left((1-\gamma\lambda_i)^s\omega_i\right)^2\right)$$
$$=\frac{1}{100\gamma^2N^2}\left(\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{0:k^*}}+\gamma^2N^2\,\|(\mathbf{I}-\gamma\mathbf{H})^s(\mathbf{w}_0-\mathbf{w}^*)\|^2_{\mathbf{H}_{k^*:\infty}}\right).$$

Summing up the two terms completes the proof.

□

2.8.2.1 Proof of Theorem 2.2.11

Proof. Plugging the bounds of the bias error and variance error in Lemmas 2.8.9 and 2.8.7 into Lemma 2.8.6 immediately completes the proof. □

2.9 Conclusions

This work considers the question of how well constant-stepsize SGD (with iterate average or tail average) generalizes for the linear regression problem in the over-parameterized regime. Our main result provides a sharp excess risk bound, stated in terms of the full eigenspectrum of the data covariance matrix. Our results reveal how a benign-overfitting phenomenon can occur under certain spectrum decay conditions on the data covariance.

There are number of more subtle points worth reflecting on:

Moving beyond the square loss. Focusing on linear regression is a means to understand phenomena that are exhibited more broadly. One natural next step here would be understand the analogues of the classical iterate averaging results [PJ92] for locally quadratic models, where decaying stepsizes are necessary for vanishing risk.

Sharper lower bounds. While our lower bound nearly matches our upper bound up to constant factors, there is notable gap in that the EffectiveVariance has a dependence on

$\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2$. This term is due to that even if $y - \mathbf{w}^\star \cdot \mathbf{x} = 0$ with probability one (i.e. the inherent noise is 0), then SGD is still not equivalent to gradient descent; here, it is the variance in SGD that contributes to this dependence in the EffectiveVariance. Our conjecture is that our lower bound can be improved to match that of our upper bound.

Relaxing the data distribution assumption. While our data distribution assumption (Assumption 2.2.2) can be satisfied if the whitened data is sub-Gaussian, it still cannot cover the simple one-hot case (i.e., $\mathbf{x} = \mathbf{e}_i$ with probability p_i, where $\sum_i p_i = 1$). Here, we conjecture that modifications of our proof can be used to establish the theoretical guarantees of SGD under the following relaxed assumption on the data distribution: assume that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{A} \mathbf{x}\mathbf{x}^\top] \leq a \operatorname{tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H} + b\|\mathbf{H}\|_2 \cdot \mathbf{H}^{1/2}\mathbf{A}\mathbf{H}^{1/2}$ for all PSD matrix $\mathbf{A}$ and some nonnegative constants a and b, which is weaker than Assumption 2.2.2 in the sense that we can allow $a = 0$; this assumption captures the case where $\mathbf{x}$ are standard basis vectors, with $a = 0$ and $b = 1$.

CHAPTER 3

Implicit Regularization of SGD for Linear Regression

3.1 Introduction

In this chapter, we seek to compare the generalization ability of SGD and ridge algorithms for *least square problems* instance-wisely. In particular, we follow the same setting in Chapter 2 and aim to estimate the optimal model parameters $\mathbf{w}^*$ that optimizes the *population risk*:

$$L(\mathbf{w}^*) = \min_{\mathbf{w}\in\mathcal{H}} L(\mathbf{w}), \quad \text{where } L(\mathbf{w}) := \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[(y - \langle \mathbf{w}, \mathbf{x}\rangle)^2\big]. \tag{3.1.1}$$

Additionally, we introduce the SGD algorithm and ridge regression solution in the following.

Constant-Stepsize SGD with Tail-Averaging. We consider the constant-stepsize SGD with tail-averaging [BM13; JKK18a; JNK17; ZWB21]: at the t-th iteration, a fresh example $(\mathbf{x}_t, y_t)$ is sampled independently from the data distribution, and SGD makes the following update on the current estimator $\mathbf{w}_{t-1} \in \mathcal{H}$,

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma \cdot \big(y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t\rangle\big)\mathbf{x}_t, \ t = 1, 2, \ldots, \qquad \mathbf{w}_0 = 0,$$

where $\gamma > 0$ is a constant stepsize. After N iterations (which is also the number of samples observed), SGD outputs the tail-averaged iterates as the final estimator:

$$\mathbf{w}_{\mathrm{sgd}}(N;\gamma) := \frac{2}{N}\sum_{t=N/2}^{N-1} \mathbf{w}_t.$$

In the underparameterized setting ($d < N$), constant-stepsize SGD with tail-averaging is known for achieving minimax optimal rate for least squares [JKK18a; JNK17]. More recently,

[ZWB21] investigate the performance of constant-stepsize SGD with tail-averaging in the overparameterized regime ($d > N$), and establish *instance-dependent*, nearly-optimal excess risk bounds under mild assumptions on the data distribution. Notably, results from [ZWB21] cover underparameterized cases ($d < N$) as well.

Ridge Regression. Given N i.i.d. samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, let us denote $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N\times d}$ and $\mathbf{y} := [y_1, \ldots, y_N]^\top \in \mathbb{R}^d$. Then ridge regression outputs the following estimator for the true parameter [Tih63]:

$$\mathbf{w}_{\mathrm{ridge}}(N; \lambda) := \arg\min_{\mathbf{w}\in\mathcal{H}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2, \tag{3.1.2}$$

where λ (which could possibly be negative) is a regularization parameter. We remark that the ridge regression estimator takes the following two equivalent form:

$$\mathbf{w}_{\mathrm{ridge}}(N; \lambda) = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N)^{-1}\mathbf{y}. \tag{3.1.3}$$

The first expression is useful in the classical, underparameterized setting ($d < N$) [HKZ12]; and the second expression is more useful in the overparameterized setting ($d > N$) where the empirical covariance $\mathbf{X}^\top\mathbf{X}$ is usually not invertible [KLS20; TB20]. As a final remark, when $\lambda = 0$, ridge estimator reduces to the *ordinary least square estimator* (OLS) [FHT01].

Contributions. Due to recent advances on sharp, *instance-dependent* excess risks bounds of both (single-pass) SGD and ridge regression for overparameterized least square problems [TB20; ZWB21], we provide a nearly complete answer to the question:

How does the generalization performance of SGD compare with that of ridge regression in least square problems?

In this work, we deliver an *instance-based* risk comparison between SGD and ridge regression in several interesting settings, including one-hot distributed data and Gaussian data. In particular, for a broad class of least squares problem instances that are natural in high-dimensional settings, we show that

- For every problem instance and for every ridge parameter, (unregularized) SGD, when provided with *logarithmically* more samples than that provided to ridge regularization, generalizes no worse than the ridge solution, provided SGD uses a tuned constant stepsize.

- Conversely, there exist instances in our problem class where optimally-tuned ridge regression requires *quadratically* more samples than SGD to achieve the same generalization performance.

Quite strikingly, the above results show that, up to some logarithmic factors, the generalization performance of SGD is always no worse than that of ridge regression in a wide range of overparameterized least square problems, and, in fact, could be much better for some problem instances. As a special case (for the above two claims), our problem class includes a setting in which: (i) the signal-to-noise is bounded and (ii) the eigenspectrum decays at a polynomial rate $1/i^{\alpha}$, for $0 \leq \alpha \leq 1$ (which permits a relatively fast decay). This one-sided near-domination phenomenon (in these natural overparameterized problem classes) could further support the preference for the implicit regularization brought by SGD over explicit ridge regularization.

Several novel technical contributions are made to make the above risk comparisons possible. For the one-hot data, we derive similar risk upper bound of SGD and risk lower bound of ridge regression. For the Gaussian data, while a sharp risk bound of SGD is borrowed from [ZWB21], we prove a sharp lower bound of ridge regression by adapting the proof techniques developed in [TB20; BLL20]. By carefully comparing these upper and lower bound results (and exhibiting particular instances to show that our sample size inflation bounds are sharp), we are able to provide nearly complete conditions that characterize when SGD generalizes better than ridge regression.

3.2 Preliminaries

We use $\mathbf{x} \in \mathcal{H}$ to denote a feature vector in a (separable) Hilbert space $\mathcal{H}$. We use d to refer to the dimensionality of $\mathcal{H}$, where $d = \infty$ if $\mathcal{H}$ is infinite-dimensional. We use $y \in \mathbb{R}$ to denote a response that is generated by

$$y = \langle \mathbf{x}, \mathbf{w}^* \rangle + \xi,$$

where $\mathbf{w}^* \in \mathcal{H}$ is an unknown true model parameter and $\xi \in \mathbb{R}$ is the model noise. The following regularity assumption is made throughout the paper.

Assumption 3.2.1 (Well-specified noise)**.** The second moment of $\mathbf{x}$, denoted by $\mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$, is strictly positive definite and has finite trace. The noise ξ is independent of $\mathbf{x}$ and satisfies

$$\mathbb{E}[\xi] = 0, \quad \text{and} \quad \mathbb{E}[\xi^2] = \sigma^2.$$

The least squares problem is to estimate the true parameter $\mathbf{w}^*$. Assumption 3.2.1 implies that $\mathbf{w}^*$ is the unique solution that minimizes the population risk $L(\mathbf{w})$. Moreover we have that $L(\mathbf{w}^*) = \sigma^2$. For an estimation $\mathbf{w}$ found by some algorithm, e.g., SGD or ridge regression, its performance is measured by the *excess risk*, $L(\mathbf{w}) - L(\mathbf{w}^*)$.

Generalizable Regime. In this chapter we will make instance-based risk comparisons between SGD and ridge regression. To make the comparison meaningful, we focus on regime where SGD and ridge regression are "generalizable", i.e, the SGD and the ridge regression estimators, with the optimally-tuned hypeparameters, can achieve excess risk that is smaller than the optimal population risk, i.e., σ^2. The formal mathematical definition is as follows.

Definition 3.2.2 (Generalizability)**.** Consider an algorithm `Alg` and a least squares problem instance `P`. Let $\texttt{Alg}(n, \boldsymbol{\theta})$ be the output of the algorithm when provided with n i.i.d. samples from the problem instance `P`, and a set of hyperparameters $\boldsymbol{\theta}$ (that could be a function on n).

Then we say that the algorithm $\texttt{Alg}$ with sample size n and hyperparameters configuration $\boldsymbol{\theta}$ is *generalizable* on problem instance $\texttt{P}$, if

$$\mathbb{E}_{\texttt{Alg},\texttt{P}}[L\big(\texttt{Alg}(n,\boldsymbol{\theta})\big)] - L(\mathbf{w}^*) \leq \sigma^2,$$

where the expectation is over the randomness of $\texttt{Alg}$ and data drawn from the problem instance $\texttt{P}$.

Clearly, the generalizable regime is defined by conditions on both the sample size, hyperparameter configuration, the problem instance, and the algorithm. For example, in the d-dimensional setting with $\|\mathbf{w}^*\|_2 = O(1)$, the ordinary least squares (OLS) solution (ridge regression with $\lambda = 0$), i.e., $\mathbf{w}_{\text{ridge}}(N; 0)$ has $\mathcal{O}(d\sigma^2/N)$ excess risk, then we can say that the ridge regression with regularization parameter $\lambda = 0$ and sample size $N = \omega(d)$ is in the generalizable regime on all problem instances in d-dimension with $\|\mathbf{w}^*\|_2 = O(1)$.

Sample Inflation vs. Risk Inflation Comparisons. This work characterizes the *sample inflation* of SGD, i.e., bounding the required sample size of SGD to achieve an instance-based comparable excess risk as ridge regression (which is essentially the notion of Bahadur statistical efficiency [Bah67; Bah71]). Another natural comparison would be examining the *risk inflation* of SGD, examining the instance-based increase in risk for any fixed sample size. Our preference for the former is due to the relative instability of the risk with respect to the sample size (in some cases, given a slightly different sample size, the risk could rapidly change.).

3.3 Warm-Up: One-Hot Least Squares Problems

Let us begin with a simpler data distribution, the *one-hot* data distribution. (inspired by settings where the input distribution is sparse). In detail, assume each input vector $\mathbf{x}$ is sampled from the set of natural basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d\}$ according to the data distribution

given by $\mathbb{P}\{\mathbf{x} = \mathbf{e}_i\} = \lambda_i$, where $0 < \lambda_i \leq 1$ and $\sum_i \lambda_i = 1$. The class of one-hot least square instances is completely characterized by the following problem set:

$$\big\{(\mathbf{w}^*; \lambda_1, \cdots, \lambda_d) : \ \mathbf{w}^* \in \mathcal{H}, \ \textstyle\sum_i \lambda_i = 1, \ 1 \geq \lambda_1 \geq \lambda_2 \geq \cdots > 0\big\}.$$

Clearly the population data covariance matrix is $\mathbf{H} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. The next two theorems give an instance-based sample inflation comparisons for this problem class.

Theorem 3.3.1 (Instance-wise comparison, one-hot data)**.** Let $\mathbf{w}_{\mathrm{sgd}}(N; \gamma)$ and $\mathbf{w}_{\mathrm{ridge}}(N; \lambda)$ be the solutions found by SGD and ridge regression when using N training examples. Then for any one-hot least square problem instance such that the ridge regression solution is generalizable and any λ, there exists a choice of stepsize γ^* for SGD such that

$$L\big[\mathbf{w}_{\mathrm{sgd}}(N_{\mathrm{sgd}}; \gamma^*)\big] - L(\mathbf{w}^*) \lesssim L\big[\mathbf{w}_{\mathrm{ridge}}(N_{\mathrm{ridge}}; \lambda)\big] - L(\mathbf{w}^*) < \sigma^2,$$

provided the sample size of SGD satisfies

$$N_{\mathrm{sgd}} \geq N_{\mathrm{ridge}}.$$

Theorem 3.3.1 suggests that for *every* one-hot problem instance, when provided with the same or more number of samples, the SGD solution with a properly tuned stepsize generalizes at most constant times worse than the optimally tuned ridge regression solution. In other words, with the same number of samples, SGD is *always* competitive with ridge regression.

Theorem 3.3.2 (Best-case comparison, one-hot data)**.** There exists an one-hot least square problem instance satisfying $\|\mathbf{w}^*\|_{\mathbf{H}}^2 = \sigma^2$, and a SGD solution with constant stepsize and sample size N_{sgd}, such that for any ridge regression solution with sample size

$$N_{\mathrm{ridge}} \leq \frac{N_{\mathrm{sgd}}^2}{\log^2(N_{\mathrm{sgd}})},$$

it holds that,

$$L\big[\mathbf{w}_{\mathrm{ridge}}(N_{\mathrm{ridge}}; \lambda)\big] - L(\mathbf{w}^*) \gtrsim L\big[\mathbf{w}_{\mathrm{sgd}}(N_{\mathrm{sgd}}; \gamma^*)\big] - L(\mathbf{w}^*).$$

Theorem 3.3.2 shows that for some one-hot least square instance, ridge regression, even with the optimally-tuned regularization, needs at least (nearly) quadratically more samples than that provided to SGD, in order to compete with the optimally-tuned SGD. In other words, ridge regression could be much worse than SGD for one-hot least squares problems.

Remark 3.3.3. The above two results together indicate a *superior* performance of the implicit regularization of SGD in comparison with the explicit regularization of ridge regression, for one-hot least squares problems. This is not the only case that SGD is always no worse than ridge estimator. In fact, we will next turn to compare SGD with ridge regression for the class of Gaussian least square instances, where both SGD and ridge regression exhibit richer behaviors but SGD still exhibits superiority over the ridge estimator.

3.4 Gaussian Least Squares Problems

In this section, we consider least squares problems with a Gaussian data distribution. In particular, assume the population distribution of the input vector $\mathbf{x}$ is Gaussian[1], i.e., $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{H})$. We further make the following regularity assumption for simplicity:

Assumption 3.4.1. $\mathbf{H}$ is strictly positive definite and has a finite trace.

Gaussian least squares problems are completely characterized by the following problem set $\big\{(\mathbf{w}^*; \mathbf{H}) : \ \mathbf{w}^* \in \mathcal{H}\big\}$.

The next theorem give an instance-based sample inflation comparison between SGD and ridge regression for Gaussian least squares instances.

Theorem 3.4.2 (Instance-wise comparison, Gaussian data)**.** Let $\mathbf{w}_{\text{sgd}}(N; \gamma)$ and $\mathbf{w}_{\text{ridge}}(N; \lambda)$ be the solutions found by SGD and ridge regression respectively. Then under Assumption

[1]We restrict ourselves to the Gaussian distribution for simplicity. Our results hold under more general assumptions, e.g., $\mathbf{H}^{-1/2}\mathbf{x}$ has sub-Gaussian tail and independent components [BLL20] and is symmetrically distributed.

3.4.1, for any Gaussian least square problem instance such that the ridge regression solution is generalizable and any λ, there exists a choice of stepsize γ^* for SGD such that

$$L\big[\mathbf{w}_{\text{sgd}}(N_{\text{sgd}};\gamma^*)\big] - L(\mathbf{w}^*) \lesssim L\big[\mathbf{w}_{\text{ridge}}(N_{\text{ridge}};\lambda)\big] - L(\mathbf{w}^*),$$

provided the sample size of SGD satisfies

$$N_{\text{sgd}} \geq (1+R^2)\cdot\kappa(N_{\text{ridge}})\cdot\log(a)\cdot N_{\text{ridge}},$$

where

$$\kappa(n) = \frac{\text{tr}(\mathbf{H})}{n\lambda_{\min\{n,d\}}}, \quad R^2 = \frac{\|\mathbf{w}^*\|_{\mathbf{H}}^2}{\sigma^2}, \quad a = \kappa(N_{\text{ridge}})R\sqrt{N}.$$

Note that the result in Theorem 3.4.2 holds for arbitrary λ. Then this theorem provides a sufficient condition for SGD such that it provably performs no worse than optimal ridge regression solution (i.e., ridge regression with optimal λ). Besides, we would also like to point out that the SGD stepsize γ^* in Theorem 3.4.2 is only a function of the regularization parameter λ and $\text{tr}(\mathbf{H})$, which can be easily estimated from training dataset without knowing the exact formula of $\mathbf{H}$.

Different from the one-hot case, here the required sample size for SGD depends on two important quantities: R^2 and $\kappa(N_{\text{ridge}})$. In particular, $R^2 = \|\mathbf{w}^*\|_{\mathbf{H}}^2/\sigma^2$ can be understood as the *signal-to-noise* ratio. The quantity $\kappa(N_{\text{ridge}})$ characterizes the flatness of the eigen-spectrum of $\mathbf{H}$ in the top N_{ridge}-dimensional subspace, which clearly satisfies $\kappa(N_{\text{ridge}}) \geq 1$. Let us further explain why we have the dependencies on R^2 and $\kappa(N_{\text{ridge}})$ in the condition of the sample inflation for SGD.

A large R^2 emphasizes the problem hardness is more from the numerical optimization instead of from the statistic learning. In particular, let us consider a special case where $\sigma = 0$ and $R^2 = \infty$, i.e., there is no noise in the least square problem, and thus solving it is purely a numerical optimization issue. In this case, ridge regression with $\lambda = 0$ achieves *zero* population risk so long as the observed data can span the whole parameter space, but constant stepsize SGD in general suffers a non-zero risk in finite steps, thus cannot be

competitive with the risk of ridge regression, which is as predicted by Theorem 3.4.2. From a learning perspective, a constant or even small R^2 is more interesting.

To explain why the dependency on $\kappa(N_{\text{ridge}})$ is unavoidable, we can consider a 2-d dimensional example where

$$\mathbf{H} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{N_{\text{ridge}} \cdot \kappa(N_{\text{ridge}})} \end{pmatrix}, \quad \mathbf{w}^* = \begin{pmatrix} 0 \\ N_{\text{ridge}} \cdot \kappa(N_{\text{ridge}}) \end{pmatrix}.$$

It is commonly known that for this problem, ridge regression with $\lambda = 0$ can achieve $\mathcal{O}(\sigma^2/N_{\text{ridge}})$ excess risk bound [FHT01]. However, this problem is rather difficult for SGD since it is hard to learn the second coordinate of $\mathbf{w}^*$ using gradient information (the gradient in the second coordinate is quite small). In fact, in order to accurately learn $\mathbf{w}^*[2]$, SGD requires at least $\Omega(1/\lambda_2) = \Omega\big(N_{\text{ridge}}\kappa(N_{\text{ridge}})\big)$ iterations/samples, which is consistent with our theory.

Then from Theorem 3.4.2 it can be observed that when the signal-to-noise ratio is nearly a constant, i.e., $R^2 = \Theta(1)$, and the eigenspectrum of $\mathbf{H}$ does not decay too fast so that $\kappa(N_{\text{ridge}}) \leq \text{polylog}(N_{\text{ridge}})$, SGD provably generalizes no worse than ridge regression, provided with logarithmically more samples than that provided to ridge regression. More specifically, the following corollary gives a family of problem instances that are in this regime.

Corollary 3.4.3. Under the same conditions as Theorem 3.4.2, let N_{ridge} be the sample size of ridge regression. Consider the problem instance that satisfies $R^2 = \Theta(1)$, $d = O(N_{\text{ridge}})$, and $\lambda_i = 1/i^\alpha$ for some $\alpha \leq 1$, then SGD, with a tuned stepsize γ^*, provably generalizes no worse than any ridge regression solution in the generalizable regime if

$$N_{\text{sgd}} \geq \log^2(N_{\text{ridge}}) \cdot N_{\text{ridge}}.$$

We would like to further point out that the comparison made in Corollary 3.4.3 concerns the worst-case result regarding $\mathbf{w}^*$ (from the perspective of SGD), while SGD could perform much better if $\mathbf{w}^*$ has a nice structure. For example, considering the same setting in Corollary

3.4.3 but assuming that the ground truth $\mathbf{w}^*$ is drawn from a prior distribution that is rotation invariant, SGD can be no worse than ridge regression provided the same or larger sample size. We formally state this result in the following corollary.

Corollary 3.4.4. Under the same conditions as Corollary 3.4.3, let N_{ridge} be the sample size of ridge regression. Consider the problem instance with random and rotation invariant $\mathbf{w}^*$, then SGD with a tuned stepsize γ^* provably generalizes no worse than any ridge regression solution in the generalizable regime if

$$N_{\text{sgd}} \geq N_{\text{ridge}}.$$

The next theorem shows that, in fact, for some instances, SGD could perform much better than ridge regression, as for the one-hot least square problems.

Theorem 3.4.5 (Best-case comparison, Gaussian data)**.** There exists a Gaussian least square problem instance satisfying $R^2 = 1$ and $\kappa(N_{\text{sgd}}) = \Theta(1)$, and an SGD solution with a constant stepsize and sample size N_{sgd}, such that for any ridge regression solution (i.e., any λ) with sample size

$$N_{\text{ridge}} \leq \frac{N_{\text{sgd}}^2}{\log^2(N_{\text{sgd}})},$$

it holds that,

$$L\big[\mathbf{w}_{\text{ridge}}(N_{\text{ridge}}; \lambda)\big] - L(\mathbf{w}^*) \gtrsim L\big[\mathbf{w}_{\text{sgd}}(N_{\text{sgd}}; \gamma^*)\big] - L(\mathbf{w}^*).$$

Besides the instance-wise comparison, it is also interesting to see under what condition SGD can provably outperform ridge regression, i.e., achieving comparable or smaller excess risk using the *same* number of samples. The following theorem shows that this occurs when the signal-to-noise ratio R^2 is a constant and there is only a small fraction of $\mathbf{w}^*$ living in the tail eigenspace of $\mathbf{H}$.

Theorem 3.4.6 (SGD outperforms ridge regression, Gaussian data)**.** Let N_{ridge} be sample size of ridge regression and $k^* = \min\left\{k : \lambda_k \leq \frac{\text{tr}(\mathbf{H})}{N_{\text{ridge}}\log(N_{\text{ridge}})}\right\}$, then if $R^2 = \Theta(1)$, and

$$\sum_{i=k^*+1}^{N_{\text{ridge}}} \lambda_i(\mathbf{w}^*[i])^2 \lesssim \frac{k^*\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N_{\text{ridge}}},$$

for any ridge regression solution that is generalizable and any λ, there exists a choice of stepsize γ^* for SGD such that

$$L\big[\mathbf{w}_{\text{sgd}}(N_{\text{sgd}};\gamma^*)\big] - L(\mathbf{w}^*) \lesssim L\big[\mathbf{w}_{\text{ridge}}(N_{\text{ridge}};\lambda)\big] - L(\mathbf{w}^*)$$

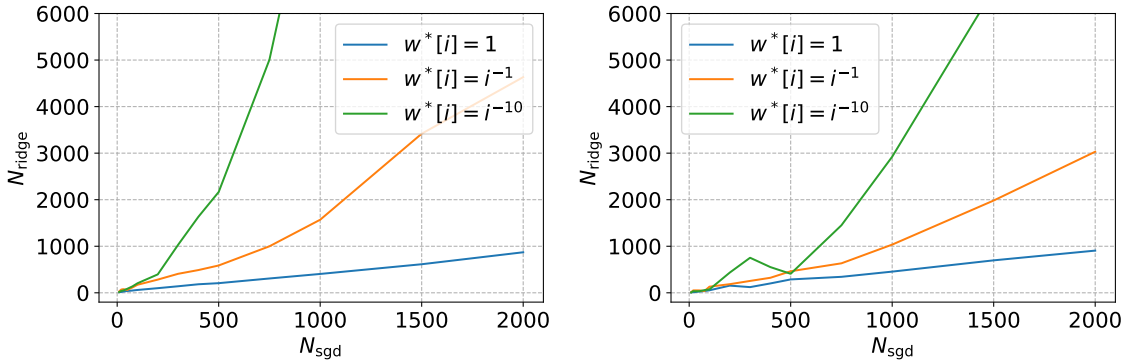
provided the sample size of SGD satisfies

$$N_{\text{sgd}} \geq N_{\text{ridge}}.$$

Experiments. We perform experiments on Gaussian least square problem. We consider 6 problem instances, which are the combinations of 2 different covariance matrices $\mathbf{H}$: $\lambda_i = i^{-1}$ and $\lambda_i = i^{-2}$; and 3 different true model parameter vectors $\mathbf{w}^*$: $\mathbf{w}^*[i] = 1$, $\mathbf{w}^*[i] = i^{-1}$, and $\mathbf{w}^*[i] = i^{-10}$. Figure 3.1 compares the required sample sizes of ridge regression and SGD that lead to the same population risk on these 6 problem instances, where the hyperparameters (i.e., γ and λ) are fine-tuned to achieve the best performance. We have two key observations: (1) in terms of the worst problem instance for SGD (i.e., $\mathbf{w}^*[i] = 1$), its sample size is only worse than ridge regression up to nearly constant factors (the curve is nearly linear); and (2) SGD can significantly outperform ridge regression when the true model $\mathbf{w}^*$ mainly lives in the head eigenspace of $\mathbf{H}$ (i.e., $\mathbf{w}^*[i] = i^{-10}$). The empirical observations are pretty consistent with our theoretical findings and again demonstrate the benefit of the implicit regularization of SGD.

3.5 An Overview of the Proof

In this section, we will sketch the proof of main Theorems for Gaussian least squares problems. Recall that we aim to show that provided certain number of training samples, SGD



(a) $\lambda_i = i^{-1}$ (b) $\lambda_i = i^{-2}$

Figure 3.1: Sample size comparison between SGD and ridge regression, where the stepsize γ and regularization parameter λ are fine-tuned to achieve the best performance. The problem dimension is $d = 200$ and the variance of model noise is $\sigma^2 = 1$. We consider 6 combinations of 2 different covariance matrices and 3 different ground truth model vectors. The plots are averaged over 20 independent runs.

is guaranteed to generalize better than ridge regression. Therefore, we will compare the risk *upper bound* of SGD [ZWB21] with the risk *lower bound* of ridge regression [TB20][2]. In particular, we first provide the following informal lemma summarizing the aforementioned risk bounds of SGD and ridge regression.

Lemma 3.5.1 (Risk bounds of SGD and ridge regression, informal)**.** Suppose Assumptions 3.2.1 and 3.4.1 hold and $\gamma \leq 1/\operatorname{tr}(\mathbf{H})$, then SGD has the following risk upper bound for

[2]The lower bound of ridge regression in our paper is a tighter variant of the lower bound in [TB20] since we consider Gaussian case and focus on the expected excess risk. [TB20] studied the sub-Gaussian case and established a high-probability risk bound.

arbitrary $k_1, k_2 \in [d]$,

$$\text{SGDRisk} \lesssim \underbrace{\frac{1}{\gamma^2 N_{\text{sgd}}^2} \cdot \left\| \exp(-N_{\text{sgd}}\gamma \mathbf{H})\mathbf{w}^* \right\|_{\mathbf{H}_{0:k_1}^{-1}}^2 + \left\|\mathbf{w}^*\right\|_{\mathbf{H}_{k_1:\infty}}^2}_{\text{SGDBiasBound}} + \underbrace{(1+R^2)\sigma^2 \cdot \left(\frac{k_2}{N_{\text{sgd}}} + N_{\text{sgd}}\gamma^2 \sum_{i>k_2} \lambda_i^2 \right)}_{\text{SGDVarianceBound}}. \tag{3.5.1}$$

Additionally, ridge regression has the following risk lower bound for a constant $\widetilde{\lambda}$, depending on λ, N_{ridge}, and $\mathbf{H}$, and $k^* = \min\{k : N_{\text{ridge}}\lambda_k \lesssim \widetilde{\lambda}\}$

$$\text{RidgeRisk} \gtrsim \underbrace{\left(\frac{\widetilde{\lambda}}{N_{\text{ridge}}} \right)^2 \left\|\mathbf{w}^*\right\|_{\mathbf{H}_{0:k^*}^{-1}}^2 + \left\|\mathbf{w}^*\right\|_{\mathbf{H}_{k^*:\infty}}^2}_{\text{RidgeBiasBound}} + \underbrace{\sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2} \sum_{i>k^*} \lambda_i^2 \right)}_{\text{RidgeVarianceBound}}. \tag{3.5.2}$$

We first highlight some useful observations in Lemma 3.5.1.

1. SGD has a condition on the stepsize: $\gamma \leq 1/\operatorname{tr}(\mathbf{H})$, while ridge regression has no condition on the regularization parameter λ.

2. Both the upper bound of SGD and the lower bound of ridge regression can be decomposed into two parts corresponding to the head and tail eigenspaces of $\mathbf{H}$. Furthermore, for the upper bound of SGD, the decomposition is arbitrary (k_1 and k_2 are arbitrary), while for the lower bound of ridge estimator, the decomposition is fixed (i.e., k^* is fixed).

3. Regarding the SGDBiasBound and SGDVarianceBound, performing the transformation $N \to \alpha N$ and $\gamma \to \alpha^{-1}\gamma$ will decrease SGDVarianceBound by a factor of α while the SGDBiasBound remains unchanged.

Based on the above useful observations, we can now interpret the proof sketch for Theorems 3.4.2, 3.4.5, and 3.4.6. We will first give the sketch for Theorem 3.4.6 and then prove Theorem 3.4.5 for the ease of presentation. We would like to emphasize that the calculation in the

proof sketch may not be the sharpest since they are presented for the ease of exposition. A preciser and sharper calculation can be found in Appendix.

Proof Sketch of Theorem 3.4.2. In order to perform instance-wise comparison, we need to take care of all possible $\mathbf{w}^* \in \mathcal{H}$. Therefore, by Observation 2, we can simply pick $k_1 = k_2 = k^*$ in the upper bound (3.5.1). Then it is clear that if setting $\gamma = \widetilde{\lambda}^{-1}$ and $N_{\mathrm{sgd}} = N_{\mathrm{ridge}}$, we have

$$\mathrm{SGDBiasBound} \leq \mathrm{RidgeBiasBound}$$

$$\mathrm{SGDVarianceBound} = (1 + R^2) \cdot \mathrm{RidgeVarianceBound}.$$

Then by Observation 3, enlarging N_{sgd} by $(1 + R^2)$ times suffices to guarantee

$$\mathrm{SGDBiasBound} + \mathrm{SGDVarianceBound} \leq \mathrm{RidgeBiasBound} + \mathrm{RidgeVarianceBound}.$$

On the other hand, according to Observation 1, there is an upper bound on the feasible stepsize of SGD: $\gamma \leq 1/\operatorname{tr}(\mathbf{H})$. Therefore, the above claim only holds when $\widetilde{\lambda} \geq \operatorname{tr}(\mathbf{H})$.

When $\widetilde{\lambda} \leq \operatorname{tr}(\mathbf{H})$, the stepsize $\widetilde{\lambda}^{-1}$ is no longer feasible and instead, we will use the largest possible stepsize: $\gamma = 1/\operatorname{tr}(\mathbf{H})$. Besides, note that we assume ridge regression solution is in the generalizable regime, then it holds that $k^* \leq N_{\mathrm{ridge}}$ since otherwise we have

$$\mathrm{RidgeRisk} \gtrsim \mathrm{RidgeVarianceBound} \geq \sigma^2.$$

Then again we set $k_1 = k_2 = k^*$ in SGDBiasBound and SGDVarianceBound. Applying the choice of stepsize $\gamma = 1/\operatorname{tr}(\mathbf{H})$ and sample size

$$N_{\mathrm{sgd}} = \frac{\log(R^2 N_{\mathrm{ridge}})}{\gamma \lambda_{k^*}} \leq N_{\mathrm{ridge}} \cdot \kappa(N_{\mathrm{ridge}}) \cdot \log(R^2 N_{\mathrm{ridge}}),$$

we get

$$\begin{aligned}
\mathrm{SGDBiasBound} &\leq \frac{(1 - N_{\mathrm{sgd}} \gamma \lambda_{k^*})^{N_{\mathrm{sgd}}}}{\gamma^2 N_{\mathrm{sgd}}^2 \lambda_{k^*}^2} \cdot \|\mathbf{w}^*\|_{\mathbf{H}_{0:k^*}}^2 + \|\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\
&\leq \frac{\sigma^2}{N_{\mathrm{ridge}}} + \|\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\
&\leq \mathrm{RidgeBiasBound} + \mathrm{RidgeVarianceBound}. \qquad (3.5.3)
\end{aligned}$$

Moreover, we can also get the following bound on SGDVarianceBound,

$$\begin{aligned}\text{SGDVarianceBound} &\leq (1+R^2)\sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{\log(R^2 N_{\text{ridge}})}{\lambda_{k^*}\operatorname{tr}(\mathbf{H})}\sum_{i>k^*}\lambda_i^2\right) \\ &\leq (1+R^2)\log(R^2 N_{\text{ridge}}) \cdot \text{RidgeVarianceBound},\end{aligned}$$

where in the second inequality we use the fact that

$$\frac{N_{\text{ridge}}}{\widetilde{\lambda}^2} \geq \frac{1}{\lambda_{k^*}\widetilde{\lambda}} \geq \frac{1}{\lambda_{k^*}\operatorname{tr}(\mathbf{H})}.$$

Therefore by Observation 3 again we can enlarge N_{sgd} properly to ensure that SGDVarianceBound remains unchanged and SGDVarianceBound $\leq$ RidgeVarianceBound. Then combining this and (3.5.3) we can get

$$\text{SGDBiasBound} + \text{SGDVarianceBound} \leq 2 \cdot \text{RidgeBiasBound} + 2 \cdot \text{RidgeVarianceBound},$$

which completes the proof.

Proof Sketch of Theorem 3.4.6. Now we will investigate in which regime SGD will generalizes no worse than ridge regression when provided with same training sample size. For simplicity in the proof we assume $R^2 = 1$. First note that we only need to deal with the case where $\widetilde{\lambda} \leq \operatorname{tr}(\mathbf{H})$ by the proof sketch of Theorem 3.4.2.

Unlike the instance-wise comparison that consider all possible $\mathbf{w}^* \in \mathcal{H}$, in this lemma we only consider the set of $\mathbf{w}^*$ that SGD performs well. Specifically, as we have shown in the proof of Theorem 3.4.2, in the worst-case comparison (in terms of $\mathbf{w}^*$), we require SGD to be able to learn the first k^* (where $k^* \leq N_{\text{ridge}}$) coordinates of $\mathbf{w}^*$ in order to be competitive with ridge regression, while SGD with sample size N_{sgd} can only be guaranteed to learn the first k^*_{sgd} coordinates of $\mathbf{w}^*$, where $k^*_{\text{sgd}} = \min\{k : N_{\text{ridge}}\lambda_k \leq \operatorname{tr}(\mathbf{H})\}$. Therefore, in the instance-wise comparison we need to enlarge N_{sgd} to $N_{\text{ridge}} \cdot \kappa(N_{\text{ridge}})$ to guarantee the learning of the top k^* coordinates of $\mathbf{w}^*$.

However, this is not required for some good $\mathbf{w}^*$'s that have small components in the k^*_{sgd}-k^* coordinates. In particular, as assumed in the theorem, we have $\sum_{i=\widehat{k}+1}^{N_{\text{ridge}}} \lambda_i(\mathbf{w}^*[i])^2 \leq$

$\widehat{k}\|\mathbf{w}^*\|_{\mathbf{H}}^2/N_{\text{ridge}}$, where $\widehat{k} := \min\{k : \lambda_k N_{\text{sgd}} \le \text{tr}(\mathbf{H})\cdot\log(N_{\text{sgd}})\}$ satisfies $\widehat{k} \le k^*_{\text{sgd}} \le k^*$. Then let $k_1 = \widehat{k}$ in SGDBiasBound, we have

$$\begin{aligned}
\text{SGDBiasBound} &= \frac{1}{\gamma^2 N_{\text{ridge}}^2}\cdot\left\|\exp(-N_{\text{ridge}}\gamma\mathbf{H})\mathbf{w}^*\right\|_{\mathbf{H}_{0:\widehat{k}}^{-1}}^2 + \left\|\mathbf{w}^*\right\|_{\mathbf{H}_{\widehat{k}:\infty}}^2 \\
&\le (1 - N_{\text{ridge}}\gamma\lambda_{\widehat{k}})^{N_{\text{ridge}}}\cdot\|\mathbf{w}^*\|_{\mathbf{H}_{0:k^*}}^2 + \|\mathbf{w}^*\|_{\mathbf{H}_{\widehat{k}:\infty}}^2 \\
&\overset{(i)}{\le} \frac{R^2\sigma^2(\widehat{k}+1)}{N_{\text{ridge}}} + \|\mathbf{w}^*\|_{\mathbf{H}_{k^*:\infty}}^2 \\
&\le 2\cdot\text{RidgeVarBound} + \text{RidgeBiasBound}.
\end{aligned}$$

where (i) is due to the condition that $\sum_{i=\widehat{k}+1}^{N_{\text{ridge}}}\lambda_i(\mathbf{w}^*[i])^2 \le \widehat{k}\|\mathbf{w}^*\|_{\mathbf{H}}^2/N_{\text{ridge}}$. Moreover, it is easy to see that given $N_{\text{sgd}} = N_{\text{ridge}}$ and $\gamma = 1/\text{tr}(\mathbf{H}) \le 1/\widetilde{\lambda}$, we have SGDVarianceBound $\le$ $2\cdot$ RidgeVarianceBound. As a consequence we can get

$$\text{SGDBiasBound} + \text{SGDVarianceBound} \le 3\cdot\text{RidgeBiasBound} + 3\cdot\text{RidgeVarianceBound}.$$

Proof Sketch of Theorem 3.4.5. We will consider the best $\mathbf{w}^*$ for SGD, which only has nonzero entry in the first coordinate. For example, consider a true model parameter vector with $\mathbf{w}^*[1] = 1$ and $\mathbf{w}^*[i] = 0$ for $i \ge 2$ and a problem instance whose spectrum of $\mathbf{H}$ has a flat tail with $\sum_{i\ge N_{\text{ridge}}}\lambda_i^2 = \Theta(1)$ and $\sum_{i\ge 2}\lambda_i^2 = \Theta(1)$. Then according to Lemma 3.5.1, we can set the stepsize as $\gamma = \Theta(\log(N_{\text{sgd}})/N_{\text{sgd}})$ and get

$$\begin{aligned}
\text{SGDRisk} &\lesssim \text{SGDBiasBound} + \text{SGDVarianceBound} \\
&= O\bigg(\frac{1}{N_{\text{sgd}}} + \frac{\log^2(N_{\text{sgd}})}{N_{\text{sgd}}}\bigg) = O\bigg(\frac{\log^2(N_{\text{sgd}})}{N_{\text{sgd}}}\bigg).
\end{aligned}$$

For ridge regression, according to Lemma 3.5.1 we have

$$\begin{aligned}
\text{RidgeRisk} &\gtrsim \text{RidgeBiasBound} + \text{RidgeVarianceBound} \\
&= \Omega\bigg(\frac{\widetilde{\lambda}^2}{N_{\text{ridge}}^2} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2}\bigg) \qquad \text{since } \sum_{i\ge k^*}\lambda_i^2 = \Theta(1) \\
&= \Omega\bigg(\frac{1}{N_{\text{ridge}}^{1/2}}\bigg). \qquad \text{by the fact that } a+b \ge \sqrt{ab}
\end{aligned}$$

Therefore, it is evident that ridge regression is guaranteed to be worse than SGD if $N_{\text{ridge}} \leq N_{\text{sgd}}^2 / \log^2(N_{\text{sgd}})$. This completes the proof.

3.6 Proof of One-hot Least Squares

3.6.1 Excess risk bound of SGD

In this part we will mainly follow the proof technique in [ZWB21] that is developed to sharply characterize the excess risk bound for SGD (with tail-averaging) when the data distribution has a nice finite fourth-moment bound. However, such condition does not hold for the one-hot case so that their results cannot be directly applied here.

Before presenting the detailed proofs, we first introduce some notations and definitions that will be repeatedly used in the subsequent analysis. Let $\mathbf{H} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ be the covariance of data distribution. It is easy to verify that $\mathbf{H}$ is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_d$. Let $\mathbf{w}_t$ be the t-th iterate of the SGD, we define $\boldsymbol{\beta}_t := \mathbf{w}_t - \mathbf{w}^*$ as the centered SGD iterate. Then we define $\boldsymbol{\beta}_t^{\text{bias}}$ and $\boldsymbol{\beta}_t^{\text{variance}}$ as the bias error and variance error respectively, which are described by the following update rule:

$$\begin{aligned} \boldsymbol{\beta}_t^{\text{bias}} &= \big(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top\big)\boldsymbol{\beta}_{t-1}^{\text{bias}}, \qquad \boldsymbol{\beta}_0^{\text{bias}} = \boldsymbol{\beta}_0, \\ \boldsymbol{\beta}_t^{\text{variance}} &= \big(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top\big)\boldsymbol{\beta}_{t-1}^{\text{bias}} + \gamma \xi_t \mathbf{x}_t, \qquad \boldsymbol{\beta}_0^{\text{variance}} = \mathbf{0}. \end{aligned} \tag{3.6.1}$$

Accordingly, we can further define the bias covariance $\mathbf{B}_t$ and variance covariance $\mathbf{C}_t$ as follows

$$\mathbf{B}_t = \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{bias}}], \qquad \mathbf{C}_t = \mathbb{E}[\boldsymbol{\beta}_t^{\text{variance}} \otimes \boldsymbol{\beta}_t^{\text{variance}}].$$

Regarding these two covariance matrices, the following lemma mathematically characterizes the upper bounds of the diagonal entries of $\mathbf{B}_t$ and $\mathbf{C}_t$.

Lemma 3.6.1. Under Assumptions 3.2.1, let $\bar{\mathbf{B}}_t = \text{diag}(\mathbf{B}_t)$ and $\bar{\mathbf{C}}_t = \text{diag}(\mathbf{C}_t)$, then if the

stepsize satisfies $\gamma \leq 1$, we have

$$\bar{\mathbf{B}}_t \preceq (\mathbf{I} - \gamma \mathbf{H})\bar{\mathbf{B}}_{t-1}, \qquad \bar{\mathbf{C}}_t \preceq (\mathbf{I} - \gamma \mathbf{H})\bar{\mathbf{C}}_{t-1} + \gamma^2 \sigma^2 \mathbf{H}.$$

Proof. According to (3.6.1), we have

$$\begin{aligned}\mathbf{B}_t = \mathbb{E}[\boldsymbol{\beta}_t^{\text{bias}} \otimes \boldsymbol{\beta}_t^{\text{bias}}] &= \mathbb{E}\big[(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\beta}_{t-1}^{\text{bias}} \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\beta}_{t-1}^{\text{bias}}\big] \\ &= \mathbf{B}_{t-1} - \gamma \mathbf{H}\mathbf{B}_{t-1} - \gamma \mathbf{B}_{t-1}\mathbf{H} + \gamma^2 \mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top \mathbf{B}_{t-1} \mathbf{x}_t \mathbf{x}_t^\top]. \end{aligned} \tag{3.6.2}$$

Note that $\mathbf{x}_t = \mathbf{e}_i$ with probability λ_i, then we have

$$\begin{aligned}\mathbb{E}[\mathbf{x}_t \mathbf{x}_t^\top \mathbf{B}_{t-1} \mathbf{x}_t \mathbf{x}_t^\top] &= \sum_i \lambda_i \cdot \mathbf{e}_i \mathbf{e}_i^\top \mathbf{B}_{t-1} \mathbf{e}_i \mathbf{e}_i^\top \\ &= \sum_i \lambda_i \cdot \mathbf{e}_i^\top \mathbf{B}_{t-1} \mathbf{e}_i \cdot \mathbf{e}_i \mathbf{e}_i^\top \\ &= \bar{\mathbf{B}}_{t-1}\mathbf{H}.\end{aligned}$$

Plugging the above equation into (3.6.2) gives

$$\mathbf{B}_t = \mathbf{B}_{t-1} - \gamma \mathbf{H}\mathbf{B}_{t-1} - \gamma \mathbf{B}_{t-1}\mathbf{H} + \gamma^2 \bar{\mathbf{B}}_{t-1}\mathbf{H}.$$

Then if only look at the diagonal entries of both sides, we have

$$\bar{\mathbf{B}}_t = \bar{\mathbf{B}}_{t-1} - 2\gamma \mathbf{H}\bar{\mathbf{B}}_{t-1} + \gamma^2 \mathbf{H}\bar{\mathbf{B}}_{t-1} \preceq (\mathbf{I} - \gamma \mathbf{H})\bar{\mathbf{B}}_{t-1},$$

where in the first equation we use the fact that $\operatorname{diag}(\mathbf{HB}) = \operatorname{diag}(\mathbf{BH}) = \mathbf{H}\bar{\mathbf{B}}$ and the inequality follows from the fact that both $\bar{\mathbf{B}}_t$ and $\mathbf{H}$ are diagonal and $\gamma \leq 1$.

Similarly, regarding $\mathbf{C}_t$ the following holds according to (3.6.1),

$$\mathbf{C}_t = \mathbb{E}\big[(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\beta}_{t-1}^{\text{variance}} \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\beta}_{t-1}^{\text{variance}}\big] + \gamma^2 \mathbb{E}[\xi_t^2 \mathbf{x}_t \mathbf{x}_t^\top],$$

where we use the fact that $\mathbb{E}[\xi_t | \mathbf{x}_t] = 0$. Similar to deriving the bound for $\bar{\mathbf{B}}_t$, we have

$$\operatorname{diag}\big(\mathbb{E}\big[(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\beta}_t^{\text{variance}} \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\beta}_t^{\text{variance}}\big]\big) \preceq (\mathbf{I} - \gamma \mathbf{H})\bar{\mathbf{C}}_{t-1}.$$

Besides, under Assumption 3.2.1 we also have $\mathbb{E}[\xi_t^2 \mathbf{x}_t \mathbf{x}_t^\top] = \sigma^2 \mathbf{H}$, which is a diagonal matrix. Based on these two results, we can get the following upper bound for $\bar{\mathbf{C}}_t$,

$$\begin{aligned}\bar{\mathbf{C}}_t &= \operatorname{diag}\big(\mathbb{E}\big[(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\beta}_{t-1}^{\text{variance}} \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\beta}_{t-1}^{\text{variance}}\big] + \gamma^2 \mathbb{E}[\xi_t^2 \mathbf{x}_t \mathbf{x}_t^\top]\big) \\ &\preceq (\mathbf{I} - \gamma \mathbf{H})\bar{\mathbf{C}}_{t-1} + \gamma^2 \sigma^2 \mathbf{H}.\end{aligned}$$

This completes the proof. □

Lemma 3.6.2 (Lemmas D.1 & D.2 in [ZWB21])**.** Let $\bar{\mathbf{w}}_{N:2N}$ be the output of tail-averaged SGD, then if the stepsize satisfied $\gamma \leq 1/\lambda_1$, it holds that

$$\mathbb{E}[L(\bar{\mathbf{w}}_{N:2N})] - L(\mathbf{w}^*) \lesssim \text{SGDBias} + \text{SGDVariance},$$

where

$$\begin{aligned}\text{SGDBias} &\leq \frac{1}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{B}_{N+t} \big\rangle \\ \text{SGDVariance} &\leq \frac{1}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \big\langle (\mathbf{I} - \gamma \mathbf{H})^{k-t} \mathbf{H}, \mathbf{C}_{N+t} \big\rangle\end{aligned}$$

Lemma 3.6.3. Under Assumptions 3.2.1, if the stepsize satisfies $\gamma \leq 1$ and set $\mathbf{w}_0 = \mathbf{0}$, then

$$\mathbb{E}[L(\bar{\mathbf{w}}_{N:2N})] - L(\mathbf{w}^*) \leq 2 \cdot \text{bias} + 2 \cdot \text{variance},$$

where

$$\begin{aligned}\text{bias} &\lesssim \frac{1}{N^2 \gamma^2} \cdot \big\| (\mathbf{I} - \gamma \mathbf{H})^{N/2} \mathbf{w}^* \big\|_{\mathbf{H}_{0:k_1}^{-1}} + \big\| (\mathbf{I} - \gamma \mathbf{H})^{N/2} \mathbf{w}^* \big\|_{\mathbf{H}_{k_1:\infty}}^2 \\ \text{variance} &\lesssim \sigma^2 \cdot \left(\frac{k_2}{N} + N \gamma^2 \sum_{i > k_2} \lambda_i^2 \right)\end{aligned}$$

for arbitrary $k_1, k_2 \in [d]$.

Proof. The first conclusion of this theorem can be directly proved via Young's inequality.

Note that $\mathbf{H}$ is a diagonal matrix, and thus $(\mathbf{I}-\gamma\mathbf{H})^{k-t}$ is also a diagonal matrix for all k and t. Therefore, by Lemma 3.6.2, it is clear that in order to calculate the upper bound of the bias and variance error, it suffices to consider the diagonal entries of $\mathbf{B}_{N+t}$ and $\mathbf{C}_{N+t}$, denoted by $\bar{\mathbf{B}}_{N+t}$ and $\bar{\mathbf{C}}_{N+t}$ (which are obtained by setting all non-diagonal entries of $\mathbf{B}_{N+t}$ and $\mathbf{C}_{N+t}$ as zero). Then by Young's inequality, Lemma 3.6.2 implies that

$$\begin{aligned}\text{bias} &\le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H},\bar{\mathbf{B}}_{N+t}\big\rangle\\ \text{variance} &\le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\big\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H},\bar{\mathbf{C}}_{N+t}\big\rangle. \end{aligned}\tag{3.6.3}$$

Now we are ready to precisely calculate the above two bounds. In particular, by Lemma 3.6.1 we have

$$\bar{\mathbf{B}}_t \preceq (\mathbf{I}-\gamma\mathbf{H})\bar{\mathbf{B}}_{t-1} \preceq (\mathbf{I}-\gamma\mathbf{H})^t\mathbf{B}_0, \tag{3.6.4}$$

$$\bar{\mathbf{C}}_t \preceq (\mathbf{I}-\gamma\mathbf{H})\bar{\mathbf{C}}_{t-1} \preceq \sum_{s=0}^{t-1}\sigma^2\gamma^2(\mathbf{I}-\gamma\mathbf{H})^s\mathbf{H} = \sigma^2\gamma\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^t\big), \tag{3.6.5}$$

where in the second inequality we use the fact that $\mathbf{C}_0 = \boldsymbol{\beta}_t^{\text{variance}}\otimes\boldsymbol{\beta}_t^{\text{variance}} = \mathbf{0}$. Then plugging (3.6.4) into (3.6.3) gives

$$\begin{aligned}\text{bias} &\le \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\langle(\mathbf{I}-\gamma\mathbf{H})^{k-t}\mathbf{H},(\mathbf{I}-\gamma\mathbf{H})^{N+t}\mathbf{B}_0\rangle\\ &= \frac{1}{N^2}\Big\langle\sum_{k=0}^{N-1-t}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{H},\sum_{t=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{N+t}\mathbf{B}_0\Big\rangle\\ &\le \frac{1}{N^2}\Big\langle\sum_{k=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^k\mathbf{H},\sum_{t=0}^{N-1}(\mathbf{I}-\gamma\mathbf{H})^{N+t}\mathbf{B}_0\Big\rangle\\ &= \frac{1}{N^2\gamma^2}\Big\langle\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N,\mathbf{H}^{-1}(\mathbf{I}-\gamma\mathbf{H})^N\big(\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N\big)\mathbf{B}_0\Big\rangle\\ &= \frac{1}{N^2\gamma^2}\Big\langle(\mathbf{I}-\gamma\mathbf{H})^N\big[\mathbf{I}-(\mathbf{I}-\gamma\mathbf{H})^N\big]^2\mathbf{H}^{-1},\mathbf{B}_0\Big\rangle \end{aligned}\tag{3.6.6}$$

Note that $(1-x)^N \ge \min\{0,1-Nx\}$ for all $x\in[0,1]$. Then for all i we have

$$\big[1-(1-\gamma\lambda_i)^N\big]^2\lambda^{-1} \le \min\left\{\frac{1}{\lambda_i}, N^2\gamma^2\lambda_i\right\}$$

where we use the fact that $\gamma \leq 1 \leq 1/\lambda_i$ for all i. This further implies that

$$\left[\mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N\right]^2 \mathbf{H}^{-1} \preceq \mathbf{H}_{0:k}^{-1} + N^2\gamma^2 \mathbf{H}_{k:\infty}$$

for all $k \in [d]$. Plugging the above results into (3.6.6) leads to

$$\text{bias} \leq \frac{1}{N^2\gamma^2} \cdot \left\langle \mathbf{H}_{0:k}^{-1}, (\mathbf{I} - \gamma\mathbf{H})^N \mathbf{B}_0 \right\rangle + \left\langle \mathbf{H}_{k:\infty}, (\mathbf{I} - \gamma\mathbf{H})^N \mathbf{B}_0 \right\rangle \tag{3.6.7}$$

for all $k \in [d]$. Further note that $\mathbf{B}_0 = (\mathbf{w}_0 - \mathbf{w}^*) \otimes (\mathbf{w}_0 - \mathbf{w}^*) = \mathbf{w}^* \otimes \mathbf{w}^*$ as we pick $\mathbf{w}_0 = \mathbf{0}$. Thus (3.6.7) implies that

$$\text{bias} \leq \frac{1}{N^2\gamma^2} \cdot \left\| (\mathbf{I} - \gamma\mathbf{H})^{N/2} \mathbf{w}^* \right\|_{\mathbf{H}_{0:k}^{-1}} + \left\| (\mathbf{I} - \gamma\mathbf{H})^{N/2} \mathbf{w}^* \right\|_{\mathbf{H}_{k:\infty}}^2.$$

Then we will deal with the variance error. Plugging (3.6.5) into (3.6.3) gives

$$\begin{aligned}
\text{variance} &\leq \frac{\sigma^2\gamma}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \langle (\mathbf{I} - \gamma\mathbf{H})^{k-t} \mathbf{H}, \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N+t} \rangle \\
&\leq \frac{\sigma^2\gamma}{N^2} \sum_{t=0}^{N-1} \left\langle \sum_{k=0}^{N-1} (\mathbf{I} - \gamma\mathbf{H})^k \mathbf{H}, \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N+t} \right\rangle \\
&= \frac{\sigma^2}{N^2} \sum_{t=0}^{N-1} \left\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^N, \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{N+t} \right\rangle \\
&\leq \frac{\sigma^2}{N} \Big\langle \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{2N}, \mathbf{I} - (\mathbf{I} - \gamma\mathbf{H})^{2N} \Big\rangle.
\end{aligned}$$

We then use the inequality $(1-x)^N \geq \min\{0, 1 - xN\}$ again and thus the above inequality further leads to

$$\begin{aligned}
\text{variance} &\leq \frac{\sigma^2}{N} \cdot \sum_i \min\{1, 4N^2\gamma^2\lambda_i^2\} \\
&\leq \frac{4\sigma^2}{N} \cdot \left(k + N^2\gamma^2 \sum_{i>k} \lambda_i^2 \right)
\end{aligned}$$

for any $k \in [d]$. □

3.6.2 Excess risk bound of ridge regression

Lemma 3.6.4. Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be the training data matrix and $\mathbf{w}_{\text{ridge}}(N;\lambda)$ be the solution of ridge regression with parameter λ and sample size N, then for any $\lambda > 0$

$$\mathbb{E}[L(\mathbf{w}_{\text{ridge}}(N;\lambda))] - L(\mathbf{w}^*) = \text{bias} + \text{variance},$$

where

$$\begin{aligned}
\text{bias} &= \lambda^2 \cdot \mathbb{E}\big[\mathbf{w}^{*\top}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{w}^*\big] \\
\text{variance} &= \sigma^2 \cdot \mathbb{E}\big[\operatorname{tr}\big((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{H}\big)\big],
\end{aligned}$$

where the expectations are taken over the randomness of the training data matrix $\mathbf{X}$.

Proof. Recall that the solution of ridge regression takes form

$$\mathbf{w}_{\text{ridge}}(N;\lambda) = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y},$$

where $\mathbf{X}$ is the data matrix and $\mathbf{y}$ is the response vector. Then according to the definition of the loss function $L(\mathbf{w})$, we have

$$\begin{aligned}
\mathbb{E}[L(\mathbf{w}_{\text{ridge}}(N;\lambda))] &= \mathbb{E}\Big[\big(y - \langle\mathbf{w}_{\text{ridge}}(N;\lambda), \mathbf{x}\rangle\big)^2\Big] \\
&= \mathbb{E}\Big[\big(\langle\mathbf{w}^*, \mathbf{x}\rangle - \langle\mathbf{w}_{\text{ridge}}(N;\lambda), \mathbf{x}\rangle\big)^2\Big] + \mathbb{E}\Big[\big(y - \langle\mathbf{w}^*, \mathbf{x}\rangle\big)^2\Big] \\
&\quad + 2\mathbb{E}\big[\big(\langle\mathbf{w}^*, \mathbf{x}\rangle - \langle\mathbf{w}_{\text{ridge}}(N;\lambda), \mathbf{x}\rangle\big) \cdot \big(y - \langle\mathbf{w}^*, \mathbf{x}\rangle\big)\big] \\
&= \mathbb{E}[\|\mathbf{w}_{\text{ridge}}(N;\lambda) - \mathbf{w}^*\|_{\mathbf{H}}^2] + L(\mathbf{w}^*),
\end{aligned}$$

where the last equation is by Assumption 3.2.1. Then regarding $\mathbb{E}[\|\mathbf{w}_{\text{ridge}}(N;\lambda) - \mathbf{w}^*\|_{\mathbf{H}}^2]$, let $\boldsymbol{\xi} = \mathbf{y} - \mathbf{X}\mathbf{w}^*$ be the model noise vector, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_{\text{ridge}}(N;\lambda) - \mathbf{w}^*\|_{\mathbf{H}}^2] &= \mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} - \mathbf{w}^*\big\|_{\mathbf{H}}^2\big] \\
&= \mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}^* + \boldsymbol{\xi}) - \mathbf{w}^*\big\|_{\mathbf{H}}^2\big] \\
&= \underbrace{\mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w}^* - \mathbf{w}^*\big\|_{\mathbf{H}}^2\big]}_{\text{bias}} + \underbrace{\mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\xi}\big\|_{\mathbf{H}}^2\big]}_{\text{variance}}.
\end{aligned}$$

where in the last inequality we again apply Assumption 3.2.1 that $\mathbb{E}[\boldsymbol{\xi}|\mathbf{X}] = \mathbf{0}$. More specifically, the bias error can be reformulated as

$$\begin{aligned}\text{bias} &= \mathbb{E}\big[\big\|\big((\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}-\mathbf{I}\big)\mathbf{w}^*\big\|_{\mathbf{H}}^2\big] \\ &= \lambda^2\mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{w}\big\|_{\mathbf{H}}^2\big] \\ &= \lambda^2\mathbb{E}\big[\mathbf{w}^{*\top}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{w}^*\big].\end{aligned}$$

In terms of the variance error, note that by Assumption 3.2.1 we have $\mathbb{E}[\boldsymbol{\xi}\boldsymbol{\xi}^\top|\mathbf{X}] = \sigma^2\mathbf{I}$, then

$$\begin{aligned}\text{variance} &= \mathbb{E}\big[\big\|(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon}\big\|_{\mathbf{H}}^2\big] \\ &= \mathbb{E}\big[\,\mathrm{tr}\,\big((\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{\xi}\boldsymbol{\xi}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{H}\big)\big] \\ &= \sigma^2\cdot\mathbb{E}\big[\,\mathrm{tr}\,\big((\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{H}\big)\big].\end{aligned}$$

□

Lemma 3.6.5. The solution of ridge regression with sample size N and regularization parameter λ satisfies

$$\mathbb{E}[L(\mathbf{w}_{\text{ridge}}(N;\lambda))] - L(\mathbf{w}^*) = \text{RidgeBias} + \text{RidgeVariance},$$

where

$$\begin{aligned}\text{RidgeBias} &\gtrsim \max\left\{\sum_i (1-\lambda_i)^N\cdot\lambda_i\mathbf{w}^*[i]^2, \sum_{i=1}^{k^*}\frac{\lambda^2\lambda_i\mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*}\lambda_i\mathbf{w}^*[i]^2\right\} \\ \text{RidgeVariance} &\gtrsim \sigma^2\cdot\left(\sum_{i=1}^{k^*}\frac{N\lambda_i^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*}\frac{N\lambda_i^2}{(1+\lambda)^2}\right),\end{aligned}$$

where $k^* = \min\{k : N\lambda_k \le 1\}$.

Proof. In the one-hot case, it is easy to verify that $\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ is a diagonal matrix. Let $\mu_1, \mu_2, \ldots, \mu_d$ be the eigenvalues of $\mathbf{X}^\top\mathbf{X}$ corresponding to the eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d$ respectively. Then by Lemma 3.6.4, we have the following results for the bias and variance

errors of ridge regression.

$$\begin{aligned}\text{RidgeBias} &= \lambda^2 \cdot \mathbb{E}\big[\mathbf{w}^{*\top}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{H}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{w}^*\big] \\ &= \lambda^2 \sum_i \mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mathbf{w}^*[i]^2}{(\mu_i+\lambda)^2}\right], \end{aligned} \tag{3.6.8}$$

where the expectation in the first equation is taken over the training data $\mathbf{X}$ and in the second inequality the expectation is equivalently taken over the eigenvalues $\mu_1, \ldots, \mu_d$. Since $\mathbf{x}_i$ can only take on natural basis, the eigenvalue μ_i can be understood as the number of training data that equals $\mathbf{e}_i$. Note that the probability of sampling $\mathbf{e}_i$ is λ_i, then we can get that μ_i has a marginal distribution $\text{Binom}(N, \lambda_i)$, where N is the sample size. Then in terms of each expectation in (3.6.8), we first have

$$\mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mathbf{w}^*[i]^2}{(\mu_i+\lambda)^2}\right] \geq \frac{\lambda_i \mathbf{w}^*[i]^2}{(\mathbb{E}[\mu_i]+\lambda)^2} = \frac{\lambda_i \mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2},$$

where the first inequality is by applying Jensen's inequality to the convex function $f(x) = 1/(x+\lambda)^2$. On the other hand, we also have

$$\mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mathbf{w}^*[i]^2}{(\mu_i+\lambda)^2}\right] \geq \frac{\lambda_i \mathbf{w}^*[i]^2}{\lambda^2} \cdot \mathbb{P}(\mu_i = 0) = \frac{\lambda_i \mathbf{w}^*[i]^2}{\lambda^2} \cdot (1-\lambda_i)^N.$$

Therefore, combining the above two lower bounds, we can get the following lower bound on the bias error by (3.6.8)

$$\text{RidgeBias} = \lambda^2 \sum_i \mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mathbf{w}^*[i]^2}{(\mu_i+\lambda)^2}\right] \geq \sum_i \max\left\{\frac{\lambda^2 \lambda_i \mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2}, \lambda_i \mathbf{w}^*[i]^2 \cdot (1-\lambda_i)^N\right\}. \tag{3.6.9}$$

Therefore, a trivial lower bound on the bias error of ridge regression is

$$\text{RidgeBias} \geq \sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2.$$

Additionally, note that $(1-\lambda_i)^N \geq 0.25$ if $\lambda_i \leq 1/N$ and $N \geq 2$. Then let $k^* = \min\{k : N\lambda_k \leq 1\}$, (3.6.9) further leads to

$$\text{RidgeBias} \geq \sum_{i=1}^{k^*} \frac{\lambda^2 \lambda_i \mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2} + 0.25 \cdot \sum_{i>k^*} \lambda_i \mathbf{w}^*[i]^2.$$

This completes the proof of the lower bound of the bias error.

By Lemma 3.6.4, we have

$$\begin{aligned}\text{RidgeVariance} &= \sigma^2 \cdot \mathbb{E}\big[\operatorname{tr}\big((\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{H}\big)\big] \\ &= \sigma^2 \cdot \sum_i \mathbb{E}_{\mu_i}\bigg[\frac{\lambda_i\mu_i}{(\mu_i+\lambda)^2}\bigg], \end{aligned} \tag{3.6.10}$$

Regarding the variance error, we cannot use the similar approach since the function $g(x) = x/(x+\lambda)^2$ is no longer convex. Instead, we will directly make use of property of the binomial distribution of μ_i to prove the desired bound. In particular, note that $\mu_i \sim \text{binom}(N, \lambda_i)$, by Bernstein inequality, we have

$$\mathbb{P}(|\mu_i - N\lambda_i| \le t) \ge 1 - 2\exp\bigg(-\frac{t^2}{2(N\lambda_i + t/3)}\bigg).$$

If $N\lambda_i \ge 6$, by set $t = \sqrt{3N\lambda_i}$, we have

$$\mathbb{P}\big(\mu_i \in \big[N\lambda_i - \sqrt{3N\lambda_i}, N\lambda_i + \sqrt{3N\lambda_i}\big]\big) \ge 1 - 2e^{-1} \ge 0.2,$$

which further implies that

$$\mathbb{P}\big(\mu_i \in \big[0.25N\lambda_i, 2N\lambda_i\big]\big) \ge 0.2,$$

where we use the fact that $\sqrt{3N\lambda_i} \le 0.75N\lambda_i$ if $N\lambda_i > 6$. Therefore, in this case, we can get

$$\mathbb{E}_{\mu_i}\bigg[\frac{\lambda_i\mu_i}{(\mu_i+\lambda)^2}\bigg] \ge 0.2\min\bigg\{\frac{0.25N\lambda_i^2}{(0.25N\lambda_i+\lambda)^2}, \frac{2N\lambda_i^2}{(2N\lambda_i+\lambda)^2}\bigg\} \ge \frac{0.05N\lambda_i^2}{(N\lambda_i+\lambda)^2}. \tag{3.6.11}$$

Then we consider the case that $N\lambda_i < 6$. In particular, we have

$$\mathbb{E}_{\mu_i}\bigg[\frac{\lambda_i\mu_i}{(\mu_i+\lambda)^2}\bigg] \ge \frac{\lambda_i}{(1+\lambda)^2} \cdot \mathbb{P}(\mu_i = 1). \tag{3.6.12}$$

Note that μ_i follows $\text{Binom}(N, \lambda_i)$ distribution, which implies that

$$\mathbb{P}(\mu_i = 1) = N\lambda_i(1-\lambda_i)^{N-1} \ge N\lambda_i\big(1 - \frac{6}{N}\big)^{N-1} \ge e^{-6}N\lambda_i.$$

Plugging this into (3.6.12) gives

$$\mathbb{E}_{\mu_i}\bigg[\frac{\lambda_i\mu_i}{(\mu_i+\lambda)^2}\bigg] \ge \frac{e^{-6}N\lambda_i^2}{(1+\lambda)^2}. \tag{3.6.13}$$

Therefore, let $k^* = \min\{k : N\lambda_k \leq 1\}$, then for all $i \leq k^*$, combining (3.6.11) and (3.6.13) gives

$$\mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mu_i}{(\mu_i + \lambda)^2}\right] \geq \frac{e^{-6} N\lambda_i^2}{(N\lambda_i + \lambda)^2}.$$

For all $i > k^*$, we can directly apply (3.6.13) to get the lower bound. Therefore, according to (3.6.10), the variance error can be lower bounded as follows,

$$\begin{aligned}\text{RidgeVariance} &= \sigma^2 \cdot \sum_i \mathbb{E}_{\mu_i}\left[\frac{\lambda_i \mu_i}{(\mu_i + \lambda)^2}\right] \\ &\geq e^{-6}\sigma^2 \cdot \bigg(\sum_{i=1}^{k^*} \frac{N\lambda_i^2}{(N\lambda_i + \lambda)^2} + \sum_{i > k^*} \frac{N\lambda_i^2}{(1+\lambda)^2}\bigg).\end{aligned}$$

This completes the proof of the lower bound of the variance error. □

3.6.3 Proof of Theorem 3.3.1

Proof. In the beginning, we first recall the excess risk upper bound of SGD (see Lemma 3.6.3) and excess risk lower bound of ridge (see Lemma 3.6.3) as follows,

$$\mathbb{E}[L(\mathbf{w}_{\text{sgd}}(N_{\text{sgd}}; \gamma))] - L(\mathbf{w}^*) \leq 2 \cdot \text{SGDBias} + 2 \cdot \text{SGDVariance},$$

where

$$\begin{aligned}\text{SGDBias} &\lesssim \frac{1}{N^2\gamma^2} \cdot \big\|(\mathbf{I} - \gamma\mathbf{H})^{N/2}\mathbf{w}^*\big\|_{\mathbf{H}_{0:k_1}^{-1}} + \big\|(\mathbf{I} - \gamma\mathbf{H})^{N/2}\mathbf{w}^*\big\|_{\mathbf{H}_{k_1:\infty}}^2 \\ \text{SGDVariance} &\lesssim \sigma^2 \cdot \bigg(\frac{k_2}{N} + N\gamma^2 \sum_{i > k_2} \lambda_i^2\bigg)\end{aligned} \tag{3.6.14}$$

for arbitrary $k_1, k_2 \in [d]$.

$$\mathbb{E}[L(\mathbf{w}_{\text{ridge}}(N; \lambda))] - L(\mathbf{w}^*) = \text{RidgeBias} + \text{RidgeVariance},$$

where

$$\begin{aligned}\text{RidgeBias} &\gtrsim \max\left\{\sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2, \sum_{i=1}^{k^*} \frac{\lambda^2 \lambda_i \mathbf{w}^*[i]^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*} \lambda_i \mathbf{w}^*[i]^2\right\} \\ \text{RidgeVariance} &\gtrsim \sigma^2 \cdot \left(\sum_{i=1}^{k^*} \frac{N\lambda_i^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*} \frac{N\lambda_i^2}{(1+\lambda)^2}\right), \end{aligned} \tag{3.6.15}$$

where $k^* = \min\{k : N\lambda_k \leq 1\}$.

Next, we will show that the excess risk of SGD can be provably upper bounded (up to constant factors) by the excess risk of ridge regression respectively, given the sample size of ridge regression N_{ridge} (which we will use N in the remaining proof for simplicity). In particular, we consider two cases regarding different λ: **Case I** $\lambda < 1$ and **Case II** $\lambda \geq 1$.

For **Case I**, (3.6.15) gives the following bias lower bound for ridge regression,

$$\begin{aligned}\text{RidgeBias} &\gtrsim \sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2 \\ &\gtrsim \sum_{i>k^*} \lambda_i \mathbf{w}^*[i]^2 \\ \text{RidgeVariance} &\gtrsim \sigma^2 \cdot \left(\sum_{i=1}^{k^*} \frac{N\lambda_i^2}{(N\lambda_i+\lambda)^2} + \sum_{i>k^*} \frac{N\lambda_i^2}{(1+\lambda)^2}\right) \\ &\overset{(i)}{\eqsim} \sigma^2 \cdot \left(\frac{k^*}{N} + N\sum_{i>k^*} \lambda_i^2\right), \end{aligned}$$

where in (i) we use the fact that $N\lambda_i + \lambda \eqsim N\lambda_i$ for all $i \leq k^*$.

Then let $R^2 = \|\mathbf{w}^*\|_2^2/\sigma^2$ denotes the signal-to-noise ratio, let's consider the following configuration for SGD:

$$N_{\text{sgd}} = N, \quad \gamma = 1.$$

Then by (3.6.14) and setting $k_1 = 0$ and $k_2 = k^*$, we get

$$\begin{aligned}
\text{SGDBias} &\lesssim \sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2 \\
\text{SGDVariance} &\lesssim \sigma^2 \cdot \left(\frac{k^*}{N_{\text{sgd}}} + N_{\text{sgd}} \gamma^2 \sum_{i>k^*} \lambda_i^2 \right) \\
&\overset{(i)}{\lesssim} \sigma^2 \cdot \left(\frac{k^*}{N} + N \sum_{i>k^*} \lambda_i^2 \right).
\end{aligned}$$

Therefore, given such choice of N_{sgd} and γ, we have

$$\begin{aligned}
\mathbb{E}[L(\mathbf{w}_{\text{sgd}}(N_{\text{sgd}};\gamma))] - L(\mathbf{w}^*) &\lesssim \text{SGDBias} + \text{SGDVariance} \\
&\lesssim \sum_i (1-\lambda_i)^N \cdot \lambda_i \mathbf{w}^*[i]^2 + \sigma^2 \cdot \left(\frac{k^*}{N} + N \sum_{i>k^*} \lambda_i^2 \right) \\
&\lesssim \text{RidgeBias} + \text{RidgeVariance} \\
&= \mathbb{E}[L(\mathbf{w}_{\text{ridge}(N;\lambda)})] - L(\mathbf{w}^*).
\end{aligned}$$

For **Case II**, we can define $\widetilde{k}^* = \min\{k : N\lambda_k \le \lambda\}$, then (3.6.15) implies

$$\begin{aligned}
\text{RidgeBias} &\gtrsim \sum_{i=1}^{k^*} \frac{\lambda^2 \lambda_i \mathbf{w}^*[i]^2}{(N\lambda_i + \lambda)^2} + \sum_{i>k^*} \lambda_i \mathbf{w}^*[i]^2 \\
&\overset{(i)}{\eqsim} \sum_{i=1}^{\widetilde{k}^*} \frac{\lambda^2 \mathbf{w}^*[i]^2}{N^2 \lambda_i} + \sum_{i>\widetilde{k}^*} \lambda_i \mathbf{w}^*[i]^2 \\
\text{RidgeVariance} &\gtrsim \sigma^2 \cdot \left(\sum_{i=1}^{k^*} \frac{N\lambda_i^2}{(N\lambda_i + \lambda)^2} + \sum_{i>k^*} \frac{N\lambda_i^2}{(1+\lambda)^2} \right) \\
&\overset{(ii)}{\eqsim} \sigma^2 \cdot \left(\frac{\widetilde{k}^*}{N} + \frac{N}{\lambda^2} \sum_{i>\widetilde{k}^*} \lambda_i^2 \right),
\end{aligned}$$

where (i) and (ii) are due to the fact that for every $i \le k^*$, we have

$$\frac{1}{(N\lambda_i + \lambda)^2} \eqsim \begin{cases} \frac{1}{N^2\lambda_i} & i \le \widetilde{k}^* \\ \frac{1}{\lambda^2} & \widetilde{k}^* < i \le k^*. \end{cases}$$

Therefore, we can apply the following configuration for SGD:

$$N_{\text{sgd}} = N, \quad \gamma = 1/\lambda.$$

Then by (3.6.14) and set $k_1 = k_2 = \widetilde{k}^*$, we have

$$\begin{aligned}
&\mathbb{E}[L(\mathbf{w}_{\mathrm{sgd}}(N_{\mathrm{sgd}};\gamma))] - L(\mathbf{w}^*) \\
&\lesssim \mathrm{SGDBias} + \mathrm{SGDVariance} \\
&\lesssim \sum_{i=1}^{\widetilde{k}^*} \frac{(1-\gamma\lambda_i)^{N_{\mathrm{sgd}}}\mathbf{w}^*[i]^2}{\lambda_i N_{\mathrm{sgd}}^2\gamma^2} + \sum_{i>\widetilde{k}^*}\lambda_i \mathbf{w}^*[i]^2 + \sigma^2\cdot\left(\frac{\widetilde{k}^*}{N_{\mathrm{sgd}}} + N_{\mathrm{sgd}}\gamma^2\sum_{i>\widetilde{k}^*}\lambda_i^2\right) \\
&\eqsim \sum_{i=1}^{\widetilde{k}^*} \frac{\lambda^2\mathbf{w}^*[i]^2}{\lambda_i N^2} + \sum_{i>\widetilde{k}^*}\lambda_i \mathbf{w}^*[i]^2 + \sigma^2\cdot\left(\frac{\widetilde{k}^*}{N} + \frac{N}{\lambda^2}\sum_{i>\widetilde{k}^*}\lambda_i^2\right) \\
&\lesssim \mathrm{RidgeBias} + \mathrm{RidgeVariance} \\
&= \mathbb{E}[L(\mathbf{w}_{\mathrm{ridge}}(N;\lambda))] - L(\mathbf{w}^*).
\end{aligned}$$

Combining the results for these two cases completes the proof. □

3.6.4 Proof of Theorem 3.3.2

Proof. For simplicity we define $N := N_{\mathrm{sgd}}$ in the proof.

- The data covariance matrix $\mathbf{H}$ has the following spectrum

$$\lambda_i = \begin{cases} \frac{\log(N)}{N^{1/2}} & i = 1, \\ \frac{1-\log(N)/N^{1/2}}{N} & 1 < i \le N, \\ 0 & N < i \le d \end{cases}$$

- The true parameter $\mathbf{w}^*$ is given by

$$\mathbf{w}^*[i] = \begin{cases} \sigma\cdot\sqrt{\frac{N^{1/2}}{\log(N)}} & i = 1, \\ 0 & 1 < i \le d. \end{cases}$$

Then it is easy to verify that $\mathrm{tr}(\mathbf{H}) = 1$. For SGD, we consider setting the stepsize as $\gamma^* = N^{-1/2}$. Then by Lemma 3.6.3 and choosing $k_1 = 1$, we have the following on the bias

error of SGD,

$$\text{SGDBias} \lesssim \sum_{i=1}^{k^*} \frac{(1-\gamma\lambda_i)^{N_{\text{sgd}}}\mathbf{w}^*[i]^2}{\lambda_i N_{\text{sgd}}^2\gamma^2} + \sum_{i>k^*}\lambda_i \mathbf{w}^*[i]^2 \lesssim \frac{(1-\log(N)/N)^N\sigma^2}{\log^2(N)} \lesssim \frac{\sigma^2}{N}.$$

For variance error, we can pick $k_2 = 1$ and get

$$\text{SGDVariance} \lesssim \sigma^2 \cdot \left(\frac{1}{N} + N\gamma^2\sum_{i>1}\lambda_i^2\right) \lesssim \sigma^2\left(\frac{1}{N} + \sum_{i>1}\lambda_i^2\right) \eqsim \frac{\sigma^2}{N}.$$

Now let us characterize the excess risk of ridge regression. In terms of the bias error, by Lemma 3.6.5 we have

$$\text{RidgeBias} \gtrsim \sum_{i=1}^{k^*} \frac{\lambda^2\lambda_i\mathbf{w}^*[i]^2}{(N_{\text{ridge}}\lambda_i+\lambda)^2} + \sum_{i>k^*}\lambda_i\mathbf{w}^*[i]^2 \eqsim \frac{\lambda^2\sigma^2}{(N_{\text{ridge}}\log(N)/N^{1/2}+\lambda)^2}, \tag{3.6.16}$$

where $k^* = \min\{k : N_{\text{ridge}}\lambda_k \le 1\}$. Then it is clear for ridge regression we must have $\lambda \lesssim N_{\text{ridge}}\log(N)/N^{1/2}$ since otherwise $\text{RidgeBias} \gtrsim \sigma^2 \gtrsim \text{SGDRisk}$. Regarding the variance, we have

$$\text{RidgeVariance} \gtrsim \sigma^2 \cdot \left(\sum_{i=1}^{k^*}\frac{N_{\text{ridge}}\lambda_i^2}{(N_{\text{ridge}}\lambda_i+\lambda)^2} + \sum_{i>k^*}\frac{N_{\text{ridge}}\lambda_i^2}{(1+\lambda)^2}\right).$$

Then we will consider two cases: (1) $N_{\text{ridge}} \lesssim N$ and (2) $N_{\text{ridge}} \gtrsim N$. In the first case we can get $k^* = 1$ and then

$$\text{RidgeVariance} \gtrsim \sigma^2 \cdot \left(\frac{N_{\text{ridge}}\log^2(N)/N^2}{(N_{\text{ridge}}\log(N)/N+\lambda)^2} + \frac{N_{\text{ridge}}}{N^2(1+\lambda)^2}\right) \ge \frac{N_{\text{ridge}}\sigma^2}{N^2(1+\lambda^2)}.$$

In this case, we can get $k^* = 1$ and thus

$$\text{RidgeVariance} \gtrsim \sigma^2 \cdot \frac{N_{\text{ridge}}\log^2(N)/N^2}{(N_{\text{ridge}}\log(N)/N+\lambda)^2} \overset{(i)}{\eqsim} \frac{\sigma^2}{N_{\text{ridge}}} \overset{(ii)}{\gtrsim} \frac{\sigma^2}{N},$$

where (i) is due to we require $\lambda \lesssim N_{\text{ridge}}\log(N)/N^{1/2}$ to guarantee vanishing bias error and (ii) is due to in this case we have $N_{\text{ridge}} \lesssim N$. As a result, ridge regression cannot achieve smaller excess risk than SGD in this case.

In the second case we can get $k^* = N$ and then

$$\begin{aligned}\text{RidgeVariance} &\gtrsim \sigma^2 \cdot \left(\frac{N_{\text{ridge}}\log^2(N)/N^2}{(N_{\text{ridge}}\log(N)/N+\lambda)^2} + \frac{(k^*-1)\cdot N_{\text{ridge}}/N^2}{(N_{\text{ridge}}/N+\lambda^2)}\right) \\ &\gtrsim \sigma^2 \cdot \frac{NN_{\text{ridge}}}{N_{\text{ridge}}^2+N^2\lambda^2},\end{aligned} \tag{3.6.17}$$

where the second inequality is due to $k^* = N$. We will again consider two cases: (a) $N_{\text{ridge}} \gtrsim N\lambda$ and (b) $N_{\text{ridge}} \lesssim N\lambda$. Regarding Case (a) we have

$$\text{RidgeVariance} \geq \frac{N\sigma^2}{N_{\text{ridge}}},$$

and it is clear that for all $N_{\text{ridge}} \lesssim N^2$ we have RidgeVariance $\gtrsim \sigma^2/N \gtrsim$ SGDRisk. Regarding Case (b), combining the lower bounds of bias (3.6.16) and variance (3.6.17) of ridge regression, we get

$$\text{RidgeRisk} \gtrsim \sigma^2 \cdot \left(\frac{\lambda^2 N}{N_{\text{ridge}}^2 \log^2(N)} + \frac{N_{\text{ridge}}}{N\lambda^2} \right) \gtrsim \frac{\sigma^2}{N_{\text{ridge}}^{1/2} \log(N)},$$

where the first inequality follows from the fact that $\lambda \lesssim N_{\text{ridge}} \log(N)/N^{1/2}$ and $N_{\text{ridge}} \lesssim N\lambda$, and the second inequality is by Cauchy-Schwartz inequality. This further suggests that RidgeRisk $\lesssim \sigma^2/N \lesssim$ SGDRisk if $N_{\text{ridge}} \leq N^2/\log^2(N)$, which completes the proof.

□

3.7 Proof of Gaussian Least Squares

3.7.1 Excess risk bounds of SGD and ridge regression

We first recall the excess risk bounds for SGD (with tail averaging) and ridge regression as follows.

SGD with tail averaging

Theorem 3.7.1 (Extension of Theorem 5.1 in [ZWB21])**.** Consider SGD with tail-averaging with initialization $\mathbf{w}_0 = \mathbf{0}$. Suppose Assumption 3.4.1 holds and the stepsize satisfies $\gamma \lesssim 1/\operatorname{tr}(\mathbf{H})$. Then the excess risk can be upper bounded as follows,

$$\mathbb{E}[L(\mathbf{w}_{\text{sgd}}(N;\gamma))] - L(\mathbf{w}^*) \leq \text{SGDBias} + \text{SGDVariance},$$

where

$$\begin{aligned}\text{SGDBias} &\lesssim \frac{1}{\gamma^2 N^2} \cdot \left\|(\mathbf{I}-\gamma\mathbf{H})^N \mathbf{w}^*\right\|_{\mathbf{H}_{0:k_1}^{-1}}^2 + \left\|(\mathbf{I}-\gamma\mathbf{H})^N \mathbf{w}^*\right\|_{\mathbf{H}_{k_1:\infty}}^2 \\ \text{SGDVariance} &\lesssim \frac{\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} \cdot \left(k_2 + N^2\gamma^2 \sum_{i>k_2} \lambda_i^2\right).\end{aligned}$$

where $k_1, k_2 \in [d]$ are arbitrary.

This theorem is a simple extension of Theorem 5.1 in [ZWB21]. In particular, we observe that though the original theorem is stated for some particular k^* and $k^\dagger$, based on the proof, their results hold for arbitrary k_1 and k_2, as stated in Theorem 3.7.1.

Ridge regression. See Appendix 3.8 for a proof of the following theorem.

Theorem 3.7.2 (Extension of Lemmas 2 & 3 in [TB20])**.** Suppose Assumption 3.4.1 holds. Let $\lambda \geq 0$ be the regularization parameter, n be the training sample size and $\widehat{\mathbf{w}}_{\text{ridge}}(N;\lambda)$ be the output of ridge regression. Then

$$\mathbb{E}\big[L(\mathbf{w}_{\text{ridge}}(N;\lambda))\big] - L(\mathbf{w}^*) = \text{RidgeBias} + \text{RidgeVariance},$$

and there is some absolute constant $b > 1$, such that for

$$k_{\text{ridge}}^* := \min\left\{k : b\lambda_{k+1} \leq \frac{\lambda + \sum_{i>k}\lambda_i}{n}\right\},$$

the following holds:

$$\begin{aligned}\text{RidgeBias} &\gtrsim \left(\frac{\lambda + \sum_{i>k_{\text{ridge}}^*}\lambda_i}{N}\right)^2 \cdot \|\mathbf{w}^*\|_{\mathbf{H}_{0:k_{\text{ridge}}^*}^{-1}}^2 + \|\mathbf{w}^*\|_{\mathbf{H}_{k_{\text{ridge}}^*:\infty}}^2, \\ \text{RidgeVariance} &\gtrsim \sigma^2 \cdot \left\{\frac{k_{\text{ridge}}^*}{N} + \frac{N\sum_{i>k_{\text{ridge}}^*}\lambda_i^2}{\left(\lambda + \sum_{i>k_{\text{ridge}}^*}\lambda_i\right)^2}\right\}.\end{aligned}$$

3.7.2 Proof of Theorem 3.4.2

Proof. For simplicity, let us fix $N := N_{\text{ridge}}$ and $k := k_{\text{ridge}}$, we will next locate γ such that the risk of SGD competes with that of Ridge. Denote $\widetilde{\lambda} := \lambda + \sum_{i>k} \lambda_i$. Then

$$\begin{aligned}\text{RidgeRisk} &= \text{RidgeBias} + \text{RidgeVariance} \\ &\gtrsim \left(\frac{\widetilde{\lambda}}{N}\right)^2 \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N}\left(k + \left(\frac{N}{\widetilde{\lambda}}\right)^2 \sum_{i>k} \lambda_i^2\right).\end{aligned}$$

Then for SGD we can set

$$N_{\text{sgd}} = (1+R^2) \cdot N \cdot (1 \vee \kappa \log a),$$

where

$$\kappa := \frac{\text{tr}(\mathbf{H})}{N\lambda_N}, \qquad a = \frac{\text{tr}(\mathbf{H})}{\lambda + \sum_{i>N} \lambda_i} \wedge (\kappa R\sqrt{N}) = \frac{\text{tr}(\mathbf{H})}{\lambda + \sum_{i>N} \lambda_i} \wedge \frac{\text{tr}(\mathbf{H})R}{\sqrt{N}\lambda_N}.$$

Next we discuss two cases:

Case I, $\widetilde{\lambda} \cdot (1 \vee \kappa \log a) \geq \text{tr}(\mathbf{H})$**.** For SGD, let us set $k_{\text{sgd}} = k$ and that

$$\gamma = \frac{1}{(1+R^2) \cdot \widetilde{\lambda} \cdot (1 \vee \kappa \log a)} \leq \frac{1}{\text{tr}(\mathbf{H})},$$

then

$$N_{\text{sgd}} \cdot \gamma = \frac{N}{\widetilde{\lambda}}.$$

Thus we obtain that

$$\begin{aligned}\text{SGDRisk} &\lesssim \frac{(1-\gamma\lambda_k)^{2N_{\text{sgd}}}}{(\gamma N_{\text{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{(1+R^2)\sigma^2}{N_{\text{sgd}}}\left(k + (\gamma N_{\text{sgd}})^2 \sum_{i>k} \lambda_i^2\right) \\ &= \frac{(1-\gamma\lambda_k)^{2N_{\text{sgd}}}}{\left(N/\widetilde{\lambda}\right)^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N(1 \vee \kappa \log a)}\left(k + \left(N/\widetilde{\lambda}\right)^2 \sum_{i>k} \lambda_i^2\right) \\ &\leq \left(\frac{\widetilde{\lambda}}{N}\right)^2 \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N}\left(k + \left(\frac{N}{\widetilde{\lambda}}\right)^2 \sum_{i>k} \lambda_i^2\right) \\ &\lesssim \text{RidgeRisk}.\end{aligned}$$

Case II, $\widetilde{\lambda} \cdot (1 \vee \kappa \log a) < \mathrm{tr}(\mathbf{H})$. For SGD, let us set $k_{\mathrm{sgd}} = k$ and that

$$\gamma = \frac{1}{(1+R^2) \cdot \mathrm{tr}(\mathbf{H})} \leq \frac{1}{\mathrm{tr}(\mathbf{H})},$$

then

$$N_{\mathrm{sgd}} \cdot \gamma = \frac{N \cdot (1 \vee \kappa \log a)}{\mathrm{tr}(\mathbf{H})} \leq \frac{N}{\widetilde{\lambda}}.$$

We obtain that

$$\begin{aligned}
\mathrm{SGDRisk} &\leq \mathrm{SGDBias} + \mathrm{SGDVariance} \\
&\lesssim \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{(1+R^2)\sigma^2}{N_{\mathrm{sgd}}} \left(k + (\gamma N_{\mathrm{sgd}})^2 \sum_{i>k} \lambda_i^2 \right) \\
&\leq \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N(1 \vee \kappa \log a)} \left(k + \left(N/\widetilde{\lambda} \right)^2 \sum_{i>k} \lambda_i^2 \right) \\
&\leq \frac{(1-\gamma\lambda_k)^{2N_{\mathrm{sgd}}}}{(\gamma N_{\mathrm{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N} \left(k + \left(\frac{N}{\widetilde{\lambda}} \right)^2 \sum_{i>k} \lambda_i^2 \right).
\end{aligned}$$

The second and the third terms match those of ridge error. As for the first term, notice that by the choice of γ and that $\lambda_k \geq \lambda_N$, we have that

$$\begin{aligned}
\frac{(1-\gamma\lambda_k)^{N_{\mathrm{sgd}}}}{\gamma N_{\mathrm{sgd}}} &\leq \left(1 - \frac{\lambda_N}{(1+R^2) \cdot \mathrm{tr}(\mathbf{H})} \right)^{N_{\mathrm{sgd}}} \cdot \frac{1}{\gamma N_{\mathrm{sgd}}} \\
&= \left(1 - \frac{1}{(1+R^2) \cdot N \cdot \kappa} \right)^{(1+R^2) \cdot N \cdot (1 \vee \kappa \log a)} \cdot \frac{\mathrm{tr}(\mathbf{H})}{N \cdot (1 \vee \kappa \log a)} \\
&\leq \left(1 - \frac{1}{(1+R^2) \cdot N \cdot \kappa} \right)^{(1+R^2) \cdot N \cdot \kappa \log a} \cdot \frac{\mathrm{tr}(\mathbf{H})}{N} \\
&\leq \frac{1}{a} \cdot \frac{\mathrm{tr}(\mathbf{H})}{N} = \frac{(\lambda + \sum_{i>N} \lambda_i) \vee (\sqrt{N}\lambda_N / R)}{\mathrm{tr}(\mathbf{H})} \cdot \frac{\mathrm{tr}(\mathbf{H})}{N} \\
&\leq \frac{\lambda + \sum_{i>k} \lambda_i}{N} \vee \frac{\lambda_k}{R \cdot \sqrt{N}} = \frac{\widetilde{\lambda}}{N} \vee \frac{\lambda_k}{R \cdot \sqrt{N}}.
\end{aligned}$$

If $\frac{(1-\gamma\lambda_k)^{N_{\text{sgd}}}}{\gamma N_{\text{sgd}}} \leq \frac{\widetilde{\lambda}}{N}$, then

$$\begin{aligned}
\text{SGDRisk} &\lesssim \frac{(1-\gamma\lambda_k)^{2N_{\text{sgd}}}}{(\gamma N_{\text{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N}\left(k + \left(\frac{N}{\widetilde{\lambda}}\right)^2 \sum_{i>k} \lambda_i^2\right) \\
&\leq \left(\frac{\widetilde{\lambda}}{N}\right)^2 \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N}\left(k + \left(\frac{N}{\widetilde{\lambda}}\right)^2 \sum_{i>k} \lambda_i^2\right) \\
&\lesssim \text{RidgeRisk}.
\end{aligned}$$

If $\frac{(1-\gamma\lambda_k)^{N_{\text{sgd}}}}{\gamma N_{\text{sgd}}} \leq \frac{\lambda_k}{R\cdot\sqrt{N}}$, then

$$\frac{(1-\gamma\lambda_k)^{2N_{\text{sgd}}}}{(\gamma N_{\text{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} \leq \frac{\lambda_k^2}{R^2\cdot N} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} \leq \frac{\|\mathbf{w}^*\|^2_{\mathbf{H}}}{R^2\cdot N} \leq \frac{\sigma^2}{N},$$

and

$$\begin{aligned}
\text{SGDRisk} &\lesssim \frac{(1-\gamma\lambda_k)^{2N_{\text{sgd}}}}{(\gamma N_{\text{sgd}})^2} \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N}\left(k + \left(\frac{N}{\widetilde{\lambda}}\right)^2 \sum_{i>k} \lambda_i^2\right) \\
&\leq \frac{\sigma^2}{N} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k:\infty}} + \frac{\sigma^2}{N}\left(k + \left(\frac{N}{\widetilde{\lambda}}\right)^2 \sum_{i>k} \lambda_i^2\right) \\
&\lesssim 2\cdot \text{RidgeRisk}.
\end{aligned}$$

These complete the proof. □

3.7.3 Proof of Corollary 3.4.3

Proof. By Theorem 3.4.2, we only need to verify that $\kappa(N_{\text{ridge}}) \lesssim \log(N_{\text{ridge}})$. Recall that $\lambda_i = 1/i^\alpha$ for $0 < \alpha \leq 1$, and $d \lesssim N_{\text{ridge}}$. For $\alpha = 1$, then

$$\text{tr}(\mathbf{H}) = \sum_{i=1}^{d} i^{-\alpha} \lesssim \log d \lesssim \log(N_{\text{ridge}}),$$

thus

$$\kappa(N_{\text{ridge}}) = \frac{\text{tr}(\mathbf{H})}{N_{\text{ridge}}\lambda_{\min\{d, N_{\text{ridge}}\}}} \lesssim \frac{\log(N_{\text{ridge}})}{N_{\text{ridge}}\cdot N_{\text{ridge}}^{-1}} = \log(N_{\text{ridge}}).$$

For $\alpha < 1$, then

$$\text{tr}(\mathbf{H}) = \sum_{i=1}^{d} i^{-\alpha} \lesssim d^{1-\alpha} \lesssim N_{\text{ridge}}^{1-\alpha},$$

thus

$$\kappa(N_{\text{ridge}}) = \frac{\text{tr}(\mathbf{H})}{N_{\text{ridge}}\lambda_{\{N_{\text{ridge}},d\}}} \lesssim \frac{N_{\text{ridge}}^{1-\alpha}}{N_{\text{ridge}} \cdot N_{\text{ridge}}^{-\alpha}} = 1.$$

□

3.7.4 Proof of Corollary 3.4.4

Proof. Note that given random $\mathbf{w}^*$, the expected risk considered in our paper will be including the expectation over both random data $\mathbf{x}$ and random ground-truth $\mathbf{w}^*$. Since the distribution of $\mathbf{w}^*$ is rotation invariant, the expectation of $\mathbf{w}^*[i]$ will be the same for all $i \in [d]$. Therefore, let $B = \mathbb{E}[(\mathbf{w}^*[i])^2]$, the following holds according to (3.5.2)

$$\begin{aligned}\text{RidgeRisk} &\gtrsim \left(\frac{\widetilde{\lambda}}{N_{\text{ridge}}}\right)^2 \cdot \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}}\big] + \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big] + \sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2}\sum_{i>k^*}\lambda_i^2\right) \\ &= B\left(\frac{\widetilde{\lambda}}{N_{\text{ridge}}}\right)^2 \cdot \sum_{i=1}^{k^*} i^{\alpha} + B \cdot \sum_{i=k^*+1} i^{-\alpha} + \sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2}\sum_{i>k^*}\lambda_i^2\right)\end{aligned}$$

where $k^* = \min\{k : N_{\text{ridge}}\lambda_k \le \widetilde{\lambda}\}$. Then note that $\lambda_i = i^{-\alpha}$, we have $k^* = (N_{\text{ridge}}/\widetilde{\lambda})^{1/\alpha}$, which implies that

$$\begin{aligned}\text{RidgeRisk} &\gtrsim B\left(\frac{\widetilde{\lambda}}{N_{\text{ridge}}}\right)^2 \cdot (k^*)^{1+\alpha} + B \cdot \big[d^{1-\alpha} - (k^*)^{1-\alpha}\big] + \sigma^2 \cdot \left(\frac{k^*}{N_{\text{ridge}}} + \frac{N_{\text{ridge}}}{\widetilde{\lambda}^2}\sum_{i>k^*}\lambda_i^2\right) \\ &\gtrsim N_{\text{ridge}}^{1-\alpha} \cdot B\end{aligned}$$

where we use the fact that $d = \Theta(N)$. Note that constant SNR $R = \Theta(1)$ implies that

$$\sigma^2 \eqsim B\sum_{i=1}^{d}\lambda_i \eqsim N_{\text{ridge}}^{1-\alpha}B.$$

Then by (3.5.1) and set $N_{\text{sgd}} = N_{\text{ridge}} = N$ and $k_1 = k_2 = N_{\text{ridge}}$, we have

$$\begin{aligned}\text{SGDRisk} &\lesssim \frac{1}{\gamma^2 N_{\text{sgd}}^2} \cdot \mathbb{E}\big[\big\|\exp(-N_{\text{sgd}}\gamma\mathbf{H})\mathbf{w}^*\big\|^2_{\mathbf{H}^{-1}_{0:k_1}}\big] + \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}_{k_1:\infty}}\big] \\ &\quad + (1+R^2)\sigma^2 \cdot \left(\frac{k_2}{N_{\text{sgd}}} + N_{\text{sgd}}\gamma^2\sum_{i>k_2}\lambda_i^2\right) \\ &= \frac{1}{\gamma^2 N^2} \cdot \mathbb{E}\big[\big\|\mathbf{w}^*\big\|^2_{\mathbf{H}^{-1}_{0:N}}\big] + \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}_{N:d}}\big] + BN^{1-\alpha} \cdot \left(1 + N\gamma^2\sum_{i>N}\lambda_i^2\right).\end{aligned}$$

Note that we have

$$\mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:N}}\big] = BN^{1+\alpha},\ \mathbb{E}\big[\|\mathbf{w}^*\|^2_{\mathbf{H}_{N:d}}\big] = BN^{1-\alpha}.$$

Then we can set $\gamma \eqsim 1/\operatorname{tr}(\mathbf{H}) \eqsim N^{\alpha-1}$ and get

$$\begin{aligned}\text{SGDRisk} &\lesssim \frac{B}{N^{2\alpha}} \cdot N^{1+\alpha} + BN^{1-\alpha} + BN^{1-\alpha} \cdot \left(1 + N\gamma^2 \sum_{i>N} \lambda_i^2\right) \\ &\lesssim BN^{1-\alpha} \\ &\lesssim \text{RidgeRisk}.\end{aligned}$$

This implies that SGD can be no worse than ridge regression as long as provided same or larger sample size, which completes the proof.

□

3.7.5 Proof of Theorem 3.4.5

Proof. For simplicity we fix $N := N_{\text{sgd}}$. Let us consider the following problem instance:

- The data covariance matrix $\mathbf{H}$ has the following spectrum

$$\lambda_i = \begin{cases} 1 & i = 1, \\ \frac{1}{N \log N} & 1 < i \le N^2, \\ 0 & N^2 < i \le d \end{cases}$$

  where we require the dimension $d \ge N^2$. We note that $\operatorname{tr}(\mathbf{H}) = 1 + N/\log N \eqsim N/\log N$.

- The true parameter $\mathbf{w}^*$ is given by

$$\mathbf{w}^*[i] = \begin{cases} \sigma & i = 1, \\ 0 & 1 < i \le d. \end{cases}$$

Then for SGD, we choose stepsize as $\gamma = \log(N)/(2N) \leq 1/\operatorname{tr}(\mathbf{H})$. By Lemma 3.7.1, we have the following excess risk bound for $\mathbf{w}_{\mathrm{sgd}}(N;\gamma^*)$,

$$L\big[\mathbf{w}_{\mathrm{sgd}}(N;\gamma)\big] - L(\mathbf{w}^*) \leq \mathrm{SGDBias} + \mathrm{SGDVariance},$$

where

$$\begin{aligned}
\mathrm{SGDBias} &\lesssim \sigma^2 \cdot \frac{(1-\gamma)^N}{(\gamma N)^2} \lesssim \sigma^2 \cdot \log^2 N \cdot \left(1 - \frac{\log N}{2N}\right)^N \lesssim \frac{\sigma^2 \log^2 N}{N^2} \lesssim \frac{\sigma^2}{N}, \\
\mathrm{SGDVariance} &\lesssim \frac{\sigma^2}{N} \cdot \left(1 + (N\gamma)^2 \sum_{i>1} \lambda_i^2\right) \eqsim \frac{\sigma^2}{N},
\end{aligned}$$

where we use the fact that $\sum_{i>1} \lambda_i^2 = \frac{1}{\log^2 N}$. This implies that SGD with sample size N achieves at most $\mathcal{O}(\sigma^2/N)$ excess risk on this example.

Then we calculate the excess risk lower bound of ridge regression. By Lemma 3.7.2 and let $\widetilde{\lambda} = \lambda + \sum_{i>k^*_{\mathrm{ridge}}} \lambda_i$, we have

$$\begin{aligned}
L\big[\mathbf{w}_{\mathrm{ridge}}(N;\lambda)\big] - L(\mathbf{w}^*) &= \mathrm{RidgeBias} + \mathrm{RidgeVariance} \\
&\gtrsim \sigma^2 \cdot \left(\frac{\widetilde{\lambda}^2}{N_{\mathrm{ridge}}^2} + \frac{k^*_{\mathrm{ridge}}}{N_{\mathrm{ridge}}} + \frac{N_{\mathrm{ridge}} \sum_{i>k^*_{\mathrm{ridge}}} \lambda_i^2}{\widetilde{\lambda}^2}\right).
\end{aligned}$$

If $k^*_{\mathrm{ridge}} > N$, then

$$L\big[\mathbf{w}_{\mathrm{ridge}}(N;\lambda)\big] - L(\mathbf{w}^*) \gtrsim \frac{\sigma^2 k^*_{\mathrm{ridge}}}{N_{\mathrm{ridge}}} \geq \frac{\sigma^2 N}{N_{\mathrm{ridge}}} \geq \frac{\sigma^2}{N}, \quad \text{for } N_{\mathrm{ridge}} < \frac{N^2}{\log^2 N}.$$

If $k^*_{\mathrm{ridge}} \leq N$, then $\sum_{i>k^*_{\mathrm{ridge}}} \lambda_i^2 \geq \sum_{N<i\leq N^2} \frac{1}{N^2 \log^2 N} \eqsim \frac{1}{\log^2 N}$, which implies that

$$\begin{aligned}
L\big[\mathbf{w}_{\mathrm{ridge}}(N;\lambda)\big] - L(\mathbf{w}^*) &\gtrsim \sigma^2 \cdot \left(\frac{\widetilde{\lambda}^2}{N_{\mathrm{ridge}}^2} + \frac{N_{\mathrm{ridge}}}{\widetilde{\lambda}^2} \cdot \frac{1}{\log^2 N}\right) \\
&\geq \frac{\sigma^2}{N_{\mathrm{ridge}}^{1/2} \log N} \\
&\geq \frac{\sigma^2}{N}, \quad \text{for } N_{\mathrm{ridge}} < \frac{N^2}{\log^2 N}.
\end{aligned}$$

To sum up, we have show that

$$L\big[\mathbf{w}_{\mathrm{ridge}}(N_{\mathrm{ridge}};\lambda)\big] - L(\mathbf{w}^*) \gtrsim \frac{\sigma^2}{N} \gtrsim L\big[\mathbf{w}_{\mathrm{sgd}}(N;\lambda)\big] - L(\mathbf{w}^*), \quad \text{for } N_{\mathrm{ridge}} < \frac{N^2}{\log^2 N}.$$

This completes the proof. □

3.7.6 Proof of Theorem 3.4.6

Proof. The proof of Theorem 3.4.6 is similar to that of Theorem 3.4.2. In particular, we still consider two cases: (1) $\lambda \gtrsim \mathrm{tr}(\mathbf{H})$ and (2) $\lambda \lesssim \mathrm{tr}(\mathbf{H})$. For the first case, we can use the identical proof in Theorem 3.4.2 and get that SGD with sample size $N_{\mathrm{sgd}} \eqsim (1+R^2)\cdot N_{\mathrm{ridge}}$ to achieve better excess risk than ridge regression. Note that we have assumed $R^2 = \Theta(1)$, therefore, we can claim that SGD outperforms ridge regression, as long as the sample size is at least in the same order of N_{ridge}.

For the second case that $\lambda \lesssim \mathrm{tr}(\mathbf{H})$, for simplicity we denote $N := N_{\mathrm{ridge}}$ and we can directly set $\gamma = 1/\mathrm{tr}(\mathbf{H})$ and $N_{\mathrm{sgd}} = N$. Let $k^* = \min\big\{k : \lambda_k \le \frac{\mathrm{tr}(\mathbf{H})\log(N)}{N}\big\}$, then by the definition of k^*_{ridge} in Lemma 3.7.2 and the assumption that ridge regression is in the generalizable regime, we have $k^* \le k^*_{\mathrm{ridge}} \le N_{\mathrm{ridge}}$. Therefore, applying Lemma 3.7.1 with $k_1 = k^*$, we have the following bound on the effective bias of SGD,

$$\begin{aligned}
\mathrm{SGDBias} &\lesssim \sum_{i=1}^{k^*} \frac{(1-\gamma\lambda_i)^N (\mathbf{w}^*[i])^2}{\lambda_i \gamma^2 N^2} + \sum_{i>k^*} \lambda_i (\mathbf{w}[i])^2 \\
&\lesssim \sum_{i=1}^{k^*} \frac{\big(1-\frac{\log(N)}{N}\big)^N (\mathbf{w}^*[i])^2}{\lambda_i N^2} + \sum_{i>k^*} \lambda_i (\mathbf{w}[i])^2 \\
&\lesssim \frac{\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} + \sum_{i>k^*} \lambda_i (\mathbf{w}[i])^2.
\end{aligned}$$

Then by our assumption that

$$\sum_{i=k^*}^{N_{\mathrm{ridge}}} \lambda_i \big(\mathbf{w}[i]\big)^2 \lesssim \frac{k^* \|\mathbf{w}^*\|_{\mathbf{H}}^2}{N},$$

we further have

$$\begin{aligned}
\mathrm{SGDBias} &\lesssim \frac{\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} + \sum_{i>k^*} \lambda_i (\mathbf{w}[i])^2 \\
&\lesssim \sum_{i>k^*_{\mathrm{ridge}}} \lambda_i (\mathbf{w}[i])^2 + \frac{(k^*_{\mathrm{ridge}}+1)\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N},
\end{aligned}$$

where in the second inequality we use the fact that $k^* \leq k^*_{\text{ridge}} \leq N_{\text{ridge}}$. Regarding the variance of SGD, applying Lemma 3.7.1 with $k_2 = k^*_{\text{ridge}}$ gives

$$\begin{aligned}
\text{SGDVariance} &\lesssim (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2) \cdot \left(\frac{k^*_{\text{ridge}}}{N} + \frac{N}{(\text{tr}(\mathbf{H})^2)} \cdot \sum_{i \geq k^*_{\text{ridge}}} \lambda_i^2 \right) \\
&\lesssim (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2) \cdot \left(\frac{k^*_{\text{ridge}}}{N} + \frac{N}{(\lambda + \sum_{i > k^*_{\text{ridge}}} \lambda_i)^2} \cdot \sum_{i \geq k^*_{\text{ridge}}} \lambda_i^2 \right),
\end{aligned}$$

where the last inequality is due to the fact that $\lambda \lesssim \text{tr}(\mathbf{H})$. Combining the above upper bounds for the bias and variance of SGD, we have that the output of SGD, with sample size $N_{\text{sgd}} = N$ and learning rate $\gamma = 1/\text{tr}(\mathbf{H})$, satisfies

$$\begin{aligned}
\text{SGDRisk} &\lesssim \text{SGDBias} + \text{SGDVariance} \\
&\lesssim \sum_{i > k^*_{\text{ridge}}} \lambda_i (\mathbf{w}[i])^2 + \frac{(k^*_{\text{ridge}} + 1)\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} \\
&\quad + (\sigma^2 + \|\mathbf{w}^*\|_{\mathbf{H}}^2) \cdot \left(\frac{k^*_{\text{ridge}}}{N_{\text{ridge}}} + \frac{N_{\text{ridge}} \gamma^2}{(\lambda + \sum_{i > N_{\text{ridge}}} \lambda_i)^2} \cdot \sum_{i \geq k^*_{\text{ridge}}} \lambda_i^2 \right) \\
&\eqsim \sum_{i > k^*_{\text{ridge}}} \lambda_i (\mathbf{w}[i])^2 + \frac{(k^*_{\text{ridge}} + 1)\|\mathbf{w}^*\|_{\mathbf{H}}^2}{N} \\
&\quad + \sigma^2 \cdot \left(\frac{k^*_{\text{ridge}}}{N_{\text{ridge}}} + \frac{N_{\text{ridge}} \gamma^2}{(\lambda + \sum_{i > N_{\text{ridge}}} \lambda_i)^2} \cdot \sum_{i \geq k^*_{\text{ridge}}} \lambda_i^2 \right) \\
&\lesssim \text{RidgeBias} + \text{RidgeVariance}, \qquad (3.7.1)
\end{aligned}$$

where the last equality holds since we assume that $\|\mathbf{w}\|_{\mathbf{H}}^2 / \sigma^2 = \Theta(1)$. Note that the R.H.S. of (3.7.1) is exactly the lower bound of the excess risk of ridge regression. Therefore, we can conclude that as long as $N_{\text{sgd}} = N$, SGD with a tuned stepsize γ will be no worse than ridge regression for all λ (up to constant factors). This completes the proof.

□

3.8 Proof of Theorem 3.7.2

In this section we always make Assumption 3.4.1. The results and techniques are either explicitly or implicitly presented in [BLL20; TB20]. For self-completeness, we provide a formal proof here.

Notation. Following [TB20] and [BLL20], we define the following notations:

- $\mathbf{v} := \mathbf{H}^{-\frac{1}{2}}\mathbf{x} \in \mathbb{R}^d$, then $\mathbf{v}$ is sub-Gaussian and has independent components.
- Let $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times d}$. Let $\mathbf{X} = (\mathbf{X}_{0:k}\ \mathbf{X}_{k:\infty})$
- Let $\mathbf{X} = \left(\sqrt{\lambda_1}\mathbf{z}_1, \ldots, \sqrt{\lambda_d}\mathbf{z}_d\right) \in \mathbb{R}^{n\times d}$, then by Assumption 3.4.1, $\mathbf{z}_j$ is 1-sub-Gaussian and has independent components.
- Let $\widetilde{\mathbf{A}} := \mathbf{X}\mathbf{X}^\top = \sum_{i=1}^d \lambda_i \mathbf{z}_i\mathbf{z}_i^\top \in \mathbb{R}^{n\times n}$. Let $\mathbf{A} := \widetilde{\mathbf{A}} + \lambda_n\mathbf{I}_n = \mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n$.
- Let $\widetilde{\mathbf{A}}_k := \mathbf{X}_{k:\infty}\mathbf{X}_{k:\infty}^\top = \sum_{i\le k} \lambda_i \mathbf{z}_i\mathbf{z}_i^\top \in \mathbb{R}^{n\times n}$. Let $\mathbf{A}_k := \widetilde{\mathbf{A}}_k + \lambda\mathbf{I}_n = \mathbf{X}_{k:\infty}\mathbf{X}_{k:\infty}^\top + \lambda\mathbf{I}_n$.
- Let $\widetilde{\mathbf{A}}_{-j} := \sum_{i\neq j} \lambda_i \mathbf{z}_i\mathbf{z}_i^\top \in \mathbb{R}^{n\times n}$. Let $\mathbf{A}_{-j} := \widetilde{\mathbf{A}}_{-j} + \lambda\mathbf{I}_n$.
- Let $\rho_k := \frac{\lambda + \sum_{i>k}\lambda_i}{\lambda_{k+1}}$.
- Let $\mathbf{C} := \mathbf{A}^{-1}\mathbf{X}\mathbf{H}\mathbf{X}^\top\mathbf{A}^{-1}$.
- Let $\mathbf{B} := \left(\mathbf{I}_d - \mathbf{X}^\top\mathbf{A}^{-1}\mathbf{X}\right)\mathbf{H}\left(\mathbf{I}_d - \mathbf{X}^\top\mathbf{A}^{-1}\mathbf{X}\right)$.
- We use $\mathbb{E}_{\mathbf{X}}[\cdot]$ and $\mathbb{E}_{\boldsymbol{\epsilon}}[\cdot]$ to denote the expectation with respect to the randomness of drawing $\mathbf{X}$ and the randomness of noise, respectively.

Under the above notations and from [BLL20; TB20], we have

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\text{ridge error}] = \mathbb{E}_{\mathbf{X}}[\text{RidgeBias}] + \mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\text{RidgeVariance}],$$

where

$$\text{RidgeBias} := (\mathbf{w}^*)^\top \mathbf{B}\mathbf{w}^*, \qquad \text{RidgeVariance} := \boldsymbol{\epsilon}^\top \mathbf{C}\boldsymbol{\epsilon}.$$

We next provide lower bounds for each terms respectively.

Lemma 3.8.1 (Variant of Lemma 10 in [BLL20])**.** There are constants $b, c \geq 1$ such that for every $k \geq 0$, with probability at least 0.1,

1. for all $i \geq 1$,

$$\mu_{k+1}(\mathbf{A}_{-i}) \leq \mu_{k+1}(\mathbf{A}) \leq \mu_1(\mathbf{A}_k) \leq c\left(\lambda + \sum_{j>k} \lambda_j + \lambda_{k+1} n\right),$$

2. for all $1 \leq i \leq k$,

$$\frac{1}{c}\left(\lambda + \sum_{j>k} \lambda_j\right) - c\lambda_{k+1} n \leq \mu_n(\mathbf{A}_k) \leq \mu_n(\mathbf{A}_{-i}) \leq \mu_n(\mathbf{A}),$$

3. if $\rho_k \geq bn$, then

$$\frac{1}{c}\lambda_{k+1}\rho_k \leq \mu_n(\mathbf{A}_k) \leq \mu_1(\mathbf{A}_k) \leq c\lambda_{k+1}\rho_k.$$

4. if $\rho_k \geq bn$, then for all $i > k$,

$$\mu_n(\mathbf{A}_{-i}) \geq \frac{1}{c}\lambda_{k+1}\rho_k$$

Proof. The first two claims are proved by noticing that $\mathbf{A} = \lambda\mathbf{I} + \widetilde{\mathbf{A}}$, $\mathbf{A}_k = \lambda\mathbf{I} + \widetilde{\mathbf{A}}_k$, $\mathbf{A}_{-i} = \lambda\mathbf{I} + \widetilde{\mathbf{A}}_{-i}$, and applying Lemma 10 in [BLL20] to $\widetilde{\mathbf{A}}, \widetilde{\mathbf{A}}_k, \widetilde{\mathbf{A}}_{-j}$.

The third claim is proved by using the first two claims and that $\rho_k \geq bn$ to obtain that

$$\mu_1(\mathbf{A}_k) \leq c\left(\lambda + \sum_{i>k} \lambda_i + \lambda_{k+1} n\right) \leq \left(c + \frac{c}{b}\right) \cdot \left(\lambda + \sum_{i>k} \lambda_i\right),$$
$$\mu_n(\mathbf{A}_k) \geq \frac{1}{c}\left(\lambda + \sum_{i>k} \lambda_i\right) - c\lambda_{k+1} n \geq \left(\frac{1}{c} - \frac{c}{b}\right) \cdot \left(\lambda + \sum_{i>k} \lambda_i\right),$$

and by re-scaling the constants.

The fourth claim is used in Lemma 3 in [TB20], which can be proved under Assumption 3.4.1 as follows. Let $i > k$ and $\widetilde{\mathbf{A}}_{k,-i} = \sum_{j>k, j\neq i} \lambda_j \mathbf{z}_j \mathbf{z}_j^\top$. Then by Lemma 10 in [BLL20] there is an absolute constant $c \geq 1$ such that

$$\mu_n(\widetilde{\mathbf{A}}_{-i}) \geq \mu_n(\widetilde{\mathbf{A}}_{k,-i}) \geq \frac{1}{c} \sum_{j>k, j\neq i} \lambda_j - c\lambda_{k+1} n$$

holds with probability at least $1 - 2e^{-n/c}$, which yields

$$\mu_n(\mathbf{A}_{-i}) \geq \lambda + \frac{1}{c} \sum_{j>k, j\neq i} \lambda_i - c\lambda_{k+1} n \geq \lambda + \frac{1}{2c} \sum_{j>k} \lambda_j - \left(c + \frac{1}{c}\right) \lambda_{k+1} n,$$

where the last inequality is because: (1) $\sum_{j>k, j\neq i} \lambda_j \geq \frac{1}{2} \sum_{j>k} \lambda_j$ if $i > k+1$, and (2) $\sum_{j>k, j\neq i} \lambda_j = \sum_{j>k} \lambda_j - \lambda_{k+1}$ if $i = k+1$. Finally, using the condition that $\rho_k \geq bn$ we obtain that for $i > k$,

$$\mu_n(\mathbf{A}_{-i}) \geq \lambda + \frac{1}{2c} \sum_{j>k} \lambda_j - (c + \frac{1}{c}) \lambda_{k+1} n \geq \left(\frac{1}{2c} - \frac{c}{b} - \frac{1}{cb}\right) \cdot \left(\lambda + \sum_{j>k} \lambda_j\right),$$

which completes the proof by letting $b > 4c^2$ and $c \geq 1$

□

Variance Lower Bounds. According to Lemma 7 in [BLL20], and note that $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$, has zero mean, and is σ-sub-Gaussian, we have that

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\text{RidgeVariance}] = \mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}^\top \mathbf{C} \boldsymbol{\epsilon}] = \operatorname{tr}\left(\mathbf{C} \cdot \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top]\right) \geq \frac{1}{c} \sigma^2 \operatorname{tr}(\mathbf{C}) \tag{3.8.1}$$

for some constant $c > 1$. In the following we lower bound $\operatorname{tr}(\mathbf{C})$.

Lemma 3.8.2 (Variant of Lemma 8 in [BLL20])**.**

$$\operatorname{tr}(\mathbf{C}) = \sum_i \lambda_i^2 \mathbf{z}_i^\top \mathbf{A}^{-2} \mathbf{z}_i = \sum_i \frac{\lambda_i^2 \mathbf{z}_i^\top \mathbf{A}_{-i}^{-2} \mathbf{z}_i}{\left(1 + \lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1} \mathbf{z}_i\right)^2}.$$

Proof. This is from the proof of Lemma 14 in [TB20], and can be proved in the same way as Lemma 8 in [BLL20]. □

Lemma 3.8.3 (Variant of Lemma 14 in [BLL20])**.** There is a constant c such that for any $i \geq 1$ with $\lambda_i > 0$, and any $0 \leq k \leq n/c$, with probability at least 0.1,

$$\frac{\lambda_i^2 \mathbf{z}_i^\top \mathbf{A}_{-i}^{-2} \mathbf{z}_i}{\left(1 + \lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1} \mathbf{z}_i\right)^2} \geq \frac{1}{cn} \cdot \left(1 + \frac{\lambda_{k+1}}{\lambda_i} \cdot \left(1 + \frac{\rho_k}{n}\right)\right)^{-2}$$

Proof. Let $\mathcal{L}_i$ be a random subspace if $\mathbb{R}^n$ of codimension k, then

$$\begin{aligned}
\mathbf{z}_i^\top \mathbf{A}_{-i}^{-1} \mathbf{z}_i &\geq \frac{1}{c_1} \cdot \frac{\|\Pi_{\mathcal{L}_i} \mathbf{z}_i\|_2^2}{\lambda + \sum_{j>k} \lambda_j + \lambda_{k+1} n} && \text{(by Lemma 3.8.1)} \\
&\geq \frac{1}{c_2} \cdot \frac{n}{\lambda + \sum_{j>k} \lambda_j + \lambda_{k+1} n} && \text{(by Corollary 13 in [BLL20])} \\
&= \frac{1}{c_2} \cdot \frac{n}{\lambda_{k+1}(\rho_k + n)},
\end{aligned}$$

where $c_1, c_2 > 1$ are constants. The above implies that

$$\begin{aligned}
\frac{\lambda_i^2 \mathbf{z}_i^\top \mathbf{A}_{-i}^{-2} \mathbf{z}_i}{\left(1 + \lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1} \mathbf{z}_i\right)^2} &= \left(1 + \left(\lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1} \mathbf{z}_i\right)^{-1}\right)^{-2} \cdot \frac{\left\|\mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\right\|_2^2}{\left(\mathbf{z}_i^\top \mathbf{A}_{-i}^{-1} \mathbf{z}_i\right)^2} \\
&\geq \left(1 + \left(\lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1} \mathbf{z}_i\right)^{-1}\right)^{-2} \cdot \frac{1}{\|\mathbf{z}_i\|_2^2} \quad \text{(by Cauchy-Schwarz's inequality)} \\
&\geq \left(1 + c_2 \cdot \frac{\lambda_{k+1}(\rho_k + n)}{n \lambda_i}\right)^{-2} \cdot \frac{1}{\|\mathbf{z}_i\|_2^2}. \text{ (by the lower bound for } \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1} \mathbf{z}_i)
\end{aligned}$$

According to Corollary 13 in [BLL20], there is constant $c_3 > 1$ such that $\|\mathbf{z}_i\|_2^2 \leq \frac{1}{c_3} n$ holds with constant probability, inserting which into the above inequality and rescaling the constants complete the proof. □

Lemma 3.8.4 (Variant of Lemma 16 in [BLL20])**.** There is constant c such that for any $0 \leq k \leq n/c$ and any $b > 1$ with probability at least 0.1,

- if $\rho_k < bn$, then $\operatorname{tr}(\mathbf{C}) \geq \frac{k+1}{cb^2 n}$;
- if $\rho_k \geq bn$, then $\operatorname{tr}(\mathbf{C}) \geq \frac{1}{cb^2} \min_{\ell \leq k} \left\{ \frac{\ell}{n} + \frac{b^2 n \sum_{i>\ell} \lambda_i^2}{(\lambda_{k+1} \rho_k)^2} \right\}$.

Proof. This is proved by repeating the proof of Lemma 16 in [BLL20], where we replace Lemmas 8 and 14 in [BLL20] with our Lemmas 3.8.2 and 3.8.3 respectively. □

Theorem 3.8.5 (Restatement of Theorem 3.7.2, variance part)**.** There exist absolute constants $b, c, c_1 > 1$ for the following to hold: let

$$k^* := \min\{k : \lambda + \sum_{i>k} \lambda_i \geq bn\lambda_{k+1}\},$$

then with probability at least 0.1:

- if $k^* \geq n/c_1$ then
$$\mathbb{E}_{\boldsymbol{\epsilon}}[\text{RidgeVariance}] \geq \frac{\sigma^2}{c};$$
- if $k^* < n/c_1$ then
$$\mathbb{E}_{\boldsymbol{\epsilon}}[\text{RidgeVariance}] \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{\lambda + \sum_{i>k^*}\lambda_i} \cdot \sum_{i>k^*} \lambda_i^2\right).$$

As a direct consequence, the expected ridge variance is lower bounded by

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}}[\text{RidgeVariance}] \geq \begin{cases} \frac{\sigma^2}{10c}, & k^* \geq n/c_1 \\ \frac{\sigma^2}{10c}\left(\frac{k^*}{n} + \frac{n}{\lambda+\sum_{i>k^*}\lambda_i} \cdot \sum_{i>k^*} \lambda_i^2\right), & k^* < n/c_1. \end{cases}$$

Proof. The high probability lower bound is proved by (3.8.1), our Lemma 3.8.4, and Lemma 17 in [BLL20]. The expectation lower bound follows immediately from the high probability lower bound by noticing the ridge variance error is non-negative. □

Bias Lower Bound. Recall the ridge bias error is [TB20]

$$\text{RidgeBias} = (\mathbf{w}^*)^\top \mathbf{B}\mathbf{w}^* = \sum_i (\mathbf{B})_{ii} (\mathbf{w}_i^*)^2 + 2\sum_{i>j} (\mathbf{B})_{ij} \mathbf{w}_i^* \mathbf{w}_j^*. \tag{3.8.2}$$

The following lemma shows the crossing terms are zero in expectation.

Lemma 3.8.6. For $i \neq j$,

$$\mathbb{E}_{\mathbf{X}}[(\mathbf{B})_{ij}] = 0.$$

Proof. Recall that

$$\mathbf{B} := \left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right) \mathbf{H} \left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right).$$

Recall that $\mathbf{X} = \left(\sqrt{\lambda_1}\mathbf{z}_1, \dots \sqrt{\lambda_d}\mathbf{z}_d\right)$, thus the i-th column of $\left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right)$ is

$$\left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right)_i = \mathbf{e}_i - \sqrt{\lambda_i}\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{z}_i.$$

Moreover recall $\mathbf{H} = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, therefore

$$\begin{aligned}
(\mathbf{B})_{ij} = \mathbf{e}_i^\top \mathbf{B}\mathbf{e}_j &= \left(\mathbf{e}_i - \sqrt{\lambda_i}\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{z}_i\right)^\top \mathbf{H} \left(\mathbf{e}_j - \sqrt{\lambda_j}\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{z}_j\right) \\
&= \mathbf{e}_i^\top \mathbf{H}\mathbf{e}_j - \sqrt{\lambda_i}\mathbf{e}_j^\top \mathbf{H}\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{z}_i - \sqrt{\lambda_j}\mathbf{e}_i^\top \mathbf{H}\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{z}_j + \sqrt{\lambda_i\lambda_j}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{X}\mathbf{H}\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{z}_j \\
&= \mathbf{e}_i^\top \mathbf{H}\mathbf{e}_j - \left(\sqrt{\lambda_i\lambda_j}\lambda_j + \sqrt{\lambda_i\lambda_j}\lambda_i\right)\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_j + \sqrt{\lambda_i\lambda_j}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{X}\mathbf{H}\mathbf{X}^\top \mathbf{A}^{-1}\mathbf{z}_j.
\end{aligned}$$

The first term is zero since $\mathbf{H}$ is diagonal and $i \neq j$. We next show the second term is zero in expectation. Indeed, let

$$F(\mathbf{z}_i) := \mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_j = \mathbf{z}_i^\top \left(\mathbf{A}_{-i} + \lambda_i \mathbf{z}_i \mathbf{z}_i^\top\right)^{-1} \mathbf{z}_j,$$

where $\mathbf{A}_{-i}$ is independent of $\mathbf{z}_i$, then $F(\mathbf{z}_i) = -F(-\mathbf{z}_i)$. Also note that $\mathbf{z}_i$ follows a standard Gaussian which is symmetric, therefore $\mathbb{E}_{\mathbf{z}_i} F(\mathbf{z}_i) = 0$. In a similar manner, the third term is also zero in expectation. The proof is then completed.

□

Lemma 3.8.7 (Part of the proof of Lemma 15 in [TB20])**.** There exists absolute constant $c > 1$, such that with probability at least 0.1,

$$(\mathbf{B})_{ii} \geq \frac{1}{c} \cdot \frac{\lambda_i}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{\rho_k}\right)^2}.$$

As a direct consequence,

$$\mathbb{E}_{\mathbf{X}}[(\mathbf{B})_{ii}] \geq \frac{1}{10c} \cdot \frac{\lambda_i}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{\rho_k}\right)^2}.$$

Proof. This lemma summarizes part of the proof of Lemma 15 in [TB20]. Recall that $\mathbf{H}$ is diagonal and $\mathbf{B} := \left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right)\mathbf{H}\left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right)$, thus

$$\begin{aligned}
(\mathbf{B})_{ii} &= \lambda_i \left\|\left(\mathbf{I}_d - \mathbf{X}^\top \mathbf{A}^{-1}\mathbf{X}\right)_i\right\|_2^2 \qquad \text{(since } \mathbf{H} \text{ is diagonal)} \\
&= \lambda_i \left\|e_i^\top - \sqrt{\lambda_i}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{X}\right\|_2^2 \qquad \left(\mathbf{X} = \left(\sqrt{\lambda_1}\mathbf{z}_1, \ldots, \sqrt{\lambda_j}\mathbf{z}_j, \ldots \sqrt{\lambda_d}\mathbf{z}_d\right)\right) \\
&= \lambda_i \left\|e_i^\top - \left(\sqrt{\lambda_i\lambda_1}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_1, \ldots, \sqrt{\lambda_i\lambda_j}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_j, \ldots \sqrt{\lambda_i\lambda_d}\mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_d\right)\right\|_2^2 \\
&\geq \lambda_i \left(1 - \lambda_i \mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_i\right)^2 \qquad \text{(use Pythagorean theorem)} \\
&= \frac{\lambda_i}{\left(1 + \lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^2},
\end{aligned}$$

where in the last step we use $\mathbf{A} = \mathbf{A}_{-i} + \lambda_i \mathbf{z}_i \mathbf{z}_i^\top$ and that

$$\begin{aligned}
1 - \lambda_i \mathbf{z}_i^\top \mathbf{A}^{-1}\mathbf{z}_i &= 1 - \lambda_i \mathbf{z}_i^\top \left(\mathbf{A}_{-i} + \lambda_i \mathbf{z}_i \mathbf{z}_i^\top\right)^{-1} \mathbf{z}_i \\
&= 1 - \lambda_i \mathbf{z}_i^\top \left(\mathbf{A}_{-i}^{-1} - \lambda_i \mathbf{A}_{-i}^{-1}\mathbf{z}_i (1 + \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i)^{-1} \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\right) \mathbf{z}_i \\
&= \frac{1}{1 + \lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i}.
\end{aligned}$$

Now according to Corollary 13 in [BLL20], there exists constant $c_1 > 1$ such that

$$\|\mathbf{z}_i\|_2^2 \leq c_1 n$$

holds with constant probability; and according to Lemma 3.8.1, there exists constant $c_2 > 1$ such that for any $i \geq 1$,

$$\mu_n(\mathbf{A}_{-i}) \geq \frac{1}{c_2}\lambda_{k+1}\rho_k$$

holds with constant probability. These two facts imply that

$$\mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i \leq \mu_n(\mathbf{A}_{-i})^{-1} \|\mathbf{z}_i\|_2^2 \leq c_1 c_2 \frac{n}{\lambda_{k+1}\rho_k},$$

inserting which into the bound of $(\mathbf{B})_{ii}$, we conclude that with constant probability,

$$(\mathbf{B})_{ii} \geq \frac{\lambda_i}{\left(1 + \lambda_i \mathbf{z}_i^\top \mathbf{A}_{-i}^{-1}\mathbf{z}_i\right)^2} \geq \frac{\lambda_i}{\left(1 + c_1 c_2 \cdot \frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{\rho_k}\right)^2}.$$

Finally a rescaling of the constants completes the proof. □

Theorem 3.8.8 (Restatement of Theorem 3.7.2, bias part)**.** There exist absolute constants $b, c > 1$ for the following to hold: let

$$k^* := \min\{k : \lambda + \sum_{i>k} \lambda_i \geq bn\lambda_{k+1}\},$$

then

$$\mathbb{E}_{\mathbf{X}}[\text{RidgeBias}] \geq \frac{1}{c}\left(\frac{\lambda + \sum_{i>k^*} \lambda_i}{n^2} \cdot \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right).$$

Proof. By (3.8.2), Lemmas 3.8.6 and 3.8.7, we have that,

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\text{RidgeBias}] &= \sum_i (\mathbf{B})_{ii} (\mathbf{w}_i^*)^2 \\
&\geq \frac{1}{c_1} \sum_i \frac{1}{\left(1 + \frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}}\right)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2 \qquad (\text{choose } k = k^*) \\
&\geq \frac{1}{c_1 b^2} \sum_i \frac{1}{\left(\frac{1}{b} + \frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}}\right)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2,
\end{aligned}$$

where $c_1, b > 1$ are all absolute constants. Note that for all $i \leq k^*$, we must have $\lambda + \sum_{j>i-1} \lambda_j < bn\lambda_i$,

$$\frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}} = \frac{\lambda_i n}{\lambda + \sum_{j>k^*} \lambda_j} \geq \frac{\lambda_i n}{\lambda + \sum_{j>i-1} \lambda_j} \geq \frac{1}{b},$$

and for all $i \geq k^* + 1$, we have

$$\frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}} \leq \frac{n}{\rho_{k^*}} \leq \frac{1}{b},$$

then

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\text{RidgeBias}] &\geq \frac{1}{c_1 b^2} \sum_i \frac{1}{\left(\frac{1}{b} + \frac{\lambda_i}{\lambda_{k^*+1}} \cdot \frac{n}{\rho_{k^*}}\right)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2 \\
&\geq \frac{1}{2c_1 b^2} \cdot \left(\sum_{i \leq k^*} \frac{1}{\left(\frac{\lambda_i}{\lambda_{k+1}} \cdot \frac{n}{\rho_k}\right)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2 + \sum_{i>k^*} \frac{1}{(1/b)^2} \cdot \lambda_i (\mathbf{w}_i^*)^2\right) \\
&\geq \frac{1}{c}\left(\sum_{i \leq k^*} \frac{(\lambda_{k+1}\rho_k)^2}{n^2} \cdot \lambda_i^{-1} (\mathbf{w}_i^*)^2 + \sum_{i>k^*} \lambda_i (\mathbf{w}_i^*)^2\right) \\
&= \frac{1}{c}\left(\frac{\lambda + \sum_{i>k^*} \lambda_i}{n^2} \cdot \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right),
\end{aligned}$$

where $c > 1$ is an absolute constant. $\square$

3.9 Conclusions

We conduct an instance-based risk comparison between SGD and ridge regression for a broad class of least square problems. We show that SGD is always no worse than ridge regression provided logarithmically more samples. On the other hand, there exist some instances where even optimally-tuned ridge regression needs quadratically more samples to compete with SGD. This separation in terms of sample inflation between SGD and ridge regression suggests a provable benefit of implicit regularization over explicit regularization for least squares problems. In the future, we will explore the benefits of implicit regularization for learning other linear models and potentially nonlinear models.

Part II

Learning Over-parameterized Neural Network Models

CHAPTER 4

Optimization of Over-parameterized Deep ReLU Networks

4.1 Introduction

In this chapter, we study the optimization properties of gradient-based methods for training deep ReLU neural networks, with more realistic assumption on the training data, milder over-parameterization condition and faster convergence rate, compared to existing works [LL18; ALS19a]. In specific, we consider an L-hidden-layer fully-connected neural network with ReLU activation function. We show that GD can achieve the global minima of the training loss for any $L \geq 1$, with the aid of over-parameterization and random initialization. The high-level idea of our proof technique is to show that Gaussian random initialization followed by gradient descent generates a sequence of iterates within a small perturbation region centering around the initial weights. In addition, we will show that the empirical loss function of deep ReLU networks has very good local curvature properties inside the perturbation region, which guarantees the global convergence of gradient descent. Compared with the proof technique in [ALS19a], we provide a sharper analysis on the GD algorithm and prove that GD can be guaranteed to have sufficient descent in a larger perturbation region with a larger step size. This leads to a faster convergence rate and a milder condition on the over-paramterization. More specifically, our main contributions are summarized as follows:

- We establish the global convergence guarantee for training deep ReLU networks in

terms of classification problems. Compared with [LL18; ALS19a] our assumption on training data is more reasonable and is often satisfied by real training data. Specifically, we only require that any two data points from different classes are separated by some constant, while [LL18] assumes that the data from different classes are sampled from small balls separated by a constant margin, and [ALS19a] requires that any two data points are well separated, even though they belong to the same class.

- We show that with Gaussian random initialization on each layer, when the number of hidden nodes per layer is at least $\widetilde{\Omega}\big(n^{14}L^{16}/\phi^4\big)$, GD can achieve zero training error within $\widetilde{O}\big(n^5L^3/\phi\big)$ iterations, where ϕ is the data separation distance[1], n is the number of training examples, and L is the number of hidden layers. This significantly improves the state-of-the-art results by [ALS19a], where the authors proved that GD can converge within $\widetilde{O}\big(n^6L^2/\phi^2\big)$ iterations if the number of hidden nodes per layer is at least $\widetilde{\Omega}(n^{24}L^{12}/\phi^8)$. Compared with [DLL19], our result only has a polynomial dependency on the number of hidden layers, which is much better than their result that has an exponential dependency on the depth for fully connected deep neural networks.

4.2 Additional Related Work

Due to the huge amount of literature on deep learning theory, we are not able to include all papers in this big vein here. Instead, we review the following two additional lines of research, which are also related to our work.

One-hidden-layer neural networks with ground truth parameters Recently a series of work [Tia17; BG17; LY17; DLT18; ZYW19] studied a specific class of shallow two-layer (one-hidden-layer) neural networks, whose training data are generated by a ground truth network called "teacher network". This series of work aim to provide recovery guarantee

[1]We will define the data separation distance, training sample size n and number of hidden layers L formally in Sections 4.3 and 4.4.

for gradient-based methods to learn the teacher networks based on either the population or empirical loss functions. More specifically, [Tia17] proved that for two-layer ReLU networks with only one hidden neuron, GD with arbitrary initialization on the population loss is able to recover the hidden teacher network. [BG17] proved that GD can learn the true parameters of a two-layer network with a convolution filter. [LY17] proved that SGD can recover the underlying parameters of a two-layer residual network in polynomial time. Moreover, [DLT18] proved that both GD and SGD can recover the teacher network of a two-layer CNN with ReLU activation function. [ZYW19] showed that GD on the empirical loss function can recover the ground truth parameters of one-hidden-layer ReLU networks at a linear rate.

Deep linear networks Beyond shallow one-hidden-layer neural networks, a series of recent work [HM16; Kaw16; BHL18; GLS18b; AGC19; ACH18] focused on the optimization landscape of deep linear networks. More specifically, [HM16] showed that deep linear residual networks have no spurious local minima. [Kaw16] proved that all local minima are global minima in deep linear networks. [ACH18] showed that depth can accelerate the optimization of deep linear networks. [BHL18] proved that with identity initialization and proper regularizer, GD can converge to the least square solution on a residual linear network with quadratic loss function, while [AGC19] proved the same properties for general deep linear networks.

4.3 Preliminaries

4.3.1 Problem Setup

Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \in (\mathbb{R}^d \times \{-1, 1\})^n$ be a set of n training examples. Let $m_0 = d$. We consider L-hidden-layer neural networks as follows:

$$f_{\mathbf{W}}(\mathbf{x}) = \mathbf{v}^\top \sigma(\mathbf{W}_L^\top \sigma(\mathbf{W}_{L-1}^\top \cdots \sigma(\mathbf{W}_1^\top \mathbf{x}) \cdots)),$$

where $\sigma(x) = \max\{0, x\}$ is the entry-wise ReLU activation function, $\mathbf{W}_l = (\mathbf{w}_{l,1}, \ldots, \mathbf{w}_{l,m_l}) \in \mathbb{R}^{m_{l-1} \times m_l}$, $l = 1, \ldots, L$ are the weight matrices, and $\mathbf{v} \in \{-1, +1\}^{m_L}$ is the fixed output layer weight vector with half 1 and half -1 entries. Let $\mathbf{W} = \{\mathbf{W}_l\}_{l=1,\ldots,L}$ be the collection of matrices $\mathbf{W}_1, \ldots, \mathbf{W}_L$, we consider solving the following empirical risk minimization problem:

$$L_S(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i \widehat{y}_i) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i \mathbf{v}^\top \sigma(\mathbf{W}_L^\top \sigma(\mathbf{W}_{L-1}^\top \cdots \sigma(\mathbf{W}_1^\top \mathbf{x}_i) \cdots))\big) \tag{4.3.1}$$

where $\widehat{y}_i = f_{\mathbf{W}}(\mathbf{x}_i)$ denotes the output of neural network and $\ell(x) = \log(1 + \exp(-x))$ is the cross-entropy loss for binary classification.

4.3.2 Optimization Algorithms

In this chapter, we consider training a deep neural network with Gaussian initialization followed by gradient descent.

Gaussian initialization. We say that the weight matrices $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated from Gaussian initialization if each column of $\mathbf{W}_l$ is generated independently from the Gaussian distribution $N(\mathbf{0}, 2/m_l \mathbf{I})$ for all $l = 1, \ldots, L$. This initialization mechanism is called He-initialization, which was proposed in [HZR15].

Gradient descent. We consider solving the empirical risk minimization problem (4.3.1) with gradient descent with Gaussian initialization: let $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ be weight matrices generated from Gaussian initialization, we consider the following gradient descent update rule:

$$\mathbf{W}_l^{(k)} = \mathbf{W}_l^{(k-1)} - \eta \nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(k-1)}), \ l = 1, \ldots, L,$$

where $\nabla_{\mathbf{W}_l} L_S(\cdot)$ is the partial gradient of $L_S(\cdot)$ with respect to the l-th layer parameters $\mathbf{W}_l$, and $\eta > 0$ is the step size (a.k.a., learning rate).

4.3.3 Calculations for Neural Network Functions

Here we briefly introduce some useful notations and provide some basic calculations regarding the neural network in our setting.

- **Output after the l-th layer:** Given an input $\mathbf{x}_i$, the output of the neural network after the l-th layer is

$$\begin{aligned}\mathbf{x}_{l,i} &= \sigma(\mathbf{W}_l^\top \sigma(\mathbf{W}_{l-1}^\top \cdots \sigma(\mathbf{W}_1^\top \mathbf{x}_i)\cdots)) \\ &= \bigg(\prod_{r=1}^{l} \boldsymbol{\Sigma}_{r,i}\mathbf{W}_r^\top\bigg)\mathbf{x}_i,\end{aligned}$$

  where $\boldsymbol{\Sigma}_{1,i} = \mathrm{Diag}\big(\mathbb{1}\{\mathbf{W}_1^\top \mathbf{x}_i > 0\}\big)$[2], and $\boldsymbol{\Sigma}_{l,i} = \mathrm{Diag}[\mathbb{1}\{\mathbf{W}_l^\top(\prod_{r=1}^{l-1}\boldsymbol{\Sigma}_{r,i}\mathbf{W}_r^\top)\mathbf{x}_i > 0\}]$ for $l = 2, \ldots, L$.

- **Output of the neural network:** The output of the neural network with input $\mathbf{x}_i$ is as follows:

$$\begin{aligned}f_{\mathbf{W}}(\mathbf{x}_i) &= \mathbf{v}^\top \sigma(\mathbf{W}_L^\top \sigma(\mathbf{W}_{L-1}^\top \cdots \sigma(\mathbf{W}_1^\top \mathbf{x}_i)\cdots)) \\ &= \mathbf{v}^\top \bigg(\prod_{r=l}^{L} \boldsymbol{\Sigma}_{r,i}\mathbf{W}_r^\top\bigg)\mathbf{x}_{l-1,i},\end{aligned}$$

  where we define $\mathbf{x}_{0,i} = \mathbf{x}_i$ and the last equality holds for any $l \geq 1$.

- **Gradient of the neural network:** The partial gradient of the training loss $L_S(\mathbf{W})$ with respect to $\mathbf{W}_l$ is as follows:

$$\nabla_{\mathbf{W}_l} L_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i\widehat{y}_i)\cdot y_i \cdot \nabla_{\mathbf{W}_l}[f_{\mathbf{W}}(\mathbf{x}_i)],$$

  where the gradient of the neural network function is defined as

$$\nabla_{\mathbf{W}_l}[f_{\mathbf{W}}(\mathbf{x}_i)] = \mathbf{x}_{l-1,i}\mathbf{v}^\top \bigg(\prod_{r=l+1}^{L} \boldsymbol{\Sigma}_{r,i}\mathbf{W}_r^\top\bigg)\boldsymbol{\Sigma}_{l,i}.$$

[2]Here we slightly abuse the notation and denote $\mathbb{1}\{\mathbf{a} > 0\} = (\mathbb{1}\{\mathbf{a}_1 > 0\}, \ldots, \mathbb{1}\{\mathbf{a}_m > 0\})^\top$ for a vector $\mathbf{a} \in \mathbb{R}^m$.

In the remaining of this paper, we define the gradient $\nabla L_S(\mathbf{W})$ as the collection of partial gradients with respect to all $\mathbf{W}_l$'s, i.e.,

$$\nabla L_S(\mathbf{W}) = \{\nabla_{\mathbf{W}_1} L_S(\mathbf{W}), \nabla_{\mathbf{W}_2} L_S(\mathbf{W}), \ldots, \nabla_{\mathbf{W}_L} L_S(\mathbf{W})\}.$$

We also define the Frobenius norm of $\nabla L_S(\mathbf{W})$ as

$$\|\nabla_{\mathbf{W}_l} L_S(\mathbf{W})\|_F = \left[\sum_{l=1}^{L} \|\nabla_{\mathbf{W}_l} L_S(\mathbf{W})\|_F^2\right]^{1/2}.$$

4.4 Main Theory

In this section, we show that with random Gaussian initialization, over-parameterization helps gradient descent converge to the global minimum, i.e., find a point in the parameter space with arbitrary small training loss. We start with assumptions on the training data,

Assumption 4.4.1. $\|\mathbf{x}_i\|_2 = 1$ and $(\mathbf{x}_i)_d = \mu$ for all $i \in \{1, \ldots, n\}$, where $\mu \in (0, 1)$ is a constant.

As is shown in the assumption above, the last entry of input $\mathbf{x}$ is considered to be a constant μ. This assumption is natural because it can be seen as adding a bias term in the input layer, and learning both weight vector and bias is equivalent to adding an additional dummy variable ($(\mathbf{x}_i)_d = \mu$) to all input vectors and learning the weight vector only. The same assumption has been made in [ALS19a]. In addition, we emphasize that Assumption 4.4.1 is made in order to simplify the proof. Actually, rather than restricting the norm of all training examples to be 1, this assumption can be relaxed to be that $\|\mathbf{x}_i\|_2$ is lower and upper bounded by some constants.

Assumption 4.4.2. For all $i, i' \in \{1, \ldots, n\}$, if $y_i \neq y_{i'}$, then $\|\mathbf{x}_i - \mathbf{x}_{i'}\|_2 \geq \phi$ for some $\phi > 0$.

Assumption 4.4.2 basically requires that inputs with different labels in the training data are separated from each other by at least a constant. This assumption is often satisfied in

practice. In contrast, [ALS19a] assumes that every two different data points in the training data are separated by a constant, which is much stronger and cannot be satisfied since in classification it is allowed that the data with the same label can be arbitrarily close.

Furthermore, Assumption 4.4.2 can be easily verified based on the training data. As a comparison, the assumption made in [DLL19] assumes that certain deep compositional Gram matrix defined on the training data is strictly positive definite, which is not easy to verify, since the definition of their special Gram matrix is based on integration.

Then we have the following assumption on the structure of neural network.

Assumption 4.4.3. Define $M = \max\{m_1, \ldots, m_L\}$, $m = \min\{m_1, \ldots, m_L\}$. We assume that $M \leq 2m$.

Assumption 4.4.3 states that the number of nodes at all layers are of the same order. The constant 2 is not essential and can be replaced with an arbitrary constant greater than or equal to 1.

Under Assumptions 4.4.1-4.4.3, we are able to establish the global convergence of gradient descent for training deep ReLU networks. Specifically, we provide the following theorem which characterizes the required numbers of hidden nodes and iterations such that the gradient descent can attain the global minimum of the training loss function.

Theorem 4.4.4. Suppose $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ are generated by Gaussian initialization. Then under Assumptions 4.4.1-4.4.3, if the step size $\eta = O(M^{-1}L^{-3})$, the number of hidden nodes per layer satisfies

$$m = \widetilde{\Omega}\big(n^{14}L^{16}\phi^{-4} + n^{12}L^{16}\phi^{-4}\epsilon^{-1}\big)$$

and the maximum number of iteration satisfies

$$K = \widetilde{O}\big(n^5L^3/\phi + n^3L^3\epsilon^{-1}/\phi\big),$$

then with high probability, the last iterate of gradient descent $\mathbf{W}^{(K)}$ satisfies $L_S(\mathbf{W}^{(K)}) \leq \epsilon$.

Remark 4.4.5. Note that our bound on the required number of hidden nodes per layer, i.e., m, depends on the target accuracy ϵ. However, in practical classification tasks, we are more interested in finding some points with zero training error. In specific, the cross-entropy loss $\ell(x) = \log(1 + \exp(-x))$ is strictly decreasing in x, thus $\ell(y_i\widehat{y}_i) \leq \ell(0) = \log(2)$ implies $y_i\widehat{y}_i \geq 0$. If we set $L_S(\mathbf{W}) \leq \ell(0)/n = \log(2)/n$, it holds that $\ell(y_i\widehat{y}_i) \leq nL_S(\mathbf{W}) \leq \ell(0)$ for all $i \in [n]$, which further implies that $y_i\widehat{y}_i \geq 0$ for all $i \in [n]$, i.e., all training data are correctly classified. Therefore, Theorem 4.4.4 implies that gradient descent can find a point with zero training error if the number of hidden nodes per layer is at least $m = \widetilde{\Omega}(n^{14}L^{16}\phi^{-4})$.

Remark 4.4.6. Here we compare our theoretical results with those in [ALS19a] and [DLL19]. Specifically, [ALS19a] proves that gradient descent can achieve zero training error within $O(n^6L^2/\phi^2)$ iterations under the condition that the neural network width is at least $m = \widetilde{\Omega}(n^{24}L^{12}/\phi^8)$. As a clear comparison, our result on m is significantly better by a factor of $\widetilde{\Omega}(n^{10}L^{-4}/\phi^4)$, and our convergence rate is faster by a factor of $O(nL^{-1})$[3]. On the other hand, [DLL19] proved similar global convergence result when the neural network width is at least $\widetilde{\Omega}\big(2^{O(L)} \cdot n^4/\lambda_0^4\big)$, where λ_0 is the smallest eigenvalue of the deep compositional Gram matrix defined in their paper. Compared with their result, our condition on m has significantly better dependency in L. In addition, for real training data, λ_0 can have high degree dependency on the reciprocal of the sample size n, which makes the dependency of their result on n much worse.

4.5 Proof of the Main Theory

In this section, we provide the proof of the main theory. In specific, we decompose the proof into three steps:

Step 1: We characterize a perturbation region at the initialization, and prove that the

[3]It is worth noting that in practice we usually have $n \gg L$, thus our improvements in terms of the over-parameterization condition and convergence rate are indeed significant.

neural network attains good properties within such region.

Step 2: Based on the assumption that all iterates are staying inside the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$, we establish the convergence results of gradient descent.

Step 3: We verify that with our choice of m, until convergence all iterates of gradient descent would not escape from the perturbation region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$, which justifies the derived convergence guarantee.

Now we characterize the perturbation as follows. Given the initialization generated by Gaussian distribution $\mathbf{W}^{(0)} := \{\mathbf{W}_l^{(0)}\}_{l=1,\ldots,L}$, we define by $\mathcal{B}(\mathbf{W}^{(0)}, \tau) = \{\mathbf{W} : \|\mathbf{W}_l - \mathbf{W}_l^{(0)}\|_2 \leq \tau \text{ for all } l \in [L]\}$ the perturbation region centered at $\mathbf{W}^{(0)}$. Then we provide the following Lemmas that provides key results which are essential to establish the convergence guarantees for (stochastic) gradient descent.

Lemma 4.5.1 (Bounded initial training loss)**.** Under Assumptions 4.4.1 and 4.4.3, with probability at least $1 - \delta$, at the initialization the training loss satisfies $L_S(\mathbf{W}^{(0)}) \leq C\sqrt{\log(n/\delta)}$.

Next we are going to state the following key lemmas that characterizes some essential properties of the neural network when its weight parameters satisfies $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Firstly, the following lemma that provides the lower and upper bounds of the Frobenious norm of the partial gradient $\nabla_{\mathbf{W}_l}[L_S(\mathbf{W})]$.

Lemma 4.5.2 (Gradient lower and upper bound)**.** Under Assumptions 4.4.1, 4.4.2, and 4.4.3, if $\tau = O\big(\phi^{3/2} n^{-3} L^{-2}\big)$ and $m = \widetilde{\Omega}(n^2 \phi^{-1})$, then for all $\widetilde{\mathbf{W}} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, with probability at least $1 - \exp\big(-O(m\phi/n)\big)$, there exist positive constants C and C' such that

$$\|\nabla_{\mathbf{W}_L}[L_S(\widetilde{\mathbf{W}})]\|_F^2 \geq C\frac{m\phi}{n^5}\bigg(\sum_{i=1}^n \ell'(y_i \widetilde{y}_i)\bigg)^2,$$

$$\|\nabla_{\mathbf{W}_l}[L_S(\widetilde{\mathbf{W}})]\|_F \leq -\frac{C' L M^{1/2}}{n}\sum_{i=1}^n \ell'(y_i \widetilde{y}_i),$$

for all $l \in [L]$, where $\widetilde{y}_i = f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)$.

Then we provide the following lemma that characterizes the training loss decreasing after one-step gradient descent.

Lemma 4.5.3 (Sufficient Descent)**.** Let $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ be generated via Gaussian random initialization. Let $\mathbf{W}^{(k)} = \{\mathbf{W}_l^{(k)}\}_{l=1,\ldots,L}$ be the k-th iterate in the gradient descent and $\tau = O(L^{-11}\log^{-3/2}(M))$. If $\mathbf{W}^{(k)}, \mathbf{W}^{(k+1)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, then there exist constants C' and C'' such that with probability at least $1 - \exp\big(-O(m\phi/n)\big)$ the following holds,

$$L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) \leq -\big(\eta - C'ML^3\eta^2\big)\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 - \frac{C''L^{8/3}\tau^{1/3}\sqrt{M\log(M)}\cdot\eta\|\nabla L_S(\mathbf{W}^{(k)})\|_F}{n}\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})$$

The second term on the R.H.S. of the result in Lemma 4.5.3 is due to the non-smoothness of ReLU activation, which can be characterized by counting how many nodes would change their activation patterns during the training process. Clearly, in order to guarantee that the gradient descent can bring sufficient descent in each step, we require the radius τ to be sufficiently small. In the following, we are going to complete the proof of Theorem 4.4.4 based on Lemmas 4.5.1-4.5.3 .

Proof of Theorem 4.4.4. We first prove that GD is able to achieve ϵ training loss under the condition that all iterates are staying inside the perturbation region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Note that by Lemma 4.5.2, we know that there exists a constant c_0 such that

$$\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 \geq \|\nabla_{\mathbf{W}_L}[L_S(\mathbf{W}^{(k)})]\|_F^2 \geq \frac{c_0 m\phi}{n^5}\bigg(\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})\bigg)^2.$$

We set the radius τ and the step size η as follows,

$$\tau = \bigg(\frac{c_0^{1/2}m^{1/2}\phi^{1/2}}{4C''L^{8/3}n^{3/2}\sqrt{M\log(M)}}\bigg)^3 = \widetilde{O}(n^{-9/2}L^{-8}\phi^{3/2}),$$
$$\eta = \frac{1}{4C'ML^3} = O(M^{-1}L^{-3}).$$

Then we have

$$
\begin{aligned}
&L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) \\
&\leq -\frac{3\eta}{4}\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 - \frac{c_0\eta m^{1/2}\phi^{1/2}}{4n^{5/2}}\|\nabla L_S(\mathbf{W}^{(k)})\|_F \cdot \sum_{i=1}^n \ell'(y_i\widehat{y}_i^{(k)}) \\
&\leq -\frac{\eta}{2}\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 \\
&\leq -\eta\frac{c_0 m\phi}{2n^5}\bigg(\sum_{i=1}^n \ell'(y_i\widehat{y}_i^{(k)})\bigg)^2, \qquad (4.5.1)
\end{aligned}
$$

where the first inequality is by Lemma 4.5.3 and the choices of η and τ, the second inequality follows from Lemma 4.5.2, and the last inequality is due to the gradient lower bound we derived above. Note that $\ell(x) = \log(1+\exp(-x))$, which satisfies $-\ell'(x) = 1/(1+\exp(x)) \geq \min\big\{\alpha_0, \alpha_1\ell(x)\big\}$ where $\alpha_0 = 1/2$ and $\alpha_1 = 1/(2\log(2))$. This implies that

$$
-\sum_{i=1}^n \ell'(y_i\widehat{y}_i^{(k)}) \geq \min\bigg\{\alpha_0, \sum_{i=1}^n \alpha_1\ell(y_i\widehat{y}_i^{(k)})\bigg\} \geq \min\big\{\alpha_0, n\alpha_1 L_S(\mathbf{W}^{(k)})\big\}.
$$

Note that $\min\{a,b\} \geq 1/(1/a+1/b)$, we have the following by plugging the above inequality into (4.5.1)

$$
\begin{aligned}
L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) &\leq -\eta\min\bigg\{\frac{c_0 m\phi\alpha_0^2}{2n^5}, \frac{c_0 m\phi\alpha_1^2}{2n^3}L_S^2(\mathbf{W}^{(k)})\bigg\} \\
&\leq -\eta\bigg(\frac{2n^5}{c_0 m\phi\alpha_0^2} + \frac{2n^3}{c_0 m\phi\alpha_1^2 L_S^2(\mathbf{W}^{(k)})}\bigg)^{-1}.
\end{aligned}
$$

Rearranging terms gives

$$
\frac{2n^5}{c_0 m\phi\alpha_0^2}\big(L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)})\big) + \frac{2n^3\big(L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)})\big)}{c_0 m\phi\alpha_1^2 L_S^2(\mathbf{W}^{(k)})} \leq -\eta. \qquad (4.5.2)
$$

Applying the inequality $(x-y)/y^2 \geq y^{-1} - x^{-1}$ and taking telescope sum over k give

$$
\begin{aligned}
k\eta &\leq \frac{2n^5}{c_0 m\phi\alpha_0^2}\big(L_S(\mathbf{W}^{(0)}) - L_S(\mathbf{W}^{(k)})\big) + \frac{2n^3\big(L_S^{-1}(\mathbf{W}^{(k)}) - L_S^{-1}(\mathbf{W}^{(0)})\big)}{c_0 m\phi\alpha_1^2} \\
&\leq \frac{2n^5}{c_0 m\phi\alpha_0^2}L_S(\mathbf{W}^{(0)}) + \frac{2n^3\big(L_S^{-1}(\mathbf{W}^{(k)}) - L_S^{-1}(\mathbf{W}^{(0)})\big)}{c_0 m\phi\alpha_1^2}. \qquad (4.5.3)
\end{aligned}
$$

Now we need to guarantee that after K gradient descent steps the loss function $L_S(\mathbf{W}^{(K)})$ is smaller than the target accuracy ϵ. By Lemma 4.5.1, we know that the training loss

$L_S(\mathbf{W}^{(0)}) = \widetilde{O}(1)$. Therefore, by (4.5.3) and our choice of η, the maximum iteration number K satisfies

$$K = \widetilde{O}\big(n^5 L^3/\phi + n^3 L^3 \epsilon^{-1}/\phi\big). \tag{4.5.4}$$

Then we are going to verify the condition that all iterates stay inside the perturbation region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$. We prove this by induction. Clearly, $\mathbf{W}^{(0)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Then we are going to prove $\mathbf{W}^{(k+1)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ under the induction hypothesis that $\mathbf{W}^{(t)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ holds for all $t \leq k$. According to (4.5.1), we have

$$L_S(\mathbf{W}^{(t+1)}) - L_S(\mathbf{W}^{(t)}) \leq -\frac{\eta}{2}\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2, \tag{4.5.5}$$

for any $t < k$. Therefore, by triangle inequality, we have

$$\begin{aligned}
\|\mathbf{W}_l^{(k)} - \mathbf{W}_l^{(0)}\|_2 &\leq \eta \sum_{t=0}^{k-1} \big\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(t)})]\big\|_2 \\
&\leq \eta \sqrt{k \sum_{t=0}^{k-1} \big\|\nabla L_S(\mathbf{W}^{(t)})\big\|_F^2} \\
&\leq \sqrt{2k\eta \sum_{t=0}^{k-1} \big[L_S(\mathbf{W}^{(t)}) - L_S(\mathbf{W}^{(t+1)})\big]} \\
&\leq \sqrt{2k\eta L_S(\mathbf{W}^{(0)})}.
\end{aligned}$$

By Lemma 4.5.1, we know that $L_S(\mathbf{W}^{(0)}) = \widetilde{O}(1)$. Then applying our choices of η and K, we have

$$\|\mathbf{W}_l^{(k)} - \mathbf{W}_l^{(0)}\|_2 \leq \sqrt{2K\eta L_S(\mathbf{W}^{(0)})} = \widetilde{O}\big(n^{5/2}\phi^{-1/2}m^{-1/2} + n^{3/2}\epsilon^{-1/2}\phi^{-1/2}m^{-1/2}\big).$$

In addition, by Lemma 4.5.2 and our choice of η, we have

$$\begin{aligned}
\eta\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\|_2 &\leq -\frac{\eta C' L M^{1/2}}{n} \sum_{i=1}^{n} \ell'\big(y_i \cdot f_{\mathbf{W}^{(k)}}(\mathbf{x}_i)\big) \\
&\leq \widetilde{O}(L^{-2}M^{-1/2}),
\end{aligned}$$

where the second inequality follows from the choice of η and the fact that $-1 \leq \ell'(\cdot) \leq 0$. Then by triangle inequality, we have

$$\begin{aligned}\|\mathbf{W}_l^{(k+1)} - \mathbf{W}_l^{(0)}\|_2 &\leq \eta \|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\|_2 + \|\mathbf{W}_l^{(k)} - \mathbf{W}_l^{(0)}\|_2 \\ &= \widetilde{O}(n^{-9/2}L^{-8}\phi^{3/2}),\end{aligned}$$

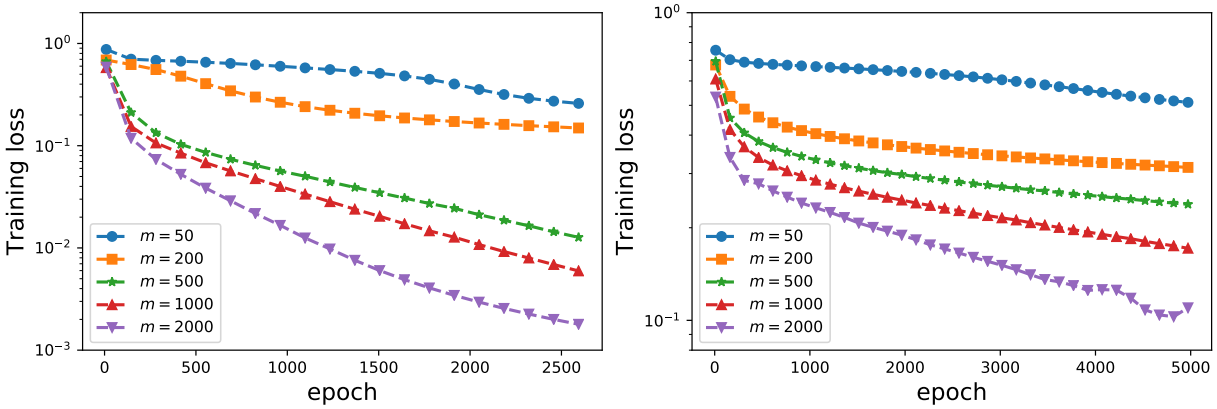
which is exactly in the same order of τ, where the last equality follows from the over-parameterization assumption $m = \widetilde{\Omega}\big(n^{14}L^{16}\phi^{-4} + n^{12}L^{16}\phi^{-4}\epsilon^{-1}\big)$. This verifies that $\mathbf{W}^{(k+1)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ and completes the induction for k. Thus we can complete the proof.

□

4.6 Experiments

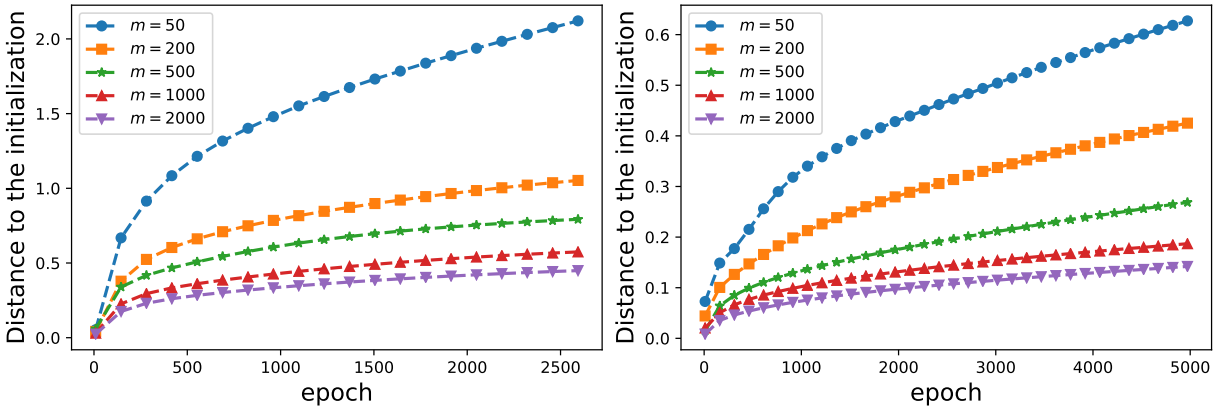
In this section we carry out experiments on two real datasets (MNIST [LBB98] and CIFAR10 [Kri09]) to support our theory. Since we mainly focus on binary classification, we extract a subset with digits 3 and 8 from the original MNIST dataset, which consists of $9,943$ training examples. In addition, we also extract two classes of images ("cat" and "ship") from the original CIFAR10 dataset, which consists of $7,931$ training examples. Regarding the neural network architecture, we use a fully-connected deep ReLU network with $L = 15$ hidden layers, each layer has width m. The network architecture is consistent with the setting of our theory.

We first demonstrate that over-parameterization indeed helps optimization. We run GD for training deep ReLU networks with different network widths and plot the training loss in Figure 4.1, where we apply cross-entropy loss on both MNIST and CIFAR10 datasets. In addition, the step sizes are set to be small enough and fixed for ReLU networks with different width. It can be observed that over-parameterization indeed speeds up the convergence of gradient descent, which is consistent with Lemmas 4.5.2 and 4.5.3, since the square of gradient norm scales with m, which further implies that wider network leads to larger



(a) *MNIST* (b) *CIFAR10*

Figure 4.1: The convergence of GD for training deep ReLU network with different network widths. (a) MNIST dataset. (b) CIFAR10 dataset.



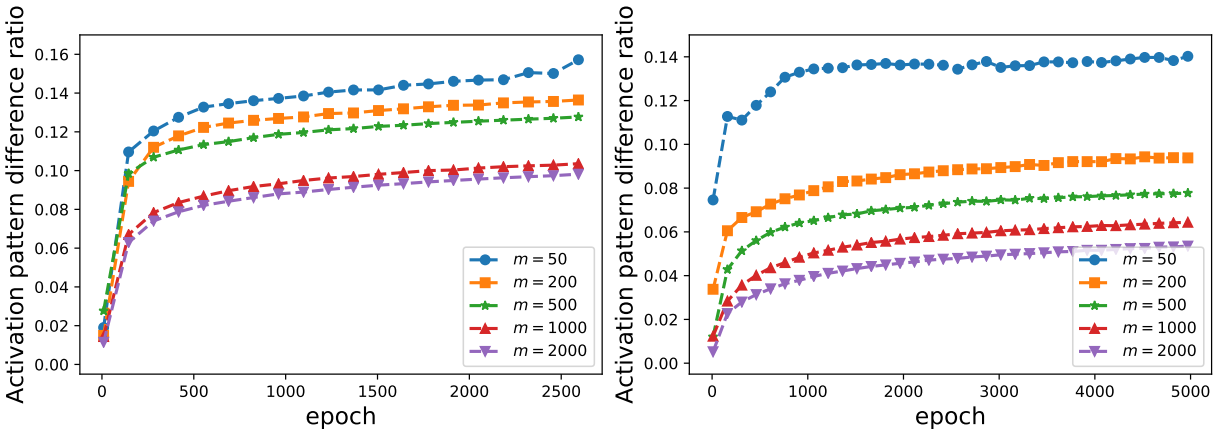
(a) *MNIST* (b) *CIFAR10*

Figure 4.2: Distance between the iterates of GD and the initialization. (a) MNIST dataset. (b) CIFAR10 dataset.

function decrease if the step size is fixed. We also display the distance between the iterates of GD and the initialization in Figure 4.2. It shows that when the network becomes wider, GD is more likely to converge to a point closer to the initialization. This suggests that the iterates of GD for training an over-parameterized deep ReLU network are harder to exceed the required perturbation region, thus can be guaranteed to converge to a global minimum.

This corroborates our theory.

Finally, we monitor the activation pattern changes of all hidden neurons during the training process, and show the results in Figure 4.3, where we use cross-entropy loss on both MNIST and CIFAR10 datasets. Specifically, in each iteration, we compare the activation status of all hidden nodes regarding all inputs with that at the initialization, and compute the number of nodes whose activation status differs from that at the initialization. From Figure 4.3 it is clear that the activation pattern difference ratio dramatically decreases as the neural network becomes wider, which brings less non-smoothness during the training process. This implies that wider ReLU network can better guarantee sufficient function decrease after one-step gradient descent, which is consistent with our theory.



(a) *MNIST* (b) *CIFAR10*

Figure 4.3: Activation pattern difference ratio between iterates of GD and the initialization. (a) MNIST dataset. (b) CIFAR10 dataset.

4.7 Proof of Lemmas in Section 4.5

In this section we provide the proof of all lemmas in Section 4.5.

4.7.1 Proof of Lemma 4.5.1

We first provide the following lemma that bounds the output of all hidden layer.

Lemma 4.7.1. With Gaussian random initialization, for any $\delta \in (0,1)$, if $m \geq \overline{C}L^2 \log(nL/\delta)$ for some large enough constant $\overline{C}$, then with probability at least $1-\delta$, the following holds for all $l \in [L]$,

$$\big|\|\mathbf{x}_{l,i}\|_2 - 1\big| \leq Cl\sqrt{\frac{\log(nL/\delta)}{m}},$$

where $m = \min\{m_1, \ldots, m_L\}$, and C is an absolute constant.

Proof of Lemma 4.5.1. Note that half of the entries of $\mathbf{v}$ are 1's and the other half of the entries are -1's. Therefore, without loss of generality, here we assume that $v_1 = \cdots = v_{m_L/2} = 1$ and $v_{m_L/2+1} = \cdots = v_{m_L} = -1$. Clearly, we have $\mathbb{E}(\widehat{y}_i) = 0$. Moreover, plugging in the value of $\mathbf{v}$ gives

$$\widehat{y}_i = \sum_{j=1}^{m_L/2} [\sigma(\mathbf{w}_{L,j}^\top \mathbf{x}_{L-1,i}) - \sigma(\mathbf{w}_{L,j+m_L/2}^\top \mathbf{x}_{L-1,i})].$$

Apparently, we have $\|\sigma(\mathbf{w}_{L,j}^\top \mathbf{x}_{L-1,i}) - \sigma(\mathbf{w}_{L,j+m_L/2}^\top \mathbf{x}_{L-1,i})\|_{\psi_2} \leq C_1 m_L^{-1/2}$ for some absolute constant C_1. Therefore by Hoeffding's inequality and Lemma 4.7.1, with probability at least $1-\delta$, it holds that

$$|\widehat{y}_i| \leq C_2\sqrt{\log(n/\delta)}$$

for all $i = 1, \ldots, n$. Then substituting the above bound into the formula of loss function $\ell(y_i\widehat{y}_i)$, we are able to complete the proof. □

4.7.2 Proof of Lemma 4.5.2

In order to prove Lemma 4.5.2, we require the following lemmas. We first establish the gradient lower bound at the initialization. Specifically, the following lemma gives a lower bound of gradient norm with respect to the weight matrix in the last hidden layer.

Lemma 4.7.2. There exist absolute constants $C, C', C'', C''' > 0$ such that, if $m \geq Cn^2\phi^{-1}\log(n)$, then with probability at least $1 - \exp(-C'm_L\phi/n)$, for any $\mathbf{a} = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n_+$, there exist at least $C''m_L\phi/n$ nodes in $\{1, \ldots, j, \ldots, m_L\}$ that satisfy

$$\bigg\| \frac{1}{n} \sum_{i=1}^{n} a_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i} \rangle) \mathbf{x}_{L-1,i} \bigg\|_2 \geq C''' \|\mathbf{a}\|_\infty / n$$

The following lemma characterizes the Lipschitz continuity of the gradients when the neural network parameters are staying inside the required perturbation region, which is essential to bound the norms of gradients.

Lemma 4.7.3 (Lemmas B.1 and B.2 in [ZCZ18])**.** Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization. For $\tau > 0$, let $\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_L$ with $\|\widetilde{\mathbf{W}}_l - \mathbf{W}_l\|_2 \leq \tau$, $l = 1, \ldots, L$ be the perturbed matrices. Let $\widetilde{\mathbf{\Sigma}}_{l,i}$, $l = 1, \ldots, L$, $i = 1, \ldots, n$ be diagonal matrices satisfying $\|\widetilde{\mathbf{\Sigma}}_{l,i} - \mathbf{\Sigma}_{l,i}\|_0 \leq s$ and $|(\widetilde{\mathbf{\Sigma}}_{l,i} - \mathbf{\Sigma}_{l,i})_{jj}|, |(\widetilde{\mathbf{\Sigma}}_{l,i})_{jj}| \leq 1$ for all $l = 1, \ldots, L$, $i = 1, \ldots, n$ and $j = 1, \ldots, m_l$. If $\tau, \sqrt{s\log(M)/m} \leq \kappa L^{-3/2}$ for some small enough absolute constant κ, then

$$\bigg\| \prod_{r=l_1}^{l_2} \widetilde{\mathbf{\Sigma}}_{r,i} \widetilde{\mathbf{W}}_r^\top \bigg\|_2 \leq C\sqrt{L}, \ \bigg\| \mathbf{v}^\top \prod_{r=l_1}^{L} \widetilde{\mathbf{\Sigma}}_{r,i} \widetilde{\mathbf{W}}_r^\top \bigg\|_2 \leq C'\sqrt{M}, \ \bigg\| \mathbf{v}^\top \prod_{r=l_1}^{L} \widetilde{\mathbf{\Sigma}}_{r,i} \widetilde{\mathbf{W}}_r^\top \mathbf{u} \bigg\|_2 \leq C''\sqrt{s\log(M)}$$

for any $1 \leq l_1 < l_2 \leq L$ and vector $\mathbf{u}$ with $\|\mathbf{u}\|_2 = 1$ and $\|\mathbf{u}\|_0 \leq s$, where C, C' and C'' are absolute constants.

We then provide the following lemma which characterizes the difference between activation patterns and outputs of all hidden layers generated by any two different neural networks.

Lemma 4.7.4 (Lemma B.3 in [ZCZ18])**.** Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization. Let $\widetilde{\mathbf{W}} = \{\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_L\}$, $\widehat{\mathbf{W}} = \{\widehat{\mathbf{W}}_1, \ldots, \widehat{\mathbf{W}}_L\}$ be two collections of weight matrices satisfying $\|\widetilde{\mathbf{W}}_l - \mathbf{W}_l\|_2, \|\widehat{\mathbf{W}}_l - \mathbf{W}_l\|_2 \leq \tau$, $l = 1, \ldots, L$. Let $\mathbf{\Sigma}_{l,i}, \widetilde{\mathbf{\Sigma}}_{l,i}, \widehat{\mathbf{\Sigma}}_{l,i}$ and $\mathbf{x}_{l,i}, \widetilde{\mathbf{x}}_{l,i}, \widehat{\mathbf{x}}_{l,i}$ be the binary matrices and hidden layer outputs at the l-th layer with parameter matrices $\mathbf{W}, \widetilde{\mathbf{W}}, \widehat{\mathbf{W}}$ respectively. If $\tau \leq \kappa' L^{-11}(\log(M))^{-3/2}$ for some small enough absolute constant $\kappa' > 0$, then there exits constants C and C' such that

$$\|\widehat{\mathbf{x}}_{l,i} - \widetilde{\mathbf{x}}_{l,i}\|_2 \leq CL \cdot \sum_{r=1}^{l} \|\widehat{\mathbf{W}}_r - \widetilde{\mathbf{W}}_r\|_2, \qquad \|\widehat{\mathbf{\Sigma}}_{l,i} - \widetilde{\mathbf{\Sigma}}_{l,i}\|_0 \leq C' L^{4/3} \tau^{2/3} m_l,$$

for all $l = 1, \ldots, L$ and $i = 1, \ldots, n$.

Now we ready to prove Lemma 4.5.2.

Proof of Lemma 4.5.2. We first prove the gradient upper bound. For the training example $(\mathbf{x}_i, y_i)$, let $\widetilde{y}_i = f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)$, the gradient $\nabla_{\mathbf{W}_l} \ell(y_i \widetilde{y}_i)$ can be written as follows,

$$\begin{aligned}
\nabla_{\mathbf{W}_l} \ell(y_i \widetilde{y}_i) &= \ell'(y_i \widetilde{y}_i) y_i \nabla_{\mathbf{W}_l} [f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)] \\
&= \ell'(y_i \widetilde{y}_i) y_i \widetilde{\mathbf{x}}_{l-1,i} \mathbf{v}^\top \bigg(\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_{r,i} \widetilde{\mathbf{W}}_l^\top \bigg) \widetilde{\boldsymbol{\Sigma}}_{l,i}.
\end{aligned}$$

Note that by Lemma 4.7.3, there exists an absolute constant C_0 such that $\| \prod_{r=l_1}^{l_2} \widetilde{\boldsymbol{\Sigma}}_{r,i} \widetilde{\mathbf{W}}_r \|_2 \leq C_0 \sqrt{L}$. Hence, we have the following upper bound on $\big\| \nabla_{\mathbf{W}_l} \ell(y_i \widetilde{y}_i) \big\|_F$,

$$\begin{aligned}
\big\| \nabla_{\mathbf{W}_l} \ell(y_i \widetilde{y}_i) \big\|_F &= \big\| \nabla_{\mathbf{W}_l} \ell(y_i \widetilde{y}_i) \big\|_2 \\
&\leq -\ell'(y_i \widetilde{y}_i) \bigg\| \prod_{r=1}^{l-1} \widetilde{\boldsymbol{\Sigma}}_{r,i} \widetilde{\mathbf{W}}_r^\top \mathbf{x}_i \bigg\|_2 \bigg\| \prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_{r,i} \widetilde{\mathbf{W}}_l^\top \bigg\|_2 \|\mathbf{v}\|_2 \\
&\leq -\ell'(y_i \widetilde{y}_i) C_0^2 L M^{1/2},
\end{aligned}$$

where the first equality holds due to the fact that the gradient of $\mathbf{W}_l$: $\nabla_{\mathbf{W}_l} \ell(y_i \widetilde{y}_i) = \ell'(y_i \widetilde{y}_i) y_i \widetilde{\mathbf{x}}_{l-1,i} \mathbf{v}^\top \big(\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_{r,i} \widetilde{\mathbf{W}}_l^\top \big) \widetilde{\boldsymbol{\Sigma}}_{l,i}$ is a rank-one matrix, and the last inequality follows from the fact that $\|\mathbf{v}\|_2 = m_L^{1/2} \leq M^{1/2}$. Moreover, we have the following for $\nabla_{\mathbf{W}_l} [L_S(\widetilde{\mathbf{W}})]$:

$$\big\| \nabla_{\mathbf{W}_l} [L_S(\widetilde{\mathbf{W}})] \big\|_F = \bigg\| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\mathbf{W}_l} \ell(y_i \widetilde{y}_i) \bigg\|_F \leq \frac{1}{n} \sum_{i=1}^{n} \big\| \nabla_{\mathbf{W}_l} \ell(y_i \widetilde{y}_i) \big\|_F \leq -\frac{C_0^2 L M^{1/2}}{n} \sum_{i=1}^{n} \ell'(y_i \widetilde{y}_i),$$

which completes the proof of gradient upper bound.

Now we are going to prove the gradient lower bound. Given initialization $\mathbf{W}^{(0)}$ and any $\widetilde{\mathbf{W}} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, let $\widetilde{y}_i = f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)$, we define

$$\mathbf{g}_j = \frac{1}{n} \sum_{i=1}^{n} \ell'(y_i \widetilde{y}_i) y_i \mathbf{v}_j \sigma'(\langle \mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i} \rangle) \mathbf{x}_{L-1,i},$$

where $\mathbf{x}_{L,i}$ denotes the output of the last hidden layer with input $\mathbf{x}_i$ at the initialization. Then since $\mathbf{W}^{(0)}$ is generated via Gaussian random initialization, by Lemma 4.7.2, we have the following holds for at least $C_2 m_L \phi/n$ nodes,

$$\|\mathbf{g}_j\|_2 \geq C_1 \max_i |\ell'(y_i \widetilde{y}_i)|/n$$

where $C_1, C_2 > 0$ are positive absolute constants. Moreover, we rewrite the gradient $\nabla_{\mathbf{W}_{L,j}} L_S(\widetilde{\mathbf{W}})$ as follows:

$$\nabla_{\mathbf{W}_{L,j}} L_S(\widetilde{\mathbf{W}}) = \frac{1}{n}\sum_{i=1}^n \ell'(y_i \widetilde{y}_i) y_i \mathbf{v}_j \sigma'(\langle \widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle)\widetilde{\mathbf{x}}_{L-1,i},$$

where $\widetilde{\mathbf{x}}_{l,i}$ denotes the output of the l-th hidden layer with input $\mathbf{x}_i$ and weight matrices $\widetilde{\mathbf{W}}$. Let $b_{i,j} = \ell'(y_i \widetilde{y}_i) y_i \mathbf{v}_j$, we have

$$\begin{aligned}
&\|\mathbf{g}_j\|_2 - \|\nabla_{\mathbf{W}_{L,j}} L_S(\widetilde{\mathbf{W}})\|_2 \\
&\leq \bigg\| \frac{1}{n}\sum_{i=1}^n b_{i,j}\big(\sigma'(\langle \widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle)\widetilde{\mathbf{x}}_{L-1,i} - \sigma'(\langle \mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\big)\bigg\|_2 \\
&\leq \bigg\| \frac{1}{n}\sum_{i=1}^n b_{i,j}\Big[\big(\sigma'(\langle \widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle) - \sigma'(\langle \mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle\big)\mathbf{x}_{L-1,i} + \sigma'(\langle \widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle)(\widetilde{\mathbf{x}}_{L-1,i} - \mathbf{x}_{L-1,i})\Big]\bigg\|_2.
\end{aligned}$$

According to Lemma 4.7.4, the number of nodes satisfying $\sigma'(\langle \widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle) - \sigma'(\langle \mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle \neq 0$ for at least one i is at most $C_3 n L^{4/3}\tau^{2/3} m_L$, where C_3 is an absolute constant. For the rest of the nodes in this layer, we have

$$\begin{aligned}
\|\mathbf{g}_j\|_2 - \|\nabla_{\mathbf{W}_{L,j}} L_S(\widetilde{\mathbf{W}})\|_2 &\leq \bigg\| \frac{1}{n}\sum_{i=1}^n b_{i,j}\sigma'(\langle \widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle)(\widetilde{\mathbf{x}}_{L-1,i} - \mathbf{x}_{L-1,i})\bigg\|_2 \\
&\leq \frac{1}{n}\sum_{i=1}^n C_4 L^2 \tau |b_{i,j}| \\
&\leq C_4 L^2 \tau \max_i |\ell'(y_i \widetilde{y}_i)|,
\end{aligned}$$

where C_4 is an absolute constant, the first inequality holds since these nodes satisfy $\sigma'(\langle \widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle) - \sigma'(\langle \mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle = 0$ for all i, the second inequality follows from Lemma 4.7.4 and triangle

inequality. Let

$$\tau \leq \left(\frac{C_2\phi}{2C_3n^2L^{4/3}}\right)^{3/2} \wedge \frac{C_1}{2nL^2C_4} = O\big(\phi^{3/2}n^{-3}L^{-2}\big).$$

Note that we have at least $C_2m_L\phi/n$ nodes satisfying $\|\mathbf{g}_j\|_2 \geq C_1 \max_i |\ell'(y_i\widetilde{y}_i)|/n$, thus there are at least $C_2m_L\phi/n - C_3nL^{4/3}\tau^{2/3}m_L = C_2m_L\phi/(2n)$ nodes satisfying

$$\|\nabla_{\mathbf{W}_{L,j}} L_S(\widetilde{\mathbf{W}})\|_2 \geq C_1 \max_i |\ell'(y_i\widetilde{y}_i)|/n - C_4L^2\tau \max_i |\ell'(y_i\widetilde{y}_i)|/n \geq \frac{C_1 \max_i |\ell'(y_i\widetilde{y}_i)|}{2n}.$$

Therefore,

$$\begin{aligned}
\|\nabla_{\mathbf{W}_L} L_S(\widetilde{\mathbf{W}})\|_F^2 &= \sum_{j=1}^{m_L} \|\nabla_{\mathbf{W}_{L,j}} L_S(\widetilde{\mathbf{W}})\|_2^2 \\
&\geq \frac{C_2\phi m_L}{2n}\left(\frac{C_1 \max_i |\ell'(y_i\widehat{y}_i^{(k)})y_i\mathbf{v}_j|}{2n}\right)^2 \\
&\geq \frac{C_2C_1^2\phi m_L}{8n^5}\left(\sum_{i=1}^n \ell'(y_i\widehat{y}_i^{(k)})\right)^2,
\end{aligned}$$

where the last inequality follows from the fact that $\ell'(\cdot) < 0$ and $|y_i\mathbf{v}_j| = 1$. Let $C = C_2C_1^2/8$, we complete the proof. □

4.7.3 Proof of Lemma 4.5.3

Proof of Lemma 4.5.3. Note that $\ell(x)$ is 1/4-smooth, thus the following holds for any Δ and x,

$$\ell(x+\Delta) \leq \ell(x) + \ell'(x)\Delta + \frac{1}{8}\Delta^2.$$

Then we have the following upper bound on $L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)})$,

$$\begin{aligned}
L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) &= \frac{1}{n}\sum_{i=1}^n \Big[\ell\big(y_i\widehat{y}_i^{(k+1)}\big) - \ell\big(y_i\widehat{y}_i^{(k)}\big)\Big] \\
&\leq \frac{1}{n}\sum_{i=1}^n \Big[\ell'(y_i\widehat{y}_i^{(k)})\Delta_i^{(k)} + \frac{1}{8}(\Delta_i^{(k)})^2\Big], \qquad (4.7.1)
\end{aligned}$$

where $\Delta_i^{(k)} = y_i\big(\widehat{y}_i^{(k+1)} - \widehat{y}_i^{(k)}\big)$. Therefore, our next goal is to bound the quantity $\Delta_i^{(k)}$.

The upper bound of $|\Delta_i^{(k)}|$ can be derived straightforwardly. By Lemma 4.7.4, we know that there exists a constant C_1 such that

$$\begin{aligned}\|\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\|_2 &\leq C_1 L \cdot \sum_{l=1}^{L} \|\mathbf{W}_l^{(k+1)} - \mathbf{W}_l^{(k)}\|_2 \\ &= C_1 L\eta \sum_{l=1}^{L} \big\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big\|_2 \\ &\leq C_1 L^{1.5}\eta \|\nabla L_S(\mathbf{W}^{(k)})\|_F. \end{aligned} \tag{4.7.2}$$

Therefore, it follows that

$$|\Delta_i^{(k)}| = |y_i \mathbf{v}^\top (\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)})| \leq \|\mathbf{v}\|_2 \|\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\|_2 \leq C_1 L^{1.5} M^{1/2} \|\nabla L_S(\mathbf{W}^{(k)})\|_F,$$

where we use the fact that $\|\mathbf{v}\|_2 \leq M^{1/2}$. In what follows we are going to prove the lower bound of $\Delta_i^{(k)}$. Note that $\Delta_i^{(k)} = y_i \mathbf{v}^\top \big(\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\big)$, thus we mainly focus on bounding the term $\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}$. For $l = 1, \ldots, L$, we define the diagonal matrix $\widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}$ as

$$\big(\widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}\big)_{jj} = \big(\boldsymbol{\Sigma}_{l,i}^{(k+1)} - \boldsymbol{\Sigma}_{l,i}^{(k)}\big)_{jj} \cdot \frac{\mathbf{w}_{l,j}^{(k)\top} \mathbf{x}_{l-1,i}^{(k)}}{\mathbf{w}_{l,j}^{(k+1)\top} \mathbf{x}_{l-1,i}^{(k+1)} - \mathbf{w}_{l,j}^{(k)\top} \mathbf{x}_{l-1,i}^{(k)}}.$$

Given the above definition of $\widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}$, we have

$$\begin{aligned}&\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)} \\ &= \big(\boldsymbol{\Sigma}_{L,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{L,i}^{(k)}\big)\big(\mathbf{W}_L^{(k+1)\top} \mathbf{x}_{L-1,i}^{(k+1)} - \mathbf{W}_L^{(k)\top} \mathbf{x}_{L-1,i}^{(k)}\big) \\ &= \big(\boldsymbol{\Sigma}_{L,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{L,i}^{(k)}\big)\mathbf{W}_L^{(k+1)\top}\big(\mathbf{x}_{L-1,i}^{(k+1)} - \mathbf{x}_{L-1,i}^{(k)}\big) + \big(\boldsymbol{\Sigma}_{L,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{L,i}^{(k)}\big)\big(\mathbf{W}_L^{(k+1)\top} - \mathbf{W}_L^{(k)\top}\big)\mathbf{x}_{L-1,i}^{(k)} \\ &= \sum_{l=1}^{L} \bigg(\prod_{r=l+1}^{L} \big(\boldsymbol{\Sigma}_{r,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{r,i}^{(k)}\big)\mathbf{W}_r^{(k+1)\top} \bigg) \big(\boldsymbol{\Sigma}_{l,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}\big)\big(\mathbf{W}_l^{(k+1)\top} - \mathbf{W}_l^{(k)\top}\big)\mathbf{x}_{l-1,i}^{(k)}. \end{aligned}$$

Then we define

$$\boldsymbol{D}_{l,i}^{(k)} = \bigg(\prod_{r=l+1}^{L} \boldsymbol{\Sigma}_{r,i}^{(k)} \mathbf{W}_r^{(k)\top} \bigg) \boldsymbol{\Sigma}_{l,i}^{(k)}, \quad \widetilde{\boldsymbol{D}}_{l,i}^{(k)} = \bigg(\prod_{r=l+1}^{L} \big(\boldsymbol{\Sigma}_{r,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{r,i}^{(k)}\big)\mathbf{W}_r^{(k+1)\top} \bigg) \big(\boldsymbol{\Sigma}_{l,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}\big).$$

Then by triangle inequality, we have

$$\big\|\mathbf{v}^\top \big(\boldsymbol{D}_{l,i}^{(k)} - \widetilde{\boldsymbol{D}}_{l,i}^{(k)}\big)\big\|_2 \leq \big\|\mathbf{v}^\top \big(\boldsymbol{D}_{l,i}^{(k)} - \boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2 + \big\|\mathbf{v}^\top \big(\boldsymbol{D}_{l,i}^{(0)} - \widetilde{\boldsymbol{D}}_{l,i}^{(k)}\big)\big\|_2.$$

Note that, it holds that

$$
\begin{aligned}
&\big\|\mathbf{v}^\top\big(\boldsymbol{D}_{l,i}^{(k)}-\boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2\\
&\le\sum_{r=l}^{L}\bigg\|\mathbf{v}^\top\Big(\prod_{t=r+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(k)}\mathbf{W}_t^{(k)\top}\Big)\Big(\boldsymbol{\Sigma}_{t,i}^{(k)}\mathbf{W}_t^{(k)\top}-\boldsymbol{\Sigma}_{t,i}^{(0)}\mathbf{W}_t^{(0)\top}\Big)\Big(\prod_{t=l+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(0)}\mathbf{W}_t^{(0)\top}\Big)\bigg\|_2\\
&\le\sum_{r=l}^{L}\bigg\|\mathbf{v}^\top\Big(\prod_{t=r+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(k)}\mathbf{W}_t^{(k)\top}\Big)\Big(\boldsymbol{\Sigma}_{t,i}^{(k)}-\boldsymbol{\Sigma}_{t,i}^{(0)}\Big)\bigg\|_2\bigg\|\mathbf{W}_t^{(0)\top}\Big(\prod_{t=l+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(0)}\mathbf{W}_t^{(0)\top}\Big)\bigg\|_2\\
&+\sum_{r=l}^{L}\bigg\|\mathbf{v}^\top\Big(\prod_{t=r+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(k)}\mathbf{W}_t^{(k)\top}\Big)\boldsymbol{\Sigma}_{t,i}^{(k)}\bigg\|_2\bigg\|\mathbf{W}_t^{(k)}-\mathbf{W}_t^{(0)\top}\bigg\|_2\bigg\|\prod_{t=l+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(0)}\mathbf{W}_t^{(0)\top}\bigg\|_2.
\end{aligned}
$$

Then by Lemma 4.7.3, and use the fact that $\|\boldsymbol{\Sigma}_{t,i}^{(k)}-\boldsymbol{\Sigma}_{t,i}^{(0)}\|_0\le O\big(L^{4/3}\tau^{2/3}M\big)$, we have

$$
\big\|\mathbf{v}^\top\big(\boldsymbol{D}_{l,i}^{(k)}-\boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2\le C_2L^{13/6}\tau^{1/3}\sqrt{M\log(M)}+C_3L^{3/2}\sqrt{M}\tau,
$$

where C_2 and C_3 are absolute constants and we use the fact that $\|\mathbf{v}\|_2\le\sqrt{M}$. Then note that $\tau\le 1$, the second term on the R.H.S. of the above inequality is dominated by the first one. Then we have

$$
\big\|\mathbf{v}^\top\big(\boldsymbol{D}_{l,i}^{(k)}-\boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2\le C_5L^{13/6}\tau^{1/3}\sqrt{M\log(M)},\tag{4.7.3}
$$

where C_5 is an absolute constant. This inequality also holds for $\big\|\mathbf{v}^\top\big(\widetilde{\boldsymbol{D}}_{l,i}^{(k)}-\boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2$. Therefore, we have

$$
\begin{aligned}
\Delta_i^{(k)}&=y_i\mathbf{v}^\top\big(\mathbf{x}_{L,i}^{(k+1)}-\mathbf{x}_{L,i}^{(k)}\big)\\
&=y_i\mathbf{v}^\top\sum_{l=1}^{L}\widetilde{\boldsymbol{D}}_{l,i}^{(k)}\big(\mathbf{W}_l^{(k+1)}-\mathbf{W}_l^{(k)}\big)\mathbf{x}_{l-1,i}^{(k)}\\
&=\underbrace{-y_i\mathbf{v}^\top\sum_{l=1}^{L}\big(\widetilde{\boldsymbol{D}}_{l,i}^{(k)}-\boldsymbol{D}_{l,i}^{(k)}\big)\big(\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big)^\top\mathbf{x}_{l-1,i}^{(k)}}_{I_{1,i}^{(k)}}\\
&\quad\underbrace{-y_i\mathbf{v}^\top\sum_{l=1}^{L}\boldsymbol{D}_{l,i}^{(k)}\big(\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big)^\top\mathbf{x}_{l-1,i}^{(k)}}_{I_{2,i}^{(k)}}.
\end{aligned}
$$

By (4.7.3), we know that

$$|I_{1,i}^{(k)}| \leq 2C_5 L^{13/6}\tau^{1/3}\sqrt{M\log(M)}\eta \cdot \sum_{l=1}^{L} \big\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big\|_2$$
$$\leq 2C_5 L^{8/3}\tau^{1/3}\sqrt{M\log(M)}\eta \cdot \big\|\nabla L_S(\mathbf{W}^{(k)})\big\|_F.$$

Moreover, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})I_{2,i}^{(k)} &= -\frac{\eta}{n}\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})y_i\mathbf{v}^\top\bigg(\prod_{r=l+1}^{L}\mathbf{\Sigma}_{r,i}^{(k)}\mathbf{W}_r^{(k)\top}\bigg)\mathbf{\Sigma}_{l,i}^{(k)}\big(\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big)^\top\mathbf{x}_{l-1,i}^{(k)} \\
&= -\frac{\eta}{n^2}\bigg\|\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})y_i\mathbf{x}_{l-1,i}^{(k)}\mathbf{v}^\top\bigg(\prod_{r=l+1}^{L}\mathbf{\Sigma}_{r,i}^{(k)}\mathbf{W}_r^{(k)\top}\bigg)\mathbf{\Sigma}_{l,i}^{(k)}\bigg\|_F^2 \\
&= -\eta\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\|_F^2.
\end{aligned}$$

Therefore, putting everything together, we have

$$\begin{aligned}
&L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\Big[\ell'(y_i\widehat{y}_i^{(k)})\Delta_i^{(k)} + \frac{1}{8}(\Delta_i^{(k)})^2\Big] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})\big(I_{1,i}^{(k)} + I_{2,i}^{(k)}\big) + C_3 ML^3\eta^2\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 \\
&\leq -\big(\eta - C_6 ML^3\eta^2\big)\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 - \frac{C_7 L^{8/3}\tau^{1/3}\sqrt{M\log(M)}\cdot\|\nabla L_S(\mathbf{W}^{(k)})\|_F}{n}\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)}),
\end{aligned}$$

where C_6 and C_7 are absolute constants. Thus we complete the proof.

□

4.8 Proof of Lemmas in Section 4.7

4.8.1 Proof of Lemma 4.7.1

Proof of Lemma 4.7.1. In order to prove the desired results, it suffices to prove the inequality

$$\big|\|\mathbf{x}_{l,i}\|_2^2 - \|\mathbf{x}_{l-1,i}\|_2^2\big| \leq C\|\mathbf{x}_{l-1,i}\|_2^2 \cdot \sqrt{\frac{\log(nL/\delta)}{m_l}}$$

for all $i = 1, \ldots, n$ and $l = 1, \ldots, L$, since this inequality implies that

$$\|\mathbf{x}_{l,i}\|_2 \leq \left[1 + C'\sqrt{\frac{\log(nL/\delta)}{m}}\right]^{1/2} \|\mathbf{x}_{l-1,i}\|_2 \leq \cdots \leq \left[1 + C'\sqrt{\frac{\log(nL/\delta)}{m}}\right]^{l/2} \|\mathbf{x}_i\|_2$$
$$\leq 1 + C'l\sqrt{\frac{\log(nL/\delta)}{m}},$$

where C' is an absolute constant, and the last inequality follows by the fact that $(1 + x)^{l/2} \leq 1 + lx$ for $x \in (0, 1/(2L))$, which is applicable here because of the assumption $m \geq \overline{C}L^2 \log(nL/\delta)$ for some large enough constant $\overline{C}$. Similarly, we can also proved that

$$\|\mathbf{x}_{l,i}\|_2 \geq 1 - C''l\sqrt{\frac{\log(nL/\delta)}{m}}$$

for some absolute constant C''. Combining the upper and lower bounds of $\|\mathbf{x}_{l,i}\|_2$ derived above gives the result of Lemma 4.7.1.

For any fixed $i \in \{1, \ldots, n\}$, $l \in \{1, \ldots, L\}$ and $j \in \{1, \ldots, m_l\}$, condition on $\mathbf{x}_{l-1,i}$ we have $\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1,i} \sim N(\mathbf{0}, 2\|\mathbf{x}_{l-1,i}\|_2^2/m_l)$. Therefore,

$$\mathbb{E}[\sigma^2(\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1,i})|\mathbf{x}_{l-1,i}] = \frac{1}{2}\mathbb{E}[(\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1,i})^2|\mathbf{x}_{l-1,i}] = \frac{1}{m_l}\|\mathbf{x}_{l-1,i}\|_2^2.$$

Since $\|\mathbf{x}_{l,i}\|_2^2 = \sum_{j=1}^{m_l} \sigma^2(\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1,i})$ and condition on $\mathbf{x}_{l-1}$, $\|\sigma^2(\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1,i})\|_{\psi_1} \leq C_1\|\mathbf{x}_{l-1,i}\|_2^2/m_l$ for some absolute constant C_1, by Bernstein inequality (See Proposition 5.16 in [Ver10]), for any $\xi \geq 0$ we have

$$\mathbb{P}\Big(\big|\|\mathbf{\Sigma}_{l,i}\mathbf{W}_l^\top \mathbf{x}_{l-1,i}\|_2^2 - \|\mathbf{x}_{l-1,i}\|_2^2\big| \geq \|\mathbf{x}_{l-1,i}\|_2^2\xi \Big| \mathbf{x}_{l-1,i}\Big) \leq 2\exp(-C_2 m_l \min\{\xi^2, \xi\}).$$

Taking union bound over l and i gives

$$\mathbb{P}\Big(\big|\|\mathbf{x}_{l,i}\|_2^2 - \|\mathbf{x}_{l-1,i}\|_2^2\big| \leq \|\mathbf{x}_{l-1,i}\|_2^2\xi, i = 1, \ldots, n, l = 1, \ldots, L\Big) \geq 1 - 2nL\exp(-C_2 m_l \min\{\xi^2, \xi\}).$$

The inequality above further implies that if $m_l \geq C_3^2 \log(nL/\delta)$, then with probability at least $1 - \delta$, we have

$$\big|\|\mathbf{x}_{l,i}\|_2^2 - \|\mathbf{x}_{l-1,i}\|_2^2\big| \leq C_3\|\mathbf{x}_{l-1,i}\|_2^2 \cdot \sqrt{\frac{\log(nL/\delta)}{m_l}}$$

for any $i = 1, \ldots, n$ and $l = 1, \ldots, L$, where C_3 is an absolute constant. This completes the proof. □

4.8.2 Proof of Lemma 4.7.2

In order to prove the gradient bounds, one key aspect is that the separation property for training data can be well preserved after passing through layers. The following lemma shows that the separation distance can be well preserved for all intermediate layers.

Lemma 4.8.1. Under the same conditions in Lemma 4.7.2, with probability at least $1-\delta$,

$$\|\overline{\mathbf{x}}_{l,i} - \overline{\mathbf{x}}_{l,i'}\|_2 \geq \phi/2$$

for all $i, i' = 1, \ldots, n$ with $y_i \neq y_{i'}$, $l = 1, \ldots, L$.

Lemma 4.8.2. Let $\mathbf{z}_1, \ldots, \mathbf{z}_n \in S^{d-1}$ be n unit vectors and $y_1, \ldots, y_n \in \{-1, 1\}$ be the corresponding labels. Assume that for any $i \neq j$ such that $y_i \neq y_j$, $\|\mathbf{z}_i - \mathbf{z}_j\|_2 \geq \widetilde{\phi}$ and $\mathbf{z}_i^\top \mathbf{z}_j \geq \widetilde{\mu}^2$ for some $\widetilde{\phi}, \widetilde{\mu} > 0$. For any $\mathbf{a} = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n_+$, let $\mathbf{h}(\mathbf{w}) = \sum_{i=1}^n a_i y_i \sigma'(\langle \mathbf{w}, \mathbf{z}_i \rangle)\mathbf{z}_i$ where $\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$ is a Gaussian random vector. If $\widetilde{\phi} \leq \widetilde{\mu}/2$, then there exist absolute constants $C, C' > 0$ such that

$$\mathbb{P}\big[\|\mathbf{h}(\mathbf{w})\|_2 \geq C\|\mathbf{a}\|_\infty\big] \geq C'\widetilde{\phi}/n.$$

The following lemma is essential to show that deep ReLU network can provide significantly large gradient at the initialization.

Proof of Lemma 4.7.2. For any given $j \in \{1, \ldots, m_L\}$ and $\widehat{\mathbf{a}}$ with $\|\widehat{\mathbf{a}}\|_\infty = 1$. By Lemma 4.8.1, we know that for any $i \neq j$ and $y_i \neq y_j$, $\|\bar{\mathbf{x}}_{L-1,i} - \bar{\mathbf{x}}_{L-1,j}\|_2 \geq \widetilde{\phi}$, where $\bar{\mathbf{x}}_{L-1,i} = \mathbf{x}_{L-1,i}/\|\mathbf{x}_{L-1,i}\|_2$ and $\bar{\mathbf{x}}_{L-1,j} = \mathbf{x}_{L-1,j}/\|\mathbf{x}_{L-1,i}\|_2$. Then by Lemma 4.8.2, we have

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \widehat{a}_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2 \geq \frac{C_1}{n}\right] \geq \frac{C_2\phi}{n},$$

where $C_1, C_2 > 0$ are absolute constants. Let $S^{n-1}_{\infty,+} = \{\mathbf{a} \in \mathbb{R}^n_+ : \|\mathbf{a}\|_\infty = 1\}$, and $\mathcal{N} = \mathcal{N}[S^{n-1}_{\infty,+}, C_1/(4n)]$ be a $C_1/(4n)$-net covering $S^{n-1}_{\infty,+}$ in ℓ_∞ norm. Then we have

$$|\mathcal{N}| \leq (4n/C_1)^n.$$

For $j = 1, \ldots, m_L$, define

$$Z_j = \mathbb{1}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \widehat{a}_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2 \geq \frac{C_1}{n}\right].$$

Let $p_\phi = C_2\phi/n$. Then by Bernstein inequality and union bound, with probability at least $1 - \exp[-C_3 m_L p_\phi + n\log(4n/C_1)] \geq 1 - \exp(C_4 m_L \phi/n)$, we have

$$\frac{1}{m_L}\sum_{j=1}^{m_L} Z_j \geq p_\phi/2, \tag{4.8.1}$$

where C_3, C_4 are absolute constants. For any $\mathbf{a} \in S_{\infty,+}^{n-1}$, there exists $\widehat{\mathbf{a}} \in \mathcal{N}$ such that

$$\|\mathbf{a} - \widehat{\mathbf{a}}\|_\infty \leq C_1/(4n).$$

Therefore, we have

$$\begin{aligned}
&\left|\left\|\frac{1}{n}\sum_{i=1}^{n} a_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2 - \left\|\frac{1}{n}\sum_{i=1}^{n} \widehat{a}_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2\right| \\
&\leq \left\|\frac{1}{n}\sum_{i=1}^{n} a_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i} - \frac{1}{n}\sum_{i=1}^{n} \widehat{a}_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n} |a_i - \widehat{a}_i| \leq \frac{C_1}{2n}.
\end{aligned} \tag{4.8.2}$$

By (4.8.1) and (4.8.2), it is clear that with probability at least $1 - \exp(C_4 m_L \phi/n)$, for any $\mathbf{a} \in S_{\infty,+}^{n-1}$, there exist at least $m_L p_\phi/2$ nodes on layer L that satisfy

$$\left\|\frac{1}{n}\sum_{i=1}^{n} a_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2 \geq \frac{C_1}{2n}.$$

This completes the proof. □

4.9 Proof of Lemmas in Section

4.9.1 Proof of Lemma 4.8.1

The following lemma is necessary for proving Lemma 4.8.1.

Lemma 4.9.1 (Lemma A.3 in [ZCZ18])**.** For $\theta > 0$, let Z_1, Z_2 be two jointly Gaussian random variables with $\mathbb{E}(Z_1) = \mathbb{E}(Z_2) = 0$, $\mathbb{E}(Z_1^2) = \mathbb{E}(Z_2^2) = 1$ and $\mathbb{E}(Z_1 Z_2) \leq 1 - \theta^2/2$. If $\theta \leq \kappa$ for some small enough absolute constant κ, then

$$\mathbb{E}[\sigma(Z_1)\sigma(Z_2)] \leq \frac{1}{2} - \frac{1}{4}\theta^2 + C\theta^3,$$

where C is an absolute constant.

Proof of Lemma 4.8.1. We first consider any fixed $l \geq 1$. Suppose that $\|\overline{\mathbf{x}}_{l-1,i} - \overline{\mathbf{x}}_{l-1,i'}\|_2 \geq [1 - (2L)^{-1}\log(2)]^{l-1}\phi$. If we can show that under this condition, with high probability

$$\|\overline{\mathbf{x}}_{l,i} - \overline{\mathbf{x}}_{l,i'}\|_2 \geq [1 - (2L)^{-1}\log(2)]^{l}\phi,$$

then the result of the lemma follows by union bound and induction. Denote

$$\phi_{l-1} = [1 - (2L)^{-1}\log(2)]^{l-1}\phi.$$

Then by assumption we have $\|\overline{\mathbf{x}}_{l-1,i} - \overline{\mathbf{x}}_{l-1,i'}\|_2^2 \geq \phi_{l-1}^2$. Therefore $\overline{\mathbf{x}}_{l-1,i}^\top \overline{\mathbf{x}}_{l-1,i'} \leq 1 - \phi_{l-1}^2/2$. It follows by direct calculation that

$$\begin{aligned}
\mathbb{E}\big(\|\mathbf{x}_{l,i} - \mathbf{x}_{l,i'}\|_2^2 \big| \mathbf{x}_{l-1,i}, \mathbf{x}_{l-1,i'}\big) &= \mathbb{E}\big(\|\mathbf{x}_{l,i}\|_2^2 + \|\mathbf{x}_{l,i'}\|_2^2 \big| \mathbf{x}_{l-1,i}, \mathbf{x}_{l-1,i'}\big) - 2\mathbb{E}\big(\mathbf{x}_{l,i}^\top \mathbf{x}_{l,i'} \big| \mathbf{x}_{l-1,i}, \mathbf{x}_{l-1,i'}\big) \\
&= (\|\mathbf{x}_{l-1,i}\|_2^2 + \|\mathbf{x}_{l-1,i'}\|_2^2) - 2\mathbb{E}\big(\mathbf{x}_{l,i}^\top \mathbf{x}_{l,i'} \big| \mathbf{x}_{l-1,i}, \mathbf{x}_{l-1,i'}\big).
\end{aligned}$$

By Lemma 4.9.1 and the assumption that $\phi_{l-1} \leq \phi \leq \kappa$, we have

$$\begin{aligned}
&\mathbb{E}\big(\mathbf{x}_{l,i}^\top \mathbf{x}_{l,i'} \big| \mathbf{x}_{l-1,i}, \mathbf{x}_{l-1,i'}\big) \\
&= \mathbb{E}\bigg[\sum_{j=1}^{m_l} \sigma(\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1,i})\sigma(\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1,i'}) \bigg| \mathbf{x}_{l-1,i}, \mathbf{x}_{l-1,i'}\bigg] \\
&= \|\mathbf{x}_{l-1,i}\|_2 \|\mathbf{x}_{l-1,i'}\|_2 \cdot \mathbb{E}\bigg[\sum_{j=1}^{m_l} \sigma(\mathbf{w}_{l,j}^\top \overline{\mathbf{x}}_{l-1,i})\sigma(\mathbf{w}_{l,j}^\top \overline{\mathbf{x}}_{l-1,i'}) \bigg| \mathbf{x}_{l-1,i}, \mathbf{x}_{l-1,i'}\bigg] \\
&\leq \frac{2}{m}\|\mathbf{x}_{l-1,i}\|_2 \|\mathbf{x}_{l-1,i'}\|_2 \cdot m \cdot \bigg(\frac{1}{2} - \frac{1}{4}\phi_{l-1}^2 + C\phi_{l-1}^3\bigg) \\
&= \|\mathbf{x}_{l-1,i}\|_2 \|\mathbf{x}_{l-1,i'}\|_2 \cdot \bigg(1 - \frac{1}{2}\phi_{l-1}^2 + 2C\phi_{l-1}^3\bigg).
\end{aligned}$$

Therefore,

$$\mathbb{E}\big(\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2\big|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\big)\geq(\|\mathbf{x}_{l-1,i}\|_2-\|\mathbf{x}_{l-1,i'}\|_2)^2+\|\mathbf{x}_{l-1,i}\|_2\|\mathbf{x}_{l-1,i'}\|_2(\phi_{l-1}^2-4C\phi_{l-1}^3). \tag{4.9.1}$$

Condition on $\mathbf{x}_{l-1,i}$ and $\mathbf{x}_{l-1,i'}$, by Lemma 5.14 in [Ver10] we have

$$\begin{aligned}\big\|[\sigma(\mathbf{w}_{l,j}^\top\mathbf{x}_{l-1,i})-\sigma(\mathbf{w}_{l,j}^\top\mathbf{x}_{l-1,i'})]^2\big\|_{\psi_1}&\leq 2\big[\big\|[\sigma(\mathbf{w}_{l,j}^\top\mathbf{x}_{l-1,i})\big\|_{\psi_2}+\big\|\sigma(\mathbf{w}_{l,j}^\top\mathbf{x}_{l-1,i'})\big\|_{\psi_2}\big]^2\\&\leq C_1(\|\mathbf{x}_{l-1,i}\|_2+\|\mathbf{x}_{l-1,i'}\|_2)^2/m_l,\end{aligned}$$

where C_1 is an absolute constant. Therefore if $m_l\geq C_2^2\log(4n^2L/\delta)$, by Bernstein inequality and union bound, with probability at least $1-\delta/(4n^2L)$ we have

$$\big|\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2-\mathbb{E}\big(\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2\big|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\big)\big|\leq C_2(\|\mathbf{x}_{l-1,i}\|_2+\|\mathbf{x}_{l-1,i'}\|_2)^2\cdot\sqrt{\frac{\log(8n^2L/\delta)}{m_l}},$$

where C_2 is an absolute constant. Therefore with probability at least $1-\delta/(4n^2L)$ we have

$$\begin{aligned}\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2\geq{}&(\|\mathbf{x}_{l-1,i}\|_2-\|\mathbf{x}_{l-1,i'}\|_2)^2+\|\mathbf{x}_{l-1,i}\|_2\|\mathbf{x}_{l-1,i'}\|_2(\phi_{l-1}^2-4C\phi_{l-1}^3)\\&-C_2(\|\mathbf{x}_{l-1,i}\|_2+\|\mathbf{x}_{l-1,i'}\|_2)^2\cdot\sqrt{\frac{\log(8n^2L/\delta)}{m_l}}.\end{aligned}$$

By union bound and Lemma 4.7.1, if $m_r\geq C_3L^4\phi_l^{-4}\log(4n^2L/\delta)$, $r=1,\ldots,l$ for some large enough absolute constant C_3 and $\phi\leq\kappa L^{-1}$ for some small enough absolute constant κ, then with probability at least $1-\delta/(2n^2L)$ we have

$$\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2\geq[1-(4L)^{-1}\log(2)]\phi_{l-1}^2\geq[1-(4L)^{-1}\log(2)]^2\phi_{l-1}^2.$$

Moreover, by Lemma 4.7.1, with probability at least $1-\delta/(2n^2L)$ we have

$$\begin{aligned}\big|\|\overline{\mathbf{x}}_{l,i}-\overline{\mathbf{x}}_{l,i'}\|_2-\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2\big|&\leq\|\overline{\mathbf{x}}_{l,i}-\mathbf{x}_{l,i}\|_2+\|\overline{\mathbf{x}}_{l,i'}-\mathbf{x}_{l,i'}\|_2\\&=\big|1-\|\mathbf{x}_{l,i}\|_2\big|+\big|1-\|\mathbf{x}_{l,i'}\|_2\big|\\&\leq(4L)^{-1}\log(2)\cdot\phi_{l-1}^2,\end{aligned}$$

and therefore with probability at least $1-\delta/(n^2L)$, we have

$$\|\overline{\mathbf{x}}_{l,i}-\overline{\mathbf{x}}_{l,i'}\|_2\geq[1-(2L)^{-1}\log(2)]\phi_{l-1}=[1-(2L)^{-1}\log(2)]^l\phi.$$

Applying union bound and induction over $l=1,\ldots,L$ completes the proof. □

4.9.2 Proof of Lemma 4.8.2

Proof of Lemma 4.8.2. Without loss of generality, assume that $a_1 = \|\mathbf{a}\|_\infty$. Since $\|\mathbf{z}_1\|_2 = 1$, we can construct an orthonormal matrix $\mathbf{Q} = [\mathbf{z}_1, \mathbf{Q}'] \in \mathbb{R}^{d\times d}$. Let $\mathbf{u} = \mathbf{Q}^\top \mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$ be a standard Gaussian random vector. Then we have

$$\mathbf{w} = \mathbf{Q}\mathbf{u} = u_1\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}',$$

where $\mathbf{u}' := (u_2, \ldots, u_d)^\top$ is independent of u_1. We define the following two events based on a parameter $\gamma \in (0, 1]$:

$$\mathcal{E}_1(\gamma) = \big\{|u_1| \le \gamma\big\},\ \mathcal{E}_2(\gamma) = \big\{|\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle| \ge \gamma \text{ for all } \mathbf{z}_i \text{ such that } \|\mathbf{z}_i - \mathbf{z}_1\|_2 \ge \widetilde{\phi}\big\}.$$

Let $\mathcal{E}(\gamma) = \mathcal{E}_1(\gamma) \cap \mathcal{E}_2(\gamma)$. We first give lower bound for $\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathcal{E}_1)\mathbb{P}(\mathcal{E}_2)$. Since u_1 is a standard Gaussian random variable, we have

$$\mathbb{P}(\mathcal{E}_1) = \frac{1}{\sqrt{2\pi}} \int_{-\gamma}^{\gamma} \exp\Big(-\frac{1}{2}x^2\Big)\mathrm{d}x \ge \sqrt{\frac{2}{\pi e}}\gamma.$$

Moreover, by definition, for any $i = 1, \ldots, n$ we have

$$\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle \sim N\big[0, 1 - (\mathbf{z}_1^\top \mathbf{z}_i)^2\big].$$

Let $\mathcal{I} = \{i : \|\mathbf{z}_i - \mathbf{z}_1\|_2 \ge \widetilde{\phi}\}$. By the assumption that $\widetilde{\phi} \le \widetilde{\mu}/2$, for any $i \in \mathcal{I}$, we have

$$-1 + \widetilde{\phi}^2/2 \le -(1 - \widetilde{\mu}^2) + \widetilde{\mu}^2 \le \langle \mathbf{z}_i, \mathbf{z}_1\rangle \le 1 - \widetilde{\phi}^2/2,$$

and if $\widetilde{\phi}^2 \le 2$, then

$$1 - (\mathbf{z}_1^\top \mathbf{z}_i)^2 \ge \widetilde{\phi}^2 - \widetilde{\phi}^4/4 \ge \widetilde{\phi}^2/2.$$

Therefore for any $i \in \mathcal{I}$,

$$\mathbb{P}[|\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle| < \gamma] = \frac{1}{\sqrt{2\pi}} \int_{-[1-(\mathbf{z}_1^\top \mathbf{z}_i)^2]^{-1/2}\gamma}^{[1-(\mathbf{z}_1^\top \mathbf{z}_i)^2]^{-1/2}\gamma} \exp\Big(-\frac{1}{2}x^2\Big)\mathrm{d}x \le \sqrt{\frac{2}{\pi}} \frac{\gamma}{[1 - (\mathbf{z}_1^\top \mathbf{z}_i)^2]^{1/2}} \le \frac{2}{\sqrt{\pi}}\gamma\widetilde{\phi}^{-1}.$$

By union bound over $\mathcal{I}$, we have

$$\mathbb{P}(\mathcal{E}_2) = \mathbb{P}[|\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle| \geq \gamma, i \in \mathcal{I}] \geq 1 - \frac{2}{\sqrt{\pi}} n\gamma\widetilde{\phi}^{-1}.$$

Therefore we have

$$\mathbb{P}(\mathcal{E}) \geq \sqrt{\frac{2}{\pi e}}\gamma \cdot \bigg(1 - \frac{2}{\sqrt{\pi}} n\gamma\widetilde{\phi}^{-1}\bigg).$$

Setting $\gamma = \sqrt{\pi}\widetilde{\phi}/(4n)$, we obtain $\mathbb{P}(\mathcal{E}) \geq \widetilde{\phi}/(\sqrt{32e}n)$. Now let $\mathcal{I}' = [n] \setminus (\mathcal{I} \cup \{1\})$. Then conditioning on event $\mathcal{E}$, we have

$$\begin{aligned}
\mathbf{h}(\mathbf{w}) &= \sum_{i=1}^{n} a_i y_i \sigma'(\langle \mathbf{w}, \mathbf{z}_i\rangle)\mathbf{z}_i \\
&= a_1 y_1 \sigma'(u_1)\mathbf{z}_1 + \sum_{i\in\mathcal{I}} a_i y_i \sigma'\big(u_1\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big)\mathbf{z}_i + \sum_{i\in\mathcal{I}'} a_i y_i \sigma'\big(u_1\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}, \mathbf{z}_i\rangle\big)\mathbf{z}_i \\
&= a_1 y_1 \sigma'(u_1)\mathbf{z}_1 + \sum_{i\in\mathcal{I}} a_i y_i \sigma'\big(\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big)\mathbf{z}_i + \sum_{i\in\mathcal{I}'} a_i y_i \sigma'\big(u_1\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big)\mathbf{z}_i,
\end{aligned} \tag{4.9.2}$$

where the last equality follows from the fact that conditioning on event $\mathcal{E}$, for all $i \in \mathcal{I}$, it holds that $|\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle| \geq |u_1| \geq |u_1\langle \mathbf{z}_1, \mathbf{z}_i\rangle|$. We then consider two cases: $u_1 > 0$ and $u_1 < 0$, which occur equally likely conditioning on $\mathcal{E}$. Therefore we have

$$\mathbb{P}\bigg[\|\mathbf{h}(\mathbf{w})\|_2 \geq \inf_{u_1^{(1)}>0, u_1^{(2)}<0} \max\big\{\big\|\mathbf{h}(u_1^{(1)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}')\big\|_2, \big\|\mathbf{h}(u_1^{(2)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}')\big\|_2\big\}\bigg|\mathcal{E}\bigg] \geq 1/2.$$

By the inequality $\max\{\|\mathbf{a}\|_2, \|\mathbf{b}\|_2\} \geq \|\mathbf{a} - \mathbf{b}\|_2/2$, we have

$$\mathbb{P}\bigg[\|\mathbf{h}(\mathbf{w})\|_2 \geq \inf_{u_1^{(1)}>0, u_1^{(2)}<0} \big\|\mathbf{h}(u_1^{(1)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}') - \mathbf{h}(u_1^{(2)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}')\big\|_2/2\bigg|\mathcal{E}\bigg] \geq 1/2. \tag{4.9.3}$$

For any $u_1^{(1)} > 0$ and $u_1^{(2)} < 0$, denote $\mathbf{w}_1 = u_1^{(1)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}'$, $\mathbf{w}_2 = u_1^{(2)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}'$. We now proceed to give lower bound for $\|\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2)\|_2$. By (4.9.2), we have

$$\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2) = a_1 y_1 \mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i' y_i \mathbf{z}_i, \tag{4.9.4}$$

where

$$a_i' = a_i\big[\sigma'\big(u_1^{(1)}\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big) - \sigma'\big(u_1^{(2)}\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big)\big].$$

Note that for all $i \in \mathcal{I}'$, we have $y_i = y_1$ and $\langle \mathbf{z}_1, \mathbf{z}_i\rangle \geq 1 - \widetilde{\phi}^2/2 \geq 0$. Therefore, since $u_1^{(1)} > 0 > u_1^{(2)}$, we have

$$\sigma'\big(u_1^{(1)}\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big) - \sigma'\big(u_1^{(2)}\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big) \geq 0.$$

Therefore $a_i' \geq 0$ for all $i \in \mathcal{I}'$ and

$$\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2) = a_1 y_1 \mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i' y_1 \mathbf{z}_i = y_1\bigg(a_1\mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i'\mathbf{z}_i\bigg),$$

We have shown that $\langle \mathbf{z}_i, \mathbf{z}_1\rangle \geq 0$ for all $i \in \mathcal{I}'$. Therefore we have

$$\|\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2)\|_2 \geq \bigg\|y_1\bigg(a_1\mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i'\mathbf{z}_i\bigg)\bigg\|_2 \geq \bigg\langle a_1\mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i'\mathbf{z}_i, \mathbf{z}_1\bigg\rangle \geq a_1.$$

Since the inequality above holds for any $u_1^{(1)} > 0$ and $u_1^{(2)} < 0$, taking infimum gives

$$\inf_{u_1^{(1)}>0, u_1^{(2)}<0} \|\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2)\|_2 \geq a_1. \tag{4.9.5}$$

Plugging (4.9.5) back to (4.9.3), we obtain

$$\mathbb{P}\big[\|\mathbf{h}(\mathbf{w})\|_2 \geq a_1/2\big|\mathcal{E}\big] \geq 1/2,$$

Since $a_1 = \|\mathbf{a}\|_\infty$ and $\mathbb{P}(\mathcal{E}) \geq \widetilde{\phi}/(\sqrt{32e}n)$, we have

$$\mathbb{P}\big[\|\mathbf{h}(\mathbf{w})\|_2 \geq C\|\mathbf{a}\|_\infty\big] \geq C'\widetilde{\phi}/n,$$

where C and C' are absolute constants. This completes the proof. □

4.10 Conclusions

In this chapter, we studied training deep neural networks by gradient descent. We proved that gradient descent can achieve global minima of the training loss for over-parameterized

deep ReLU networks with random initialization, with milder assumption on the training data. Compared with the state-of-the-art results, our theoretical guarantees are sharper in terms of both over-parameterization condition and convergence rate. Our result can also be extended to stochastic gradient descent (SGD) and other loss functions (e.g., square hinge loss and smoothed hinge loss). Such extensions can be found in the longer version of this paper [ZCZ18]. In the future, we will further improve the over-parameterization condition such that it is closer to width of neural networks used in practice. Our proof technique can also be extended to other neural network architectures including convolutional neural networks (CNNs) [KSH12], residual networks (ResNets) [HZR16] and recurrent neural networks (RNNs) [HS97], and give sharper over-parameterization conditions than existing results for CNNs, ResNets [DLL19; ALS19a] and RNNs [ALS19b]. Moreover, it is also interesting to explore how our optimization guarantees of over-parameterized neural networks can be integrated with existing universal approximation ability results such as [Hor91; Tel16; LJ18; Zho19].

CHAPTER 5

Generalization of Deep ReLU Networks in the NTK Regime

5.1 Introduction

Although existing results in the neural tangent kernel regime [ALL19; ADH19a; CG19] have provided important insights into the learning of deep neural networks, they require the neural network to be extremely wide. The typical requirement on the network width is a high degree polynomial of the training sample size n and the inverse of the target error ϵ^{-1}. As there still remains a huge gap between such network width requirement and the practice, many attempts, including ours, have been made to improve the over-parameterization condition under various conditions on the training data and model initialization [OS19; ZG19; KH19; BL19]. For two-layer ReLU networks, a recent work [JT20] showed that when the training data are well separated, polylogarithmic width is sufficient to guarantee good optimization and generalization performances. However, their results cannot be extended to deep ReLU networks since their proof technique largely relies on the fact that the network model is 1-homogeneous, which cannot be satisfied by DNNs. Therefore, whether deep neural networks can be learned with such a mild over-parameterization is still an open problem.

In this work, we resolve this open problem by showing that polylogarithmic network width is sufficient to learn DNNs. In particular, unlike the existing works that require the DNNs to behave very close to a linear model (up to some small approximation error), we show that a constant linear approximation error is sufficient to establish nice optimization

and generalization guarantees for DNNs. Thanks to the relaxed requirement on the linear approximation error, a milder condition on the network width and tighter bounds on the convergence rate and generalization error can be proved. We summarize our contributions as follows:

- We establish the global convergence guarantee of GD for training deep ReLU networks based on the so-called NTRF function class [CG19], a set of linear functions over random features. Specifically, we prove that GD can learn deep ReLU networks with width $m = \text{poly}(R)$ to compete with the best function in NTRF function class, where R is the radius of the NTRF function class.

- We also establish the generalization guarantees for both GD and SGD in the same setting. Specifically, we prove a diminishing statistical error for a wide range of network width $m \in (\widetilde{\Omega}(1), \infty)$, while most of the previous generalization bounds in the NTK regime only works in the setting where the network width m is much greater than the sample size n. Moreover, we establish $\widetilde{\mathcal{O}}(\epsilon^{-2})$ and $\widetilde{\mathcal{O}}(\epsilon^{-1})$ sample complexities for GD and SGD respectively, which are tighter than existing bounds for learning deep ReLU networks [CG19], and match the best results when reduced to the two-layer cases [ADH19b; JT20].

- We further generalize our theoretical analysis to the scenarios with different data separability assumptions in the literature. We show if a large fraction of the training data are well separated, the best function in the NTRF function class with radius $R = \widetilde{\mathcal{O}}(1)$ can learn the training data with error up to ϵ. This together with our optimization and generalization guarantees immediately suggests that deep ReLU networks can be learned with network width $m = \widetilde{\Omega}(1)$, which has a logarithmic dependence on the target error ϵ and sample size n. Compared with existing results [CG20; JT20] which require all training data points to be separated in the NTK regime, our result is stronger since it allows the NTRF function class to misclassify a small proportion of the training

data.

For the ease of comparison, we summarize our results along with the most related previous results in Table 5.1, in terms of data assumption, the over-parameterization condition and sample complexity. It can be seen that under data separation assumption (See Sections 5.4.1, 5.4.2), our result improves existing results for learning deep neural networks by only requiring a polylog(n, ϵ^{-1}) network width.

Table 5.1: Comparison of neural network learning results in terms of over-parameterization condition and sample complexity. Here ϵ is the target error rate, n is the sample size, L is the network depth.

	Assumptions	Algorithm	Over-para. Condition	Sample Complexity	Network
[ZCZ19]	Data nondegeneration	GD	$\widetilde{\Omega}\big(n^{12}L^{16}(n^2+\epsilon^{-1})\big)$	-	Deep
this work	Data nondegeneration	GD	$\widetilde{\Omega}\big(L^{22}n^{12}\big)$	-	Deep
[CG20]	Data separation	GD	$\widetilde{\Omega}(\epsilon^{-14})\cdot e^{\Omega(L)}$	$\widetilde{\mathcal{O}}(\epsilon^{-4})\cdot e^{O(L)}$	Deep
[JT20]	Data separation	GD	polylog(n, ϵ^{-1})	$\widetilde{\mathcal{O}}(\epsilon^{-2})$	Shallow
this work	Data separation	GD	polylog$(n, \epsilon^{-1})\cdot$ poly(L)	$\widetilde{\mathcal{O}}(\epsilon^{-2})\cdot e^{O(L)}$	Deep
[CG19]	Data separation	SGD	$\widetilde{\Omega}(\epsilon^{-14})\cdot$ poly(L)	$\widetilde{\mathcal{O}}(\epsilon^{-2})\cdot$ poly(L)	Deep
[JT20]	Data separation	SGD	polylog(ϵ^{-1})	$\widetilde{\mathcal{O}}(\epsilon^{-1})$	Shallow
this work	Data separation	SGD	polylog$(\epsilon^{-1})\cdot$ poly(L)	$\widetilde{\mathcal{O}}(\epsilon^{-1})\cdot$ poly(L)	Deep

5.2 Preliminaries on learning neural networks

In this section, we introduce the problem setting in this work, including definitions of the neural network and loss functions, and the training algorithms, i.e., GD and SGD with random initialization.

Loss function. Given training dataset $\{\mathbf{x}_i, y_i\}_{i=1,\ldots,n}$ with input $\mathbf{x}_i \in \mathbb{R}^d$ and output $y_i \in \{-1, +1\}$, we define the training loss function as

$$L_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} L_i(\mathbf{W}),$$

Algorithm 1 Gradient descent with random initialization

Input: Number of iterations T, step size η, training set $S = \{(\mathbf{x}_i, y_i)_{i=1}^n\}$, initialization $\mathbf{W}^{(0)}$

for $t = 1, 2, \ldots, T$ **do**

Update $\mathbf{W}^{(t)} = \mathbf{W}^{(t-1)} - \eta \cdot \nabla_{\mathbf{W}} L_S(\mathbf{W}^{(t-1)})$.

end for

Output: $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(T)}$.

Algorithm 2 Stochastic gradient desecent (SGD) with random initialization

Input: Number of iterations n, step size η, initialization $\mathbf{W}^{(0)}$

for $i = 1, 2, \ldots, n$ **do**

Draw $(\mathbf{x}_i, y_i)$ from $\mathcal{D}$ and compute the corresponding gradient $\nabla_{\mathbf{W}} L_i(\mathbf{W}^{(i-1)})$.

Update $\mathbf{W}^{(i)} = \mathbf{W}^{(i-1)} - \eta \cdot \nabla_{\mathbf{W}} L_i(\mathbf{W}^{(i-1)})$.

end for

Output: Randomly choose $\widehat{\mathbf{W}}$ uniformly from $\{\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(n-1)}\}$.

where $L_i(\mathbf{W}) = \ell\big(y_i f_{\mathbf{W}}(\mathbf{x}_i)\big) = \log\big(1 + \exp(-y_i f_{\mathbf{W}}(\mathbf{x}_i))\big)$ is defined as the cross-entropy loss.

Algorithms. We consider both GD and SGD with Gaussian random initialization. The gradient descent algorithm is the same as that considered in Chapter 4. The SGD algorithm consider in this chapter uses use a new training data point in each iteration and run the algorithm for n steps. We summarize the algorithm in Algorithms 1 and 2 respectively. Specifically, the entries in $\mathbf{W}_1^{(0)}, \cdots, \mathbf{W}_{L-1}^{(0)}$ are generated independently from univariate Gaussian distribution $N(0, 2/m)$ and the entries in $\mathbf{W}_L^{(0)}$ are generated independently from $N(0, 1/m)$.

Note that our initialization method in Algorithms 1, 2 is the same as the widely used He initialization [HZR15]. Our neural network parameterization is also consistent with the parameterization used in prior work on NTK [JGH18; ALS19a; DLL19; ADH19b; CG19].

5.3 Main theory

In this section, we present the optimization and generalization guarantees of GD and SGD for learning deep ReLU networks. For simplicity, we make the following assumption on the training data points.

Assumption 5.3.1. All training data points satisfy $\|\mathbf{x}_i\|_2 = 1$, $i = 1, \ldots, n$.

This assumption has been widely made in many previous works [ALS19a; ALS19b; DZP18; DLL19; ZCZ19] in order to simplify the theoretical analysis. This assumption can be relaxed to be upper bounded and lower bounded by some constant.

In the following, we give the definition of Neural Tangent Random Feature (NTRF) [CG19], which characterizes the functions learnable by over-parameterized ReLU networks.

Definition 5.3.2 (Neural Tangent Random Feature, [CG19])**.** Let $\mathbf{W}^{(0)}$ be the initialization weights, and $F_{\mathbf{W}^{(0)},\mathbf{W}}(\mathbf{x}) = f_{\mathbf{W}^{(0)}}(\mathbf{x}) + \langle \nabla f_{\mathbf{W}^{(0)}}(\mathbf{x}), \mathbf{W} - \mathbf{W}^{(0)} \rangle$ be a function with respect to the input $\mathbf{x}$. Then the NTRF function class is defined as follows

$$\mathcal{F}(\mathbf{W}^{(0)}, R) = \big\{ F_{\mathbf{W}^{(0)},\mathbf{W}}(\cdot) : \mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, R \cdot m^{-1/2}) \big\}.$$

The function class $F_{\mathbf{W}^{(0)},\mathbf{W}}(\mathbf{x})$ consists of linear models over random features defined based on the network gradients at the initialization. Therefore it captures the key "almost linear" property of wide neural networks in the NTK regime [LXS19; CG19]. In this work, we use the NTRF function class as a reference class to measure the difficulty of a learning problem. In what follows, we deliver our main theoretical results regarding the optimization and generalization guarantees of learning deep ReLU networks. We study both GD and SGD with random initialization (presented in Algorithms 1 and 2).

5.3.1 Gradient descent

The following theorem establishes the optimization guarantee of GD for training deep ReLU networks for binary classification.

Theorem 5.3.3. For $\delta, R > 0$, let $\epsilon_{\mathrm{NTRF}} = \inf_{F \in \mathcal{F}(\mathbf{W}^{(0)}, R)} n^{-1} \sum_{i=1}^{n} \ell[y_i F(\mathbf{x}_i)]$ be the minimum training loss achievable by functions in $\mathcal{F}(\mathbf{W}^{(0)}, R)$. Then there exists

$$m^*(\delta, R, L) = \widetilde{\mathcal{O}}\big(\mathrm{poly}(R, L) \cdot \log^{4/3}(n/\delta)\big),$$

such that if $m \geq m^*(\delta, R, L)$, with probability at least $1 - \delta$ over the initialization, GD with step size $\eta = \Theta(L^{-1} m^{-1})$ can train a neural network to achieve at most $3\epsilon_{\mathrm{NTRF}}$ training loss within $T = \mathcal{O}\big(L^2 R^2 \epsilon_{\mathrm{NTRF}}^{-1}\big)$ iterations.

Theorem 5.3.3 shows that the deep ReLU network trained by GD can compete with the best function in the NTRF function class $\mathcal{F}(\mathbf{W}^{(0)}, R)$ if the network width has a polynomial dependency in R and L and a logarithmic dependency in n and $1/\delta$. Moreover, if the NTRF function class with $R = \widetilde{\mathcal{O}}(1)$ can learn the training data well (i.e., ϵ_{NTRF} is less than a small target error ϵ), a polylogarithmic (in terms of n and ϵ^{-1}) network width suffices to guarantee the global convergence of GD, which directly improves over-paramterization condition in the most related work [CG19]. Besides, we remark here that this assumption on the NTRF function class can be easily satisfied when the training data admits certain separability conditions, which we discuss in detail in Section 5.4.

Compared with the results in [JT20] which give similar network width requirements for two-layer networks, our result works for deep networks. Moreover, while [JT20] essentially required all training data to be separable by a function in the NTRF function class with a constant margin, our result does not require such data separation assumptions, and allows the NTRF function class to misclassify a small proportion of the training data points[1].

We now characterize the generalization performance of neural networks trained by GD. We denote $L_{\mathcal{D}}^{0-1}(\mathbf{W}) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}}[\mathbb{1}\{f_{\mathbf{W}}(\mathbf{x}) \cdot y < 0\}]$ as the expected 0-1 loss (i.e., expected error) of $f_{\mathbf{W}}(\mathbf{x})$.

[1] A detailed discussion is given in Section 5.4.2.

Theorem 5.3.4. Under the same assumptions as Theorem 5.3.3, with probability at least $1-\delta$, the iterate $\mathbf{W}^{(t)}$ of Algorithm 1 satisfies that

$$L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) \leq 2L_S(\mathbf{W}^{(t)}) + \widetilde{\mathcal{O}}\left(4^L L^2 R\sqrt{\frac{m}{n}} \wedge \left(\frac{L^{3/2}R}{\sqrt{n}} + \frac{L^{11/3}R^{4/3}}{m^{1/6}}\right)\right) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

for all $t = 0, \ldots, T$.

Theorem 5.3.4 shows that the test error of the trained neural network can be bounded by its training error plus statistical error terms. Note that the statistical error terms is in the form of a minimum between two terms $4^L L^2 R\sqrt{m/n}$ and $L^{3/2}R/\sqrt{n} + L^{11/3}R^{4/3}/m^{1/6}$. Depending on the network width m, one of these two terms will be the dominating term and diminishes for large n: (1) if $m = o(n)$, the statistical error will be $4^L L^2 R\sqrt{m/n}$, and diminishes as n increases; and (2) if $m = \Omega(n)$, the statistical error is $L^{3/2}R/\sqrt{n} + L^{11/3}R^{4/3}/m^{1/6}$, and again goes to zero as n increases. Moreover, in this work we have a specific focus on the setting $m = \widetilde{\mathcal{O}}(1)$, under which Theorem 5.3.4 gives a statistical error of order $\widetilde{\mathcal{O}}(n^{-1/2})$. This distinguishes our result from previous generalization bounds for deep networks [CG20; CG19], which cannot be applied to the setting $m = \widetilde{\mathcal{O}}(1)$.

We note that for two-layer ReLU networks (i.e., $L = 2$) [JT20] proves a tighter $\widetilde{O}(1/n^{1/2})$ generalization error bound regardless of the neural networks width m, while our result (Theorem 5.3.4), in the two-layer case, can only give $\widetilde{O}(1/n^{1/2})$ generalization error bound when $m = \widetilde{O}(1)$ or $m = \widetilde{\Omega}(n^3)$. However, different from our proof technique that basically uses the (approximated) linearity of the neural network function, their proof technique largely relies on the 1-homogeneous property of the neural network, which restricted their theory in two-layer cases. An interesting research direction is to explore whether a $\widetilde{O}(1/n^{1/2})$ generalization error bound can be also established for deep networks (regardless of the network width), which we will leave it as a future work.

5.3.2 Stochastic gradient descent

Here we study the performance of SGD for training deep ReLU networks. The following theorem establishes a generalization error bound for the output of SGD.

Theorem 5.3.5. For $\delta, R > 0$, let $\epsilon_{\mathrm{NTRF}} = \inf_{F \in \mathcal{F}(\mathbf{W}^{(0)}, R)} n^{-1} \sum_{i=1}^{n} \ell[y_i F(\mathbf{x}_i)]$ be the minimum training loss achievable by functions in $\mathcal{F}(\mathbf{W}^{(0)}, R)$. Then there exists

$$m^*(\delta, R, L) = \widetilde{\mathcal{O}}\big(\mathrm{poly}(R, L) \cdot \log^{4/3}(n/\delta)\big),$$

such that if $m \geq m^*(\delta, R, L)$, with probability at least $1 - \delta$, SGD with step size $\eta = \Theta\big(m^{-1} \cdot (LR^2 n^{-1} \epsilon_{\mathrm{NTRF}}^{-1} \wedge L^{-1})\big)$ achieves

$$\mathbb{E}[L_{\mathcal{D}}^{0-1}(\widehat{\mathbf{W}})] \leq \frac{8L^2R^2}{n} + \frac{8\log(2/\delta)}{n} + 24\epsilon_{\mathrm{NTRF}},$$

where the expectation is taken over the uniform draw of $\widehat{\mathbf{W}}$ from $\{\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(n-1)}\}$.

For any $\epsilon > 0$, Theorem 5.3.5 gives a $\widetilde{\mathcal{O}}(\epsilon^{-1})$ sample complexity for deep ReLU networks trained with SGD to achieve $O(\epsilon_{\mathrm{NTRF}} + \epsilon)$ test error. Our result extends the result for two-layer networks proved in [JT20] to multi-layer networks. Theorem 5.3.5 also provides sharper results compared with [ALL19; CG19] in two aspects: (1) the sample complexity is improved from $n = \widetilde{\mathcal{O}}(\epsilon^{-2})$ to $n = \widetilde{\mathcal{O}}(\epsilon^{-1})$; and (2) the overparamterization condition is improved from $m \geq \mathrm{poly}(\epsilon^{-1})$ to $m = \widetilde{\Omega}(1)$.

5.4 Discussion on the NTRF Class

Our theoretical results in Section 5.3 rely on the radius (i.e., R) of the NTRF function class $\mathcal{F}(\mathbf{W}^{(0)}, R)$ and the minimum training loss achievable by functions in $\mathcal{F}(\mathbf{W}^{(0)}, R)$, i.e., ϵ_{NTRF}. Note that a larger R naturally implies a smaller ϵ_{NTRF}, but also leads to worse conditions on m. In this section, for any (arbitrarily small) target error rate $\epsilon > 0$, we discuss various data assumptions studied in the literature under which our results can lead to $\mathcal{O}(\epsilon)$ training/test errors, and specify the network width requirement.

5.4.1 Data Separability by Neural Tangent Random Feature

In this subsection, we consider the setting where a large fraction of the training data can be linearly separated by the neural tangent random features. The assumption is stated as follows.

Assumption 5.4.1. There exists a collection of matrices $\mathbf{U}^* = \{\mathbf{U}_1^*, \cdots, \mathbf{U}_L^*\}$ satisfying $\sum_{l=1}^{L} \|\mathbf{U}_l^*\|_F^2 = 1$, such that for at least $(1-\rho)$ fraction of training data we have

$$y_i \langle \nabla f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \mathbf{U}^* \rangle \geq m^{1/2}\gamma,$$

where γ is an absolute positive constant[2] and $\rho \in [0, 1)$.

The following corollary provides an upper bound of ϵ_{NTRF} under Assumption 5.4.1 for some R.

Proposition 5.4.2. Under Assumption 5.4.1, for any $\epsilon, \delta > 0$, if $R \geq C\big[\log^{1/2}(n/\delta) + \log(1/\epsilon)\big]/\gamma$ for some absolute constant C, then with probability at least $1-\delta$,

$$\epsilon_{\text{NTRF}} := \inf_{F \in \mathcal{F}(\mathbf{W}^{(0)}, R)} n^{-1} \sum_{i=1}^{n} \ell\big(y_i F(\mathbf{x}_i)\big) \leq \epsilon + \rho \cdot \mathcal{O}(R).$$

Proposition 5.4.2 covers the setting where the NTRF function class is allowed to misclassify training data, while most of existing work typically assumes that all training data can be perfectly separated with constant margin (i.e., $\rho = 0$) [JT20; Sha21]. Our results show that for sufficiently small misclassification ratio $\rho = \mathcal{O}(\epsilon)$, we have $\epsilon_{\text{NTRF}} = \widetilde{O}(\epsilon)$ by choosing the radius parameter R logarithimic in n, δ^{-1}, and ϵ^{-1}. Substituting this result into Theorems 5.3.3, 5.3.4 and 5.3.5, it can be shown that a neural network with width $m = \text{poly}(L, \log(n/\delta), \log(1/\epsilon))\big)$ suffices to guarantee good optimization and generalization performances for both GD and SGD. Consequently, we can obtain that the bounds on the test error for GD and SGD are $\widetilde{\mathcal{O}}(n^{-1/2})$ and $\widetilde{\mathcal{O}}(n^{-1})$ respectively.

[2]The factor $m^{1/2}$ is introduced here since $\|\nabla_{\mathbf{W}^{(0)}} f(\mathbf{x}_i)\|_F$ is typically of order $O(m^{1/2})$.

5.4.2 Data Separability by Shallow Neural Tangent Model

In this subsection, we study the data separation assumption made in [JT20] and show that our results cover this particular setting. We first restate the assumption as follows.

Assumption 5.4.3. There exists $\overline{\mathbf{u}}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ and $\gamma \geq 0$ such that $\|\overline{\mathbf{u}}(\mathbf{z})\|_2 \leq 1$ for all $\mathbf{z} \in \mathbb{R}^d$, and

$$y_i \int_{\mathbb{R}^d} \sigma'(\langle \mathbf{z}, \mathbf{x}_i \rangle) \cdot \langle \bar{\mathbf{u}}(\mathbf{z}), \mathbf{x}_i \rangle \mathrm{d}\mu_N(\mathbf{z}) \geq \gamma$$

for all $i \in [n]$, where $\mu_N(\cdot)$ denotes the standard normal distribution.

Assumption 5.4.3 is related to the linear separability of the gradients of the first layer parameters at random initialization, where the randomness is replaced with an integral by taking the infinite width limit. Note that similar assumptions have also been studied in [CG20; NS19; FCG19]. The assumption made in [CG20; FCG19] uses gradients with respect to the second layer weights instead of the first layer ones. In the following, we mainly focus on Assumption 5.4.3, while our result can also be generalized to cover the setting in [CG20; FCG19].

In order to make a fair comparison, we reduce our results for multilayer networks to the two-layer setting. In this case, the neural network function takes form

$$f_{\mathbf{W}}(\mathbf{x}) = m^{1/2} \mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x}).$$

Then we provide the following proposition, which states that Assumption 5.4.3 implies a certain choice of $R = \widetilde{\mathcal{O}}(1)$ such the the minimum training loss achieved by the function in the NTRF function class $\mathcal{F}(\mathbf{W}^{(0)}, R)$ satisfies $\epsilon_{\mathrm{NTRF}} = O(\epsilon)$, where ϵ is the target error.

Proposition 5.4.4. Suppose the training data satisfies Assumption 5.4.3. For any $\epsilon, \delta > 0$, let $R = C\big[\log(n/\delta) + \log(1/\epsilon)\big]/\gamma$ for some large enough absolute constant C. If the neural network width satisfies $m = \Omega\big(\log(n/\delta)/\gamma^2\big)$, then with probability at least $1 - \delta$, there exist $F_{\mathbf{W}^{(0)}, \overline{\mathbf{W}}}(\mathbf{x}_i) \in \mathcal{F}(\mathbf{W}^{(0)}, R)$ such that $\ell\big(y_i \cdot F_{\mathbf{W}^{(0)}, \overline{\mathbf{W}}}(\mathbf{x}_i)\big) \leq \epsilon, \forall i \in [n]$.

Proposition 5.4.4 shows that under Assumption 5.4.3, there exists $F_{\mathbf{W}^{(0)},\overline{\mathbf{W}}}(\cdot) \in \mathcal{F}(\mathbf{W}^{(0)}, R)$ with $R = \widetilde{\mathcal{O}}(1/\gamma)$ such that the cross-entropy loss of $F_{\mathbf{W}^{(0)},\overline{\mathbf{W}}}(\cdot)$ at each training data point is bounded by ϵ. This implies that $\epsilon_{\text{NTRF}} \leq \epsilon$. Moreover, by applying Theorem 5.3.3 with $L = 2$, the condition on the neural network width becomes $m = \widetilde{\Omega}(1/\gamma^8)$[3], which matches the results proved in [JT20]. Moreover, plugging these results on m and ϵ_{NTRF} into Theorems 5.3.4 and 5.3.5, we can conclude that the bounds on the test error for GD and SGD are $\widetilde{\mathcal{O}}(n^{-1/2})$ and $\widetilde{\mathcal{O}}(n^{-1})$ respectively.

5.4.3 Class-dependent Data Nondegeneration

In previous subsections, we have shown that under certain data separation conditions ϵ_{NTRF} can be sufficiently small while the corresponding NTRF function class has R of order $\widetilde{\mathcal{O}}(1)$. Thus neural networks with polylogarithmic width enjoy nice optimization and generalization guarantees. In this part, we consider the following much milder data separability assumption made in [ZCZ19].

Assumption 5.4.5. For all $i \neq i'$ if $y_i \neq y_{i'}$, then $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \geq \phi$ for some absolute constant ϕ.

In contrast to the conventional data nondegeneration assumption (i.e., no duplicate data points) made in [ALS19a; DZP18; DLL19; ZG19][4], Assumption 5.4.5 only requires that the data points from different classes are nondegenerate, thus we call it class-dependent data nondegeneration.

We have the following proposition which shows that Assumption 5.4.5 also implies the existence of a good function that achieves ϵ training error, in the NTRF function class with a certain choice of R.

[3]We have shown in the proof of Theorem 5.3.3 that $m = \widetilde{\Omega}(R^8)$ (see (5.7.1) for more detail).

[4]Specifically, [ALS19a; ZG19] require that any two data points (rather than data points from different classes) are separated by a positive distance. [ZG19] shows that this assumption is equivalent to those made in [DZP18; DLL19], which require that the composite kernel matrix is strictly positive definite.

Proposition 5.4.6. Under Assumption 5.4.5, if

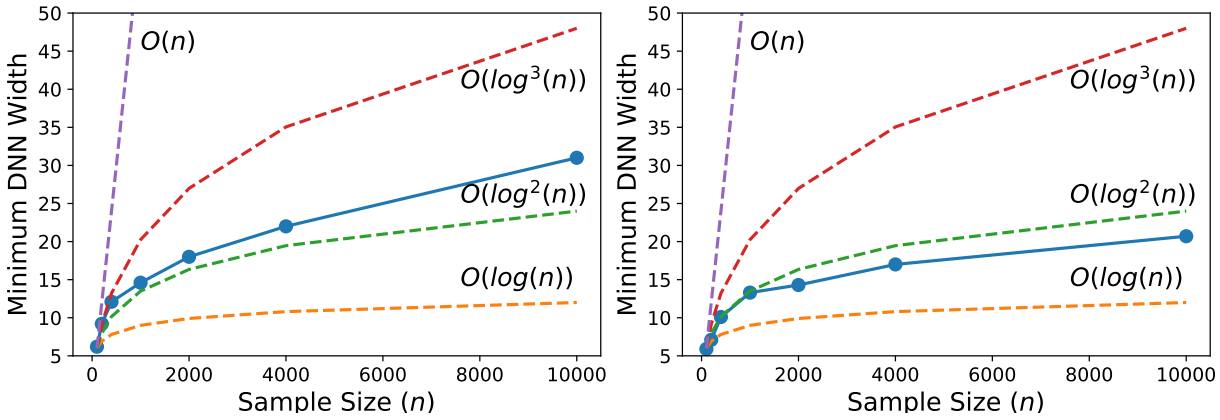
$$R = \Omega\big(n^{3/2}\phi^{-1/2}\log(n\delta^{-1}\epsilon^{-1})\big), \qquad m = \widetilde{\Omega}\big(L^{22}n^{12}\phi^{-4}\big),$$

we have $\epsilon_{\text{NTRF}} \leq \epsilon$ with probability at least $1-\delta$.

Proposition 5.4.6 suggests that under Assumption 5.4.5, in order to guarantee $\epsilon_{\text{NTRF}} \leq \epsilon$, the size of NTRF function class needs to be $\Omega(n^{3/2})$. Plugging this into Theorems 5.3.4 and 5.3.5 leads to vacuous bounds on the test error. This makes sense since Assumption 5.4.5 basically covers the "random label" setting, which is impossible to be learned with small generalization error. Moreover, we would like to point out our theoretical analysis leads to a sharper over-parameterization condition than that proved in [ZCZ19], i.e., $m = \widetilde{\Omega}\big(n^{14}L^{16}\phi^{-4} + n^{12}L^{16}\phi^{-4}\epsilon^{-1}\big)$, if the network depth satisfies $L \leq \widetilde{\mathcal{O}}(n^{1/3} \vee \epsilon^{-1/6})$.

5.5 Experiments

In this section, we conduct some simple experiments to validate our theory. Since our paper mainly focuses on binary classification, we use a subset of the original CIFAR10 dataset [Kri09], which only has two classes of images. We train a 5-layer fully-connected ReLU network on this binary classification dataset with different sample sizes, and plot the minimal neural network width that is required to achieve zero training error in Figure 5.1 (solid line). We also plot $\mathcal{O}(n), \mathcal{O}(\log^3(n)), \mathcal{O}(\log^2(n))$ and $\mathcal{O}(\log(n))$ in dashed line for reference. It is evident that the required network width to achieve zero training error is polylogarithmic on the sample size n, which is consistent with our theory.



(a) "cat" vs. "dog"

(b) "cat" vs. "ship"

Figure 5.1: Minimum network width that is required to achieve zero training error with respect to the training sample size (blue solid line). The hidden constants in all $O(\cdot)$ notations are adjusted to ensure their plots (dashed lines) start from the same point.

5.6 Proof sketch of the main theory

In this section, we introduce a key technical lemma in Section 5.6.1, based on which we provide a proof sketch of Theorems 5.3.3.

5.6.1 A key technical lemma

Here we introduce a key technical lemma used in the proof of Theorem 5.3.3.

Our proof is based on the key observation that near initialization, the neural network function can be approximated by its first-order Taylor expansion. In the following, we first give the definition of the linear approximation error in a τ-neighborhood around initialization.

$$\epsilon_{\text{app}}(\tau) := \sup_{i=1,\ldots,n} \sup_{\mathbf{W}', \mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)} \big| f_{\mathbf{W}'}(\mathbf{x}_i) - f_{\mathbf{W}}(\mathbf{x}_i) - \langle \nabla f_{\mathbf{W}}(\mathbf{x}_i), \mathbf{W}' - \mathbf{W} \rangle \big|.$$

If all the iterates of GD stay inside a neighborhood around initialization with small linear approximation error, then we may expect that the training of neural networks should be similar to the training of the corresponding linear model, where standard optimization techniques

can be applied. Motivated by this, we also give the following definition on the gradient upper bound of neural networks around initialization, which is related to the Lipschitz constant of the optimization objective function.

$$M(\tau) := \sup_{i=1,\ldots,n} \sup_{l=1,\ldots,L} \sup_{\mathbf{W}\in\mathcal{B}(\mathbf{W}^{(0)},\tau)} \|\nabla_{\mathbf{W}_l} f_{\mathbf{W}}(\mathbf{x}_i)\|_F.$$

By definition, we can choose $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, Rm^{-1/2})$ such that $n^{-1}\sum_{i=1}^n \ell\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}^*}(\mathbf{x}_i)\big) = \epsilon_{\mathrm{NTRF}}$. Then we have the following lemma.

Lemma 5.6.1. Set $\eta = \mathcal{O}(L^{-1}M(\tau)^{-2})$. Suppose that $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ and $\mathbf{W}^{(t)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ for all $0 \le t \le t'-1$. Then it holds that

$$\frac{1}{t'}\sum_{t=0}^{t'-1} L_{\mathcal{S}}(\mathbf{W}^{(t)}) \le \frac{\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t')} - \mathbf{W}^*\|_F^2 + 2t'\eta\epsilon_{\mathrm{NTRF}}}{t'\eta\big(\frac{3}{2} - 4\epsilon_{\mathrm{app}}(\tau)\big)}.$$

Lemma 5.6.1 plays a central role in our proof. In specific, if $\mathbf{W}^{(t)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ for all $t \le t'$, then Lemma 5.6.1 implies that the average training loss is in the same order of ϵ_{NTRF} as long as the linear approximation error $\epsilon_{\mathrm{app}}(\tau)$ is bounded by a positive constant. This is in contrast to the proof in [CG19], where $\epsilon_{\mathrm{app}}(\tau)$ appears as an additive term in the upper bound of the training loss, thus requiring $\epsilon_{\mathrm{app}}(\tau) = \mathcal{O}(\epsilon_{\mathrm{NTRF}})$ to achieve the same error bound as in Lemma 5.6.1. Since we can show that $\epsilon_{\mathrm{app}} = \widetilde{\mathcal{O}}(m^{-1/6})$ (See Section 5.7.1), this suggests that $m = \widetilde{\Omega}(1)$ is sufficient to make the average training loss in the same order of ϵ_{NTRF}.

Compared with the recent results for two-layer networks by [JT20], Lemma 5.6.1 is proved with different techniques. In specific, the proof by [JT20] relies on the 1-homogeneous property of the ReLU activation function, which limits their analysis to two-layer networks with fixed second layer weights. In comparison, our proof does not rely on homogeneity, and is purely based on the linear approximation property of neural networks and some specific properties of the loss function. Therefore, our proof technique can handle deep networks, and is potentially applicable to non-ReLU activation functions and other network architectures (e.g, Convolutional neural networks and Residual networks).

5.6.2 Proof sketch of Theorem 5.3.3

Here we provide a proof sketch of Theorem 5.3.3. The proof consists of two steps: (i) showing that all T iterates stay close to initialization, and (ii) bounding the empirical loss achieved by gradient descent. Both of these steps are proved based on Lemma 5.6.1.

Proof sketch of Theorem 5.3.3. Recall that we choose $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, Rm^{-1/2})$ such that $n^{-1}\sum_{i=1}^n \ell\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}^*}(\mathbf{x}_i)\big) = \epsilon_{\mathrm{NTRF}}$. We set $\tau = \widetilde{\mathcal{O}}(L^{1/2}m^{-1/2}R)$, which is chosen slightly larger than $m^{-1/2}R$ since Lemma 5.6.1 requires the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$ to include both $\mathbf{W}^*$ and $\{\mathbf{W}^{(t)}\}_{t=0,\ldots,t'}$. Then by Lemmas 4.1 and B.3 in [CG19] we know that $\epsilon_{\mathrm{app}}(\tau) = \widetilde{\mathcal{O}}(\tau^{4/3}m^{1/2}L^3) = \widetilde{\mathcal{O}}(R^{4/3}L^{11/3}m^{-1/6})$. Therefore, we can set $m = \widetilde{\Omega}(R^8L^{22})$ to ensure that $\epsilon_{\mathrm{app}}(\tau) \leq 1/8$.

Then we proceed to show that all iterates stay inside the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Since the L.H.S. of Lemma 5.6.1 is strictly positive and $\epsilon_{\mathrm{app}}(\tau) \leq 1/8$, we have for all $t \leq T$,

$$\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 \geq -2t\eta\epsilon_{\mathrm{NTRF}},$$

which gives an upper bound of $\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F$. Then by the choice of η, T, triangle inequality, and a simple induction argument, we see that $\|\mathbf{W}^{(t)} - \mathbf{W}^{(0)}\|_F \leq m^{-1/2}R + \sqrt{2T\eta\epsilon_{\mathrm{NTRF}}} = \widetilde{\mathcal{O}}(L^{1/2}m^{-1/2}R)$, which verifies that $\mathbf{W}^{(t)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ for $t = 0, \ldots, T-1$.

The second step is to show that GD can find a neural network with at most $3\epsilon_{\mathrm{NTRF}}$ training loss within T iterations. To show this, by the bound given in Lemma 5.6.1 with $\epsilon_{\mathrm{app}} \leq 1/8$, we drop the terms $\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2$ and rearrange the inequality to obtain

$$\frac{1}{T}\sum_{t=0}^{T-1} L_S(\mathbf{W}^{(t)}) \leq \frac{1}{\eta T}\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 + 2\epsilon_{\mathrm{NTRF}}.$$

We see that T is large enough to ensure that the first term in the bound above is smaller than ϵ_{NTRF}. This implies that the best iterate among $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(T-1)}$ achieves an empirical loss at most $3\epsilon_{\mathrm{NTRF}}$. □

5.7 Proof of Main Theorems

In this section we provide the full proof of Theorems 5.3.3, 5.3.4 and 5.3.5.

5.7.1 Proof of Theorem 5.3.3

We first provide the following lemma which is useful in the subsequent proof.

Lemma 5.7.1 (Lemmas 4.1 and B.3 in [CG19])**.** There exists an absolute constant κ such that, with probability at least $1 - \mathcal{O}(nL^2)\exp[-\Omega(m\tau^{2/3}L)]$, for any $\tau \leq \kappa L^{-6}[\log(m)]^{-3/2}$, it holds that

$$\epsilon_{\text{app}}(\tau) \leq \widetilde{\mathcal{O}}\big(\tau^{4/3}L^3m^{1/2}\big), \quad M(\tau) \leq \widetilde{\mathcal{O}}(\sqrt{m}).$$

Proof of Theorem 5.3.3. Recall that $\mathbf{W}^*$ is chosen such that

$$\frac{1}{n}\sum_{i=1}^{n}\ell\big(y_iF_{\mathbf{W}^{(0)},\mathbf{W}^*}(\mathbf{x}_i)\big) = \epsilon_{\text{NTRF}}$$

and $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, Rm^{-1/2})$. Note that to apply Lemma 5.6.1, we need the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$ to include both $\mathbf{W}^*$ and $\{\mathbf{W}^{(t)}\}_{t=0,\ldots,t'}$. This motivates us to set $\tau = \widetilde{\mathcal{O}}(L^{1/2}m^{-1/2}R)$, which is slightly larger than $m^{-1/2}R$. With this choice of τ, by Lemma 5.7.1 we have $\epsilon_{\text{app}}(\tau) = \widetilde{\mathcal{O}}(\tau^{4/3}m^{1/2}L^3) = \widetilde{\mathcal{O}}(R^{4/3}L^{11/3}m^{-1/6})$. Therefore, we can set

$$m = \widetilde{\Omega}(R^8L^{22}) \tag{5.7.1}$$

to ensure that $\epsilon_{\text{app}}(\tau) \leq 1/8$, where $\widetilde{\Omega}(\cdot)$ hides polylogarithmic dependencies on network depth L, NTRF function class size R, and failure probability parameter δ. Then by Lemma 5.6.1, we have with probability at least $1-\delta$, we have

$$\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t')} - \mathbf{W}^*\|_F^2 \geq \eta\sum_{t=0}^{t'-1}L_S(\mathbf{W}^{(t)}) - 2t'\eta\epsilon_{\text{NTRF}} \tag{5.7.2}$$

as long as $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(t'-1)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$. In the following proof we choose $\eta = \Theta(L^{-1}m^{-1})$ and $T = \lceil LR^2m^{-1}\eta^{-1}\epsilon_{\text{NTRF}}^{-1}\rceil$.

We prove the theorem by two steps: 1) we show that all iterates $\{\mathbf{W}^{(0)}, \cdots, \mathbf{W}^{(T)}\}$ will stay inside the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$; and 2) we show that GD can find a neural network with at most $3\epsilon_{\text{NTRF}}$ training loss within T iterations.

All iterates stay inside $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$. We prove this part by induction. Specifically, given $t' \leq T$, we assume the hypothesis $\mathbf{W}^{(t)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ holds for all $t < t'$ and prove that $\mathbf{W}^{(t')} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$. First, it is clear that $\mathbf{W}^{(0)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Then by (5.7.2) and the fact that $L_S(\mathbf{W}) \geq 0$, we have

$$\|\mathbf{W}^{(t')} - \mathbf{W}^*\|_F^2 \leq \|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 + 2\eta t' \epsilon_{\text{NTRF}}$$

Note that $T = \lceil LR^2 m^{-1} \eta^{-1} \epsilon_{\text{NTRF}}^{-1} \rceil$ and $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, R \cdot m^{-1/2})$, we have

$$\sum_{l=1}^{L} \|\mathbf{W}_l^{(t')} - \mathbf{W}_l^*\|_F^2 = \|\mathbf{W}^{(t')} - \mathbf{W}^*\|_F^2 \leq CLR^2 m^{-1},$$

where $C \geq 4$ is an absolute constant. Therefore, by triangle inequality, we further have the following for all $l \in [L]$,

$$\begin{aligned} \|\mathbf{W}_l^{(t')} - \mathbf{W}_l^{(0)}\|_F &\leq \|\mathbf{W}_l^{(t')} - \mathbf{W}_l^*\|_F + \|\mathbf{W}_l^{(0)} - \mathbf{W}_l^*\|_F \\ &\leq \sqrt{CL} R m^{-1/2} + R m^{-1/2} \\ &\leq 2\sqrt{CL} R m^{-1/2}. \end{aligned} \tag{5.7.3}$$

Therefore, it is clear that $\|\mathbf{W}_l^{(t')} - \mathbf{W}_l^{(0)}\|_F \leq 2\sqrt{CL} R m^{-1/2} \leq \tau$ based on our choice of τ previously. This completes the proof of the first part.

Convergence of gradient descent. (5.7.2) implies

$$\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(T)} - \mathbf{W}^*\|_F^2 \geq \eta \bigg(\sum_{t=0}^{T-1} L_S(\mathbf{W}^{(t)}) - 2T\epsilon_{\text{NTRF}} \bigg).$$

Dividing by ηT on the both sides, we get

$$\frac{1}{T} \sum_{t=0}^{T-1} L_S(\mathbf{W}^{(t)}) \leq \frac{\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2}{\eta T} + 2\epsilon_{\text{NTRF}} \leq \frac{LR^2 m^{-1}}{\eta T} + 2\epsilon_{\text{NTRF}} \leq 3\epsilon_{\text{NTRF}},$$

where the second inequality is by the fact that $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, R \cdot m^{-1/2})$ and the last inequality is by our choices of T and η which ensure that $T\eta \geq LR^2 m^{-1}\epsilon_{\text{NTRF}}^{-1}$. Notice that $T = \lceil LR^2 m^{-1}\eta^{-1}\epsilon_{\text{NTRF}}^{-1} \rceil = \mathcal{O}(L^2 R^2 \epsilon_{\text{NTRF}}^{-1})$. This completes the proof of the second part, and we are able to complete the proof. □

5.7.2 Proof of Theorem 5.3.4

Following [CG20], we first introduce the definition of surrogate loss of the network, which is defined by the derivative of the loss function.

Definition 5.7.2. We define the empirical surrogate error $\mathcal{E}_S(\mathbf{W})$ and population surrogate error $\mathcal{E}_{\mathcal{D}}(\mathbf{W})$ as follows:

$$\mathcal{E}_S(\mathbf{W}) := -\frac{1}{n}\sum_{i=1}^{n} \ell'\big[y_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)\big], \quad \mathcal{E}_{\mathcal{D}}(\mathbf{W}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big\{ -\ell'\big[y \cdot f_{\mathbf{W}}(\mathbf{x})\big]\big\}.$$

The following lemma gives uniform-convergence type of results for $\mathcal{E}_S(\mathbf{W})$ utilizing the fact that $-\ell'(\cdot)$ is bounded and Lipschitz continuous.

Lemma 5.7.3. For any $\widetilde{R}, \delta > 0$, suppose that $m = \widetilde{\Omega}(L^{12}\widetilde{R}^2) \cdot [\log(1/\delta)]^{3/2}$. Then with probability at least $1-\delta$, it holds that

$$|\mathcal{E}_{\mathcal{D}}(\mathbf{W}) - \mathcal{E}_S(\mathbf{W})| \leq \widetilde{\mathcal{O}}\bigg(\min\bigg\{ 4^L L^{3/2}\widetilde{R}\sqrt{\frac{m}{n}}, \frac{L\widetilde{R}}{\sqrt{n}} + \frac{L^3\widetilde{R}^{4/3}}{m^{1/6}} \bigg\}\bigg) + \mathcal{O}\bigg(\sqrt{\frac{\log(1/\delta)}{n}}\bigg)$$

for all $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \widetilde{R} \cdot m^{-1/2})$

We are now ready to prove Theorem 5.3.4, which combines the trajectory distance analysis in the proof of Theorem 5.3.3 with Lemma 5.7.3.

Proof of Theorem 5.3.4. With exactly the same proof as Theorem 5.3.3, by (5.7.3) and induction we have $\mathbf{W}^{(0)}, \mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(T)} \in \mathcal{B}(\mathbf{W}^{(0)}, \widetilde{R}m^{-1/2})$ with $\widetilde{R} = \mathcal{O}(\sqrt{L}R)$. Therefore by Lemma 5.7.3, we have

$$|\mathcal{E}_{\mathcal{D}}(\mathbf{W}^{(t)}) - \mathcal{E}_S(\mathbf{W}^{(t)})| \leq \widetilde{\mathcal{O}}\bigg(\min\bigg\{ 4^L L^2 R\sqrt{\frac{m}{n}}, \frac{L^{3/2}R}{\sqrt{n}} + \frac{L^{11/3}R^{4/3}}{m^{1/6}} \bigg\}\bigg) + \mathcal{O}\bigg(\sqrt{\frac{\log(1/\delta)}{n}}\bigg)$$

for all $t = 0, 1, \ldots, T$. Note that we have $\mathbb{1}\{z < 0\} \leq -2\ell'(z)$. Therefore,

$$\begin{aligned}\mathbb{E}L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) &\leq 2\mathcal{E}_{\mathcal{D}}(\mathbf{W}^{(t)}) \\ &\leq 2L_S(\mathbf{W}^{(t)}) + \widetilde{\mathcal{O}}\bigg(\min\bigg\{4^L L^2 R\sqrt{\frac{m}{n}}, \frac{L^{3/2}R}{\sqrt{n}} + \frac{L^{11/3}R^{4/3}}{m^{1/6}}\bigg\}\bigg) + \mathcal{O}\bigg(\sqrt{\frac{\log(1/\delta)}{n}}\bigg)\end{aligned}$$

for $t = 0, 1, \ldots, T$, where the last inequality is by $\mathcal{E}_S(\mathbf{W}) \leq L_S(\mathbf{W})$ because $-\ell'(z) \leq \ell(z)$ for all $z \in R$. This finishes the proof. □

5.7.3 Proof of Theorem 5.3.5

In this section we provide the full proof of Theorem 5.3.5. We first give the following result, which is the counterpart of Lemma 5.6.1 for SGD. Again we pick $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, Rm^{-1/2})$ such that the loss of the corresponding NTRF model $F_{\mathbf{W}^{(0)},\mathbf{W}^*}(\mathbf{x})$ achieves ϵ_{NTRF}.

Lemma 5.7.4. Set $\eta = \mathcal{O}(L^{-1}M(\tau)^{-2})$. Suppose that $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ and $\mathbf{W}^{(n')} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ for all $0 \leq n' \leq n-1$. Then it holds that

$$\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(n')} - \mathbf{W}^*\|_F^2 \geq \Big(\frac{3}{2} - 4\epsilon_{\text{app}}(\tau)\Big)\eta\sum_{i=1}^{n'} L_i(\mathbf{W}^{(i-1)}) - 2n\eta\epsilon_{\text{NTRF}}.$$

We introduce a surrogate loss $\mathcal{E}_i(\mathbf{W}) = -\ell'[y_i \cdot f_{\mathbf{W}}(\boldsymbol{x}_i)]$ and its population version $\mathcal{E}_{\mathcal{D}}(\mathbf{W}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[-\ell'[y \cdot f_{\mathbf{W}}(\boldsymbol{x})]]$, which have been used in [JT19; CG19; JT20]. Our proof is based on the application of Lemma 5.7.4 and an online-to-batch conversion argument [CCG04; CG19; JT20]. We introduce a surrogate loss $\mathcal{E}_i(\mathbf{W}) = -\ell'[y_i \cdot f_{\mathbf{W}}(\boldsymbol{x}_i)]$ and its population version $\mathcal{E}_{\mathcal{D}}(\mathbf{W}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[-\ell'(y \cdot f_{\mathbf{W}}(\boldsymbol{x}))]$, which have been used in [JT19; CG19; NS19; JT20].

Proof of Theorem 5.3.5. Recall that $\mathbf{W}^*$ is chosen such that

$$\frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}^*}(\mathbf{x}_i)\big) = \epsilon_{\text{NTRF}}$$

and $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, Rm^{-1/2})$. To apply Lemma 5.7.4, we need the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$ to include both $\mathbf{W}^*$ and $\{\mathbf{W}^{(t)}\}_{t=0,\ldots,t'}$. This motivates us to set $\tau = \widetilde{\mathcal{O}}(L^{1/2}m^{-1/2}R)$, which

is slightly larger than $m^{-1/2}R$. With this choice of τ, by Lemma 5.7.1 we have $\epsilon_{\text{app}}(\tau) = \widetilde{\mathcal{O}}(\tau^{4/3}m^{1/2}L^3) = \widetilde{\mathcal{O}}(R^{4/3}L^{11/3}m^{-1/6})$. Therefore, we can set

$$m = \widetilde{\Omega}(R^8 L^{22})$$

to ensure that $\epsilon_{\text{app}}(\tau) \leq 1/8$, where $\widetilde{\Omega}(\cdot)$ hides polylogarithmic dependencies on network depth L, NTRF function class size R, and failure probability parameter δ.

Then by Lemma 5.7.4, we have with probability at least $1-\delta$,

$$\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(n')} - \mathbf{W}^*\|_F^2 \geq \eta \sum_{i=1}^{n'} L_i(\mathbf{W}^{(i-1)}) - 2n\eta\epsilon_{\text{NTRF}} \tag{5.7.4}$$

as long as $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(n'-1)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$.

We then prove Theorem 5.3.5 in two steps: 1) all iterates stay inside $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$; and 2) convergence of online SGD.

All iterates stay inside $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Similar to the proof of Theorem 5.3.3, we prove this part by induction. Assuming $\mathbf{W}^{(i)}$ satisfies $\mathbf{W}^{(i)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ for all $i \leq n' - 1$, by (5.7.4), we have

$$\begin{aligned}
\|\mathbf{W}^{(n')} - \mathbf{W}^*\|_F^2 &\leq \|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 + 2n\eta\epsilon_{\text{NTRF}} \\
&\leq LR^2 \cdot m^{-1} + 2n\eta\epsilon_{\text{NTRF}},
\end{aligned}$$

where the last inequality is by $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, Rm^{-1/2})$. Then by triangle inequality, we further get

$$\begin{aligned}
\|\mathbf{W}_l^{(n')} - \mathbf{W}_l^{(0)}\|_F &\leq \|\mathbf{W}_l^{(n')} - \mathbf{W}_l^*\|_F + \|\mathbf{W}_l^* - \mathbf{W}_l^{(0)}\|_F \\
&\leq \|\mathbf{W}^{(n')} - \mathbf{W}^*\|_F + \|\mathbf{W}_l^* - \mathbf{W}_l^{(0)}\|_F \\
&\leq \mathcal{O}(\sqrt{L}Rm^{-1/2} + \sqrt{n\eta\epsilon_{\text{NTRF}}}).
\end{aligned}$$

Then by our choices of $\eta = \Theta\big(m^{-1} \cdot (LR^2 n^{-1}\epsilon_{\text{NTRF}}^{-1} \wedge L^{-1})\big)$, we have $\|\mathbf{W}^{(n')} - \mathbf{W}^{(0)}\|_F \leq 2\sqrt{L}Rm^{-1/2} \leq \tau$. This completes the proof of the first part.

Convergence of online SGD. By (5.7.4), we have

$$\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(n)} - \mathbf{W}^*\|_F^2 \geq \eta\bigg(\sum_{i=1}^n L_i(\mathbf{W}^{(i-1)}) - 2n\epsilon_{\mathrm{NTRF}}\bigg).$$

Dividing by ηn on the both sides and rearranging terms, we get

$$\frac{1}{n}\sum_{i=1}^n L_i(\mathbf{W}^{(i-1)}) \leq \frac{\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(n)} - \mathbf{W}^*\|_F^2}{\eta n} + 2\epsilon_{\mathrm{NTRF}} \leq \frac{L^2R^2}{n} + 3\epsilon_{\mathrm{NTRF}},$$

where the second inequality follows from facts that $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, R\cdot m^{-1/2})$ and $\eta = \Theta\big(m^{-1}\cdot (LR^2n^{-1}\epsilon_{\mathrm{NTRF}}^{-1} \wedge L^{-1})\big)$. By Lemma 4.3 in [JT20] and the fact that $\mathcal{E}_i(\mathbf{W}^{(i-1)}) \leq L_i(\mathbf{W}^{(i-1)})$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(i-1)}) &\leq \frac{2}{n}\sum_{i=1}^n \mathcal{E}_{\mathcal{D}}(\mathbf{W}^{(i-1)}) \\
&\leq \frac{8}{n}\sum_{i=1}^n \mathcal{E}_i(\mathbf{W}^{(i-1)}) + \frac{8\log(1/\delta)}{n} \\
&\leq \frac{8L^2R^2}{n} + \frac{8\log(1/\delta)}{n} + 24\epsilon_{\mathrm{NTRF}}.
\end{aligned}$$

This completes the proof of the second part. □

5.8 Proof of Results in Section 5.4

5.8.1 Proof of Proposition 5.4.2

We first provide the following lemma which gives an upper bound of the neural network output at the initialization.

Lemma 5.8.1 (Lemma 4.4 in [CG19])**.** Under Assumption 5.3.1, if $m \geq \bar{C}L\log(nL/\delta)$ with some absolute constant $\bar{C}$, with probability at least $1-\delta$, we have

$$|f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)| \leq C\sqrt{\log(n/\delta)}$$

for some absolute constant C.

Proof of Proposition 5.4.2. Under Assumption 5.4.1, we can find a collection of matrices $\mathbf{U}^* = \{\mathbf{U}_1^*, \cdots, \mathbf{U}_L^*\}$ with $\sum_{l=1}^L \|\mathbf{U}_l^*\|_F^2 = 1$ such that $y_i \langle \nabla f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \mathbf{U}^* \rangle \geq m^{1/2}\gamma$ for at least $1-\rho$ fraction of the training data. By Lemma 5.8.1, for all $i \in [n]$ we have $|f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)| \leq C\sqrt{\log(n/\delta)}$ for some absolute constant C. Then for any positive constant λ, we have for at least $1-\rho$ portion of the data,

$$y_i\big(f_{\mathbf{W}^{(0)}}(\mathbf{x}_i) + \langle \nabla f_{\mathbf{W}^{(0)}}, \lambda \mathbf{U}^* \rangle\big) \geq m^{1/2}\lambda\gamma - C\sqrt{\log(n/\delta)}.$$

For this fraction of data, we can set

$$\lambda = \frac{C'\big[\log^{1/2}(n/\delta) + \log(1/\epsilon)\big]}{m^{1/2}\gamma},$$

where C' is an absolute constant, and get

$$m^{1/2}\lambda\gamma - C\sqrt{\log(n/\delta)} \geq \log(1/\epsilon).$$

Now we let $\mathbf{W}^* = \mathbf{W}^{(0)} + \lambda \mathbf{U}^*$. By the choice of R in Proposition 5.4.2, we have $\mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, R \cdot m^{-1/2})$. The above inequality implies that for at least $1-\rho$ fraction of data, we have $\ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big) \leq \epsilon$. For the rest data, we have

$$y_i\big(f_{\mathbf{W}^{(0)}}(\mathbf{x}_i) + \langle \nabla f_{\mathbf{W}^{(0)}}, \lambda \mathbf{U}^* \rangle\big) \geq -C\sqrt{\log(n/\delta)} - \lambda \|\nabla f_{\mathbf{W}^{(0)}}\|_2^2 \geq -C_1 R$$

for some absolute positive constant C_1, where the last inequality follows from fact that $\|\nabla f_{\mathbf{W}^{(0)}}\|_2 = \widetilde{\mathcal{O}}(m^{1/2})$ (see Lemma 5.7.1 for detail). Then note that we use cross-entropy loss, it follows that for this fraction of training data, we have $\ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big) \leq C_2 R$ for some constant C_2. Combining the results of these two fractions of training data, we can conclude

$$\epsilon_{\mathrm{NTRF}} \leq n^{-1} \sum_{i=1}^n \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big) \leq (1-\rho)\epsilon + \rho \cdot \mathcal{O}(R)$$

This completes the proof.

□

5.8.2 Proof of Proposition 5.4.4

Proof of Proposition 5.4.4. We are going to prove that Assumption 5.4.3 implies the existence of a good function in the NTRF function class.

By Definition 5.3.2 and the definition of cross-entropy loss, our goal is to prove that there exists a collection of matrices $\overline{\mathbf{W}} = \{\overline{\mathbf{W}}_1, \overline{\mathbf{W}}_2\}$ satisfying $\max\{\|\overline{\mathbf{W}}_1 - \mathbf{W}_1^{(0)}\|_F, \|\overline{\mathbf{W}}_2 - \mathbf{W}_2^{(0)}\|_2\} \leq R \cdot m^{-1/2}$ such that

$$y_i \cdot \big[f_{\mathbf{W}^{(0)}}(\mathbf{x}_i) + \langle \nabla_{\mathbf{W}_1} f_{\mathbf{W}^{(0)}}, \overline{\mathbf{W}}_1 - \mathbf{W}_1^{(0)} \rangle + \langle \nabla_{\mathbf{W}_2} f_{\mathbf{W}^{(0)}}, \overline{\mathbf{W}}_2 - \mathbf{W}_2^{(0)} \rangle\big] \geq \log(2/\epsilon).$$

We first consider $\nabla_{\mathbf{W}_1} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)$, which has the form

$$\big(\nabla_{\mathbf{W}_1} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\big)_j = m^{1/2} \cdot w_{2,j}^{(0)} \cdot \sigma'(\langle \mathbf{w}_{1,j}^{(0)}, \mathbf{x}_i \rangle) \cdot \mathbf{x}_i.$$

Note that $w_{2,j}^{(0)}$ and $\mathbf{w}_{1,j}^{(0)}$ are independently generated from $\mathcal{N}(0, 1/m)$ and $\mathcal{N}(0, 2\mathbf{I}/m)$ respectively, thus we have $\mathbb{P}(|w_{2,j}^{(0)}| \geq 0.47m^{-1/2}) \geq 1/2$. By Hoeffeding's inequality, we know that with probability at least $1 - \exp(-m/8)$, there are at least $m/4$ nodes, whose union is denoted by $\mathcal{S}$, satisfying $|w_{2,j}^{(0)}| \geq 0.47m^{-1/2}$. Then we only focus on the nodes in the set $\mathcal{S}$. Note that $\mathbf{W}_1^{(0)}$ and $\mathbf{W}_2^{(0)}$ are independently generated. Then by Assumption 5.4.3 and Hoeffeding's inequality, there exists a function $\overline{\mathbf{u}}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ such that with probability at least $1 - \delta'$,

$$\frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} y_i \cdot \langle \overline{\mathbf{u}}(\mathbf{w}_{1,j}^{(0)}), \mathbf{x}_i \rangle \cdot \sigma'(\langle \mathbf{w}_{1,j}^{(0)}, \mathbf{x}_i \rangle) \geq \gamma - \sqrt{\frac{2\log(1/\delta')}{|\mathcal{S}|}}.$$

Define $\mathbf{v}_j = \overline{\mathbf{u}}(\mathbf{w}_{1,j}^{(0)})/w_{2,j}$ if $|w_{2,j}| \geq 0.47m^{-1/2}$ and $\mathbf{v}_j = \mathbf{0}$ otherwise. Then we have

$$\begin{aligned}\sum_{j=1}^{m} y_i \cdot w_{2,j}^{(0)} \cdot \langle \mathbf{v}_j, \mathbf{x}_i \rangle \cdot \sigma'(\langle \mathbf{w}_{1,j}^{(0)}, \mathbf{x}_i \rangle) &= \sum_{j \in \mathcal{S}} y_i \cdot \langle \overline{\mathbf{u}}(\mathbf{w}_{1,j}^{(0)}), \mathbf{x}_i \rangle \cdot \sigma'(\langle \mathbf{w}_{1,j}^{(0)}, \mathbf{x}_i \rangle) \\ &\geq |\mathcal{S}|\gamma - \sqrt{2|\mathcal{S}|\log(1/\delta')}.\end{aligned}$$

Set $\delta = 2n\delta'$ and apply union bound, we have with probability at least $1 - \delta/2$,

$$\sum_{j=1}^{m} y_i \cdot w_{2,j}^{(0)} \cdot \langle \mathbf{v}_j, \mathbf{x}_i \rangle \cdot \sigma'(\langle \mathbf{w}_{1,j}^{(0)}, \mathbf{x}_i \rangle) \geq |\mathcal{S}|\gamma - \sqrt{2|\mathcal{S}|\log(2n/\delta)}.$$

Therefore, note that with probability at least $1-\exp(-m/8)$, we have $|\mathcal{S}| \geq m/4$. Moreover, in Assumption 5.4.3, by $y_i \in \{\pm 1\}$ and $|\sigma'(\cdot)|, \|\overline{\mathbf{u}}(\cdot)\|_2, \|\mathbf{x}_i\|_2 \leq 1$ for $i = 1, \ldots, n$, we see that $\gamma \leq 1$. Then if $m \geq 32\log(n/\delta)/\gamma^2$, with probability at least $1-\delta/2-\exp\big(-4\log(n/\delta)/\gamma^2\big) \geq 1-\delta$,

$$\sum_{j=1}^{m} y_i \cdot w_{2,j}^{(0)} \cdot \langle \mathbf{v}_j, \mathbf{x}_i \rangle \cdot \sigma'(\langle \mathbf{w}_{1,j}^{(0)}, \mathbf{x}_i \rangle) \geq |\mathcal{S}|\gamma/2.$$

Let $\mathbf{U} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m)^\top / \sqrt{m|\mathcal{S}|}$, we have

$$y_i \langle \nabla_{\mathbf{W}_1} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \mathbf{U} \rangle = \frac{1}{\sqrt{|\mathcal{S}|}} \sum_{j=1}^{m} y_i \cdot w_{2,j}^{(0)} \cdot \langle \mathbf{v}_j, \mathbf{x}_i \rangle \cdot \sigma'(\langle \mathbf{w}_{1,j}^{(0)}, \mathbf{x}_i \rangle) \geq \frac{\sqrt{|\mathcal{S}|}\gamma}{2} \geq \frac{m^{1/2}\gamma}{4},$$

where the last inequality is by the fact that $|\mathcal{S}| \geq m/4$. Besides, note that by concentration and Gaussian tail bound, we have $|f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)| \leq C\log(n/\delta)$ for some absolute constant C. Therefore, let $\overline{\mathbf{W}}_1 = \mathbf{W}_1^{(0)} + 4\big(\log(2/\epsilon) + C\log(n/\delta)\big)m^{-1/2}\mathbf{U}/\gamma$ and $\overline{\mathbf{W}}_2 = \mathbf{W}_2^{(0)}$, we have

$$y_i \cdot \big[f_{\mathbf{W}^{(0)}}(\mathbf{x}_i) + \langle \nabla_{\mathbf{W}_1} f_{\mathbf{W}^{(0)}}, \overline{\mathbf{W}}_1 - \mathbf{W}_1^{(0)} \rangle + \langle \nabla_{\mathbf{W}_2} f_{\mathbf{W}^{(0)}}, \overline{\mathbf{W}}_2 - \mathbf{W}_2^{(0)} \rangle \big] \geq \log(2/\epsilon). \quad (5.8.1)$$

Note that $\|\overline{\mathbf{u}}(\cdot)\|_2 \leq 1$, we have $\|\mathbf{U}\|_F \leq 1/0.47 \leq 2.2$. Therefore, we further have $\|\overline{\mathbf{W}}_1 - \mathbf{W}_1^{(0)}\|_F \leq 8.8\gamma^{-1}\big(\log(2/\epsilon) + C\log(n/\delta)\big) \cdot m^{-1/2}$. This implies that $\overline{\mathbf{W}} \in \mathcal{B}(\mathbf{W}^{(0)}, R)$ with $R = \mathcal{O}\big(\log\big(n/(\delta\epsilon)\big)/\gamma\big)$. Applying the inequality $\ell(\log(2/\epsilon)) \leq \epsilon$ on (5.8.1) gives

$$\ell(y_i \cdot F_{\mathbf{W}^{(0)}, \overline{\mathbf{W}}}(\mathbf{x}_i)) \leq \epsilon$$

for all $i = 1, \ldots, n$. This completes the proof. □

5.8.3 Proof of Proposition 5.4.6

Based on our theoretical analysis, the major goal is to show that there exist certain choices of R and m such that the best NTRF model in the function class $\mathcal{F}(\mathbf{W}^{(0)}, R)$ can achieve ϵ training error. In this proof, we will prove a stronger results by showing that given the quantities of R and m specificed in Proposition 5.4.6, there exists a NTRF model with parameter $\mathbf{W}^*$ that satisfies $n^{-1}\sum_{i=1}^{n} \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big) \leq \epsilon$.

In order to do so, we consider training the NTRF model via a different surrogate loss function. Specifically, we consider squared hinge loss $\widetilde{\ell}(x) = \big(\max\{\lambda - x, 0\}\big)^2$, where λ denotes the target margin. In the later proof, we choose $\lambda = \log(1/\epsilon) + 1$ such that the condition $\widetilde{\ell}(x) \leq 1$ can guarantee that $x \geq \log(\epsilon)$. Moreover, we consider using gradient flow, i.e., gradient descent with infinitesimal step size, to train the NTRF model. Therefore, in the remaining part of the proof, we consider optimizing the NTRF parameter $\mathbf{W}$ with the loss function

$$\widetilde{L}_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\ell}\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}}(\mathbf{x}_i)\big).$$

Moreover, for simplicity, we only consider optimizing parameter in the last hidden layer (i.e., $\mathbf{W}_{L-1}$). Then the gradient flow can be formulated as

$$\frac{\mathrm{d}\mathbf{W}_{L-1}(t)}{\mathrm{d}t} = -\nabla_{\mathbf{W}_{L-1}} \widetilde{L}_S(\mathbf{W}(t)), \quad \frac{\mathrm{d}\mathbf{W}_l(t)}{\mathrm{d}t} = \mathbf{0} \quad \text{for any } l \neq L-1.$$

Note that the NTRF model is a linear model, thus by Definition 5.3.2, we have

$$\begin{aligned}
\nabla_{\mathbf{W}_{L-1}} \widetilde{L}_S(\mathbf{W}(t)) &= y_i \widetilde{\ell}'\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}(t)}(\mathbf{x}_i)\big) \cdot \nabla_{\mathbf{W}_{L-1}} F_{\mathbf{W}^{(0)},\mathbf{W}(t)}(\mathbf{x}_i) \\
&= y_i \widetilde{\ell}'\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}(t)}(\mathbf{x}_i)\big) \cdot \nabla_{\mathbf{W}_{L-1}^{(0)}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i). \qquad (5.8.2)
\end{aligned}$$

Then it is clear that $\nabla_{\mathbf{W}_{L-1}} \widetilde{L}_S(\mathbf{W}(t))$ has fixed direction throughout the optimization.

In order to prove the convergence of gradient flow and characterize the quantity of R, We first provide the following lemma which gives an upper bound of the NTRF model output at the initialization.

Then we provide the following lemma which characterizes a lower bound of the Frobenius norm of the partial gradient $\nabla_{\mathbf{W}_{L-1}} \widetilde{L}_S(\mathbf{W})$.

Lemma 5.8.2 (Lemma B.5 in [ZCZ19])**.** Under Assumptions 5.3.1 and 5.4.5, if $m = \widetilde{\Omega}(n^2\phi^{-1})$, then for all $t \geq 0$, with probability at least $1 - \exp\big(-O(m\phi/n)\big)$, there exist a positive constant C such that

$$\|\nabla_{\mathbf{W}_{L-1}} \widetilde{L}_S(\mathbf{W}(t))\|_F^2 \geq \frac{Cm\phi}{n^5}\bigg[\sum_{i=1}^{n} \widetilde{\ell}'\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}(t)}(\mathbf{x}_i)\big)\bigg]^2.$$

We slightly modified the original version of this lemma since we use different models (we consider NTRF model while [ZCZ19] considers neural network model). However, by (5.8.2), it is clear that the gradient $\nabla\widetilde{L}_S(\mathbf{W})$ can be regarded as a type of the gradient for neural network model at the initialization (i.e., $\nabla_{\mathbf{W}_{L-1}} L_S(\mathbf{W}^{(0)})$) is valid. Now we are ready to present the proof.

Proof of Proposition 5.4.6. Recall that we only consider training the last hidden weights, i.e., $\mathbf{W}_{L-1}$, via gradient flow with squared hinge loss, and our goal is to prove that gradient flow is able to find a NTRF model within the function class $\mathcal{F}(\mathbf{W}^{(0)}, R)$ around the initialization, i.e., achieving $n^{-1}\sum_{i=1}^n \ell\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}^*}(\mathbf{x}_i)\big) \leq \epsilon$. Let $\mathbf{W}(t)$ be the weights at time t, gradient flow implies that

$$\frac{\mathrm{d}\widetilde{L}_S(\mathbf{W}(t))}{\mathrm{d}t} = -\|\nabla_{\mathbf{W}_{L-1}}\widetilde{L}_S(\mathbf{W}(t))\|_F^2 \leq -\frac{Cm\phi}{n^5}\bigg(\sum_{i=1}^n \widetilde{\ell}'\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}(t)}(\mathbf{x}_i)\big)\bigg)^2 = \frac{4Cm\phi\widetilde{L}_S(\mathbf{W}(t))}{n^3},$$

where the first equality is due to the fact that we only train the last hidden layer, the first inequality is by Lemma 5.8.2 and the second equality follows from the fact that $\widetilde{\ell}'(\cdot) = -2\sqrt{\widetilde{\ell}(\cdot)}$. Solving the above inequality gives

$$\widetilde{L}_S(\mathbf{W}(t)) \leq \widetilde{L}_S(\mathbf{W}(0)) \cdot \exp\bigg(-\frac{4Cm\phi t}{n^3}\bigg). \tag{5.8.3}$$

Then, set $T = \mathcal{O}\big(n^3 m^{-1}\phi^{-1} \cdot \log(\widetilde{L}_S(\mathbf{W}(0))/\epsilon')\big)$ and $\epsilon' = 1/n$, we have $\widetilde{L}_S(\mathbf{W}(t)) \leq \epsilon'$. Then it follows that $\widetilde{\ell}\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}(t)}(\mathbf{x}_i)\big) \leq 1$, which implies that $y_i F_{\mathbf{W}^{(0)},\mathbf{W}(t)}(\mathbf{x}_i) \geq \log(\epsilon)$ and thus $n^{-1}\sum_{i=1}^n \ell\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}^*}(\mathbf{x}_i)\big) \leq \epsilon$. Therefore, $\mathbf{W}(T)$ is exactly the NTRF model we are looking for.

The next step is to characterize the distance between $\mathbf{W}(T)$ and $\mathbf{W}(0)$ in order to characterize the quantity of R. Note that $\|\nabla_{\mathbf{W}_{L-1}}\widetilde{L}_S(\mathbf{W}(t))\|_F^2 \geq 4Cm\phi\widetilde{L}_S(\mathbf{W}(t))/n^3$, we have

$$\frac{\mathrm{d}\sqrt{\widetilde{L}_S(\mathbf{W}(t))}}{\mathrm{d}t} = -\frac{\|\nabla_{\mathbf{W}_{L-1}}\widetilde{L}_S(\mathbf{W}(t))\|_F^2}{2\sqrt{\widetilde{L}_S(\mathbf{W}(t))}} \leq -\|\nabla_{\mathbf{W}_{L-1}}\widetilde{L}_S(\mathbf{W}(t))\|_F \cdot \frac{C^{1/2}m^{1/2}\phi^{1/2}}{n^{3/2}}.$$

Taking integral on both sides and rearranging terms, we have

$$\int_{t=0}^{T} \|\nabla_{\mathbf{W}_{L-1}} \widetilde{L}_S(\mathbf{W}(t))\|_F \mathrm{d}t \leq \frac{n^{3/2}}{C^{1/2} m^{1/2} \phi^{1/2}} \cdot \left(\sqrt{\widetilde{L}_S(\mathbf{W}(0))} - \sqrt{\widetilde{L}_S(\mathbf{W}(t))} \right).$$

Note that the L.H.S. of the above inequality is an upper bound of $\|\mathbf{W}(t) - \mathbf{W}(0)\|_F$, we have for any $t \geq 0$,

$$\|\mathbf{W}(t) - \mathbf{W}(0)\|_F \leq \frac{n^{3/2}}{C^{1/2} m^{1/2} \phi^{1/2}} \cdot \sqrt{\widetilde{L}_S(\mathbf{W}(0))} = \mathcal{O}\bigg(\frac{n^{3/2} \log\big(n/(\delta\epsilon)\big)}{m^{1/2} \phi^{1/2}} \bigg),$$

where the second inequality is by Lemma 5.8.1 and our choice of $\lambda = \log(1/\epsilon) + 1$. This implies that there exists a point $\mathbf{W}^*$ within the class $\mathcal{F}(\mathbf{W}^{(0)}, R)$ with

$$R = \mathcal{O}\bigg(\frac{n^{3/2} \log\big(n/(\delta\epsilon)\big)}{\phi^{1/2}} \bigg)$$

such that

$$\epsilon_{\mathrm{NTRF}} := n^{-1} \sum_{i=1}^{n} \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big) \leq \epsilon.$$

Then by Theorem 5.3.3, and, more specifically, (5.7.1), we can compute the minimal required neural network width as follows,

$$m = \widetilde{\Omega}(R^8 L^{22}) = \widetilde{\Omega}\bigg(\frac{L^{22} n^{12}}{\phi^4} \bigg).$$

This completes the proof. □

5.9 Proof of Technical Lemmas

Here we provide the proof of Lemmas 5.6.1, 5.7.3 and 5.7.4.

5.9.1 Proof of Lemma 5.6.1

The detailed proof of Lemma 5.6.1 is given as follows.

Proof of Lemma 5.6.1. Based on the update rule of gradient descent, i.e., $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \nabla_{\mathbf{W}} L_S(\mathbf{W}^{(t)})$, we have the following calculation.

$$\begin{aligned}
&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \\
&\quad = \underbrace{\frac{2\eta}{n} \sum_{i=1}^{n} \langle \mathbf{W}^{(t)} - \mathbf{W}^*, \nabla_{\mathbf{W}} L_i(\mathbf{W}^{(t)}) \rangle}_{I_1} - \underbrace{\eta^2 \sum_{l=1}^{L} \|\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(t)})\|_F^2}_{I_2}, \qquad (5.9.1)
\end{aligned}$$

where the equation follows from the fact that $L_S(\mathbf{W}^{(t)}) = n^{-1} \sum_{i=1}^{n} L_i(\mathbf{W}^{(t)})$. In what follows, we first bound the term I_1 on the R.H.S. of (5.9.1) by approximating the neural network functions with linear models. By assumption, for $t = 0, \ldots, t' - 1$, $\mathbf{W}^{(t)}, \mathbf{W}^* \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Therefore by the definition of $\epsilon_{\text{app}}(\tau)$,

$$y_i \cdot \langle \nabla f_{\mathbf{W}^{(t)}}(\mathbf{x}_i), \mathbf{W}^{(t)} - \mathbf{W}^* \rangle \leq y_i \cdot \big(f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) - f_{\mathbf{W}^*}(\mathbf{x}_i) \big) + \epsilon_{\text{app}}(\tau) \qquad (5.9.2)$$

Moreover, we also have

$$\begin{aligned}
0 &\leq y_i \cdot \big(f_{\mathbf{W}^*}(\mathbf{x}_i) - f_{\mathbf{W}^{(0)}}(\mathbf{x}_i) - \langle \nabla f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \mathbf{W}^* - \mathbf{W}^{(0)} \rangle \big) + \epsilon_{\text{app}}(\tau) \\
&= y_i \cdot \big(f_{\mathbf{W}^*}(\mathbf{x}_i) - F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i) \big) + \epsilon_{\text{app}}(\tau), \qquad (5.9.3)
\end{aligned}$$

where the equation follows by the definition of $F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x})$. Adding (5.9.3) to (5.9.2) and canceling the terms $y_i \cdot f_{\mathbf{W}^*}(\mathbf{x}_i)$, we obtain that

$$y_i \cdot \langle \nabla f_{\mathbf{W}^{(t)}}(\mathbf{x}_i), \mathbf{W}^{(t)} - \mathbf{W}^* \rangle \leq y_i \cdot \big(f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) - F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i) \big) + 2\epsilon_{\text{app}}(\tau). \qquad (5.9.4)$$

We can now give a lower bound on first term on the R.H.S. of (5.9.1). For $i = 1, \ldots, n$, applying the chain rule on the loss function gradients and utilizing (5.9.4), we have

$$\begin{aligned}
\langle \mathbf{W}^{(t)} - \mathbf{W}^*, \nabla_{\mathbf{W}} L_i(\mathbf{W}^{(t)}) \rangle &= \ell'\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \big) \cdot y_i \cdot \langle \mathbf{W}^{(t)} - \mathbf{W}^*, \nabla_{\mathbf{W}} f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \rangle \\
&\geq \ell'\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \big) \cdot \big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) - y_i f_{\mathbf{W}^*}(\mathbf{x}_i) + 2\epsilon_{\text{app}}(\tau) \big) \\
&\geq (1 - 2\epsilon_{\text{app}}(\tau)) \ell\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \big) - \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i) \big), \qquad (5.9.5)
\end{aligned}$$

where the first inequality is by the fact that $\ell'\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \big) < 0$, the second inequality is by convexity of $\ell(\cdot)$ and the fact that $-\ell'\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \big) \leq \ell\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \big)$.

We now proceed to bound the term I_2 on the R.H.S. of (5.9.1). Note that we have $\ell'(\cdot) < 0$, and therefore the Frobenius norm of the gradient $\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(t)})$ can be upper bounded as follows,

$$\begin{aligned}\|\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(t)})\|_F &= \bigg\| \frac{1}{n} \sum_{i=1}^{n} \ell'\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\big) \nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(t)}}(\mathbf{x}_i) \bigg\|_F \\ &\leq \frac{1}{n} \sum_{i=1}^{n} -\ell'\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\big) \cdot \|\nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\|_F,\end{aligned}$$

where the inequality follows by triangle inequality. We now utilize the fact that cross-entropy loss satisfies the inequalities $-\ell'(\cdot) \leq \ell(\cdot)$ and $-\ell'(\cdot) \leq 1$. Therefore by definition of $M(\tau)$, we have

$$\begin{aligned}\sum_{l=1}^{L} \|\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(t)})\|_F^2 &\leq \mathcal{O}\big(LM(\tau)^2\big) \cdot \bigg(\frac{1}{n} \sum_{i=1}^{n} -\ell'\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\big) \bigg)^2 \\ &\leq \mathcal{O}\big(LM(\tau)^2\big) \cdot L_S(\mathbf{W}^{(t)}). \end{aligned} \tag{5.9.6}$$

Then we can plug (5.9.5) and (5.9.6) into (5.9.1) and obtain

$$\begin{aligned}&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \\ &\quad\geq \frac{2\eta}{n} \sum_{i=1}^{n} \Big[(1 - 2\epsilon_{\text{app}}(\tau)) \ell\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\big) - \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big) \Big] - \mathcal{O}\big(\eta^2 LM(\tau)^2\big) \cdot L_S(\mathbf{W}^{(t)}) \\ &\quad\geq \bigg[\frac{3}{2} - 4\epsilon_{\text{app}}(\tau) \bigg] \eta L_S(\mathbf{W}^{(t)}) - \frac{2\eta}{n} \sum_{i=1}^{n} \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big),\end{aligned}$$

where the last inequality is by $\eta = \mathcal{O}(L^{-1} M(\tau)^{-2})$ and merging the third term on the second line into the first term. Taking telescope sum from $t = 0$ to $t = t' - 1$ and plugging in the definition $\frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big) = \epsilon_{\text{NTRF}}$ completes the proof. □

5.9.2 Proof of Lemma 5.7.3

Proof of Lemma 5.7.3. We first denote $\mathcal{W} = \mathcal{B}(\mathbf{W}^{(0)}, \widetilde{R} \cdot m^{-1/2})$, and define the corresponding neural network function class and surrogate loss function class as $\mathcal{F} = \{f_{\mathbf{W}}(\mathbf{x}) : \mathbf{W} \in \mathcal{W}\}$ and $\mathcal{G} = \{-\ell[y \cdot f_{\mathbf{W}}(\mathbf{x})] : \mathbf{W} \in \mathcal{W}\}$ respectively.

By standard uniform convergence results in terms of empirical Rademacher complexity [BM02; MRT18; SB14], with probability at least $1-\delta$ we have

$$\begin{aligned}\sup_{\mathbf{W}\in\mathcal{W}}|\mathcal{E}_S(\mathbf{W})-\mathcal{E}_{\mathcal{D}}(\mathbf{W})| &= \sup_{\mathbf{W}\in\mathcal{W}}\left|-\frac{1}{n}\sum_{i=1}^{n}\ell'\big[y_i\cdot f_{\mathbf{W}}(\boldsymbol{x}_i)\big]+\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\ell'\big[y\cdot f_{\mathbf{W}}(\mathbf{x})\big]\right| \\ &\le 2\widehat{\mathfrak{R}}_n(\mathcal{G})+C_1\sqrt{\frac{\log(1/\delta)}{n}},\end{aligned}$$

where C_1 is an absolute constant, and

$$\widehat{\mathfrak{R}}_n(\mathcal{G})=\mathbb{E}_{\xi_i\sim\mathrm{Unif}(\{\pm1\})}\left\{\sup_{\mathbf{W}\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}\xi_i\ell'\big[y_i\cdot f_{\mathbf{W}}(\boldsymbol{x}_i)\big]\right\}$$

is the empirical Rademacher complexity of the function class $\mathcal{G}$. We now provide two bounds on $\widehat{\mathfrak{R}}_n(\mathcal{G})$, whose combination gives the final result of Lemma 5.7.3. First, by Corollary 5.35 in [Ver10], with probability at least $1-L\cdot\exp(-\Omega(m))$, $\|\mathbf{W}_l^{(0)}\|_2\le 3$ for all $l\in[L]$. Therefore for all $\mathbf{W}\in\mathcal{W}$, we have $\|\mathbf{W}_l\|_2\le 4$. Moreover, standard concentration inequalities on the norm of the first row of $\mathbf{W}_l^{(0)}$ also implies that $\|\mathbf{W}_l\|_2\ge 0.5$ for all $\mathbf{W}\in\mathcal{W}$ and $l\in[L]$. Therefore, an adaptation of the bound in [BFT17][5] gives

$$\begin{aligned}\widehat{\mathfrak{R}}_n(\mathcal{F}) &\le \widetilde{\mathcal{O}}\left(\sup_{\mathbf{W}\in\mathcal{W}}\left\{\frac{m^{1/2}}{\sqrt{n}}\cdot\left[\prod_{l=1}^{L}\|\mathbf{W}_l\|_2\right]\cdot\left[\sum_{l=1}^{L}\frac{\|\mathbf{W}_l^\top-\mathbf{W}_l^{(0)\top}\|_{2,1}^{2/3}}{\|\mathbf{W}_l\|_2^{2/3}}\right]^{3/2}\right\}\right) \\ &\le \widetilde{\mathcal{O}}\left(\sup_{\mathbf{W}\in\mathcal{W}}\left\{\frac{4^L m^{1/2}}{\sqrt{n}}\cdot\left[\sum_{l=1}^{L}(\sqrt{m}\cdot\|\mathbf{W}_l^\top-\mathbf{W}_l^{(0)\top}\|_F)^{2/3}\right]^{3/2}\right\}\right) \\ &\le \widetilde{\mathcal{O}}\left(4^L L^{3/2}\widetilde{R}\cdot\sqrt{\frac{m}{n}}\right). \qquad (5.9.7)\end{aligned}$$

We now derive the second bound on $\widehat{\mathfrak{R}}_n(\mathcal{G})$, which is inspired by the proof provided in [CG20]. Since $y\in\{+1,1\}$, $|\ell'(z)|\le 1$ and $\ell'(z)$ is 1-Lipschitz continuous, by standard

[5] [BFT17] only proved the Rademacher complexity bound for the composition of the ramp loss and the neural network function. In our setting essentially the ramp loss is replaced with the $-\ell'(\cdot)$ function, which is bounded and 1-Lipschitz continuous. The proof in our setting is therefore exactly the same as the proof given in [BFT17], and we can apply Theorem 3.3 and Lemma A.5 in [BFT17] to obtain the desired bound we present here.

empirical Rademacher complexity bounds [BM02; MRT18; SB14], we have

$$\widehat{\mathfrak{R}}_n(\mathcal{G}) \le \widehat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_{\xi_i \sim \mathrm{Unif}(\{\pm 1\})}\left[\sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^n \xi_i f_{\mathbf{W}}(\boldsymbol{x}_i)\right],$$

where $\widehat{\mathfrak{R}}_n(\mathcal{F})$ is the empirical Rademacher complexity of the function class $\mathcal{F}$. We have

$$\widehat{\mathfrak{R}}_n[\mathcal{F}] \le \underbrace{\mathbb{E}_{\boldsymbol{\xi}}\left\{\sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^n \xi_i\big[f_{\mathbf{W}}(\boldsymbol{x}_i) - F_{\mathbf{W}^{(0)},\mathbf{W}}(\boldsymbol{x}_i)\big]\right\}}_{I_1} + \underbrace{\mathbb{E}_{\boldsymbol{\xi}}\left\{\sup_{\mathbf{W}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^n \xi_i F_{\mathbf{W}^{(0)},\mathbf{W}}(\boldsymbol{x}_i)\right\}}_{I_2}, \tag{5.9.8}$$

where $F_{\mathbf{W}^{(0)},\mathbf{W}}(\mathbf{x}) = f_{\mathbf{W}^{(0)}}(\mathbf{x}) + \big\langle \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}), \mathbf{W} - \mathbf{W}^{(0)}\big\rangle$. For I_1, by Lemma 4.1 in [CG19], with probability at least $1-\delta/2$ we have

$$I_1 \le \max_{i\in[n]} \big|f_{\mathbf{W}}(\boldsymbol{x}_i) - F_{\mathbf{W}^{(0)},\mathbf{W}}(\boldsymbol{x}_i)\big| \le \mathcal{O}\big(L^3 \widetilde{R}^{4/3} m^{-1/6}\sqrt{\log(m)}\big),$$

For I_2, note that $\mathbb{E}_{\boldsymbol{\xi}}\big[\sup_{\mathbf{W}\in\mathcal{W}} \sum_{i=1}^n \xi_i f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\big] = 0$. By Cauchy-Schwarz inequality we have

$$\begin{aligned} I_2 &= \frac{1}{n}\sum_{l=1}^L \mathbb{E}_{\boldsymbol{\xi}}\left\{\sup_{\|\widetilde{\mathbf{W}}_l\|_F \le \widetilde{R} m^{-1/2}} \mathrm{Tr}\left[\widetilde{\mathbf{W}}_l^\top \sum_{i=1}^n \xi_i \nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\right]\right\} \\ &\le \frac{\widetilde{R} m^{-1/2}}{n}\sum_{l=1}^L \mathbb{E}_{\boldsymbol{\xi}}\left[\left\|\sum_{i=1}^n \xi_i \nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\right\|_F\right]. \end{aligned}$$

Therefore

$$\begin{aligned} I_2 &\le \frac{\widetilde{R} m^{-1/2}}{n}\sum_{l=1}^L \sqrt{\mathbb{E}_{\boldsymbol{\xi}}\left[\left\|\sum_{i=1}^n \xi_i \nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\right\|_F^2\right]} \\ &= \frac{\widetilde{R} m^{-1/2}}{n}\sum_{l=1}^L \sqrt{\sum_{i=1}^n \big\|\nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\big\|_F^2} \\ &\le \mathcal{O}\bigg(\frac{L\cdot\widetilde{R}}{\sqrt{n}}\bigg), \end{aligned}$$

where we apply Jensen's inequality to obtain the first inequality, and the last inequality follows by Lemma B.3 in [CG19]. Combining the bounds of I_1 and I_2 gives

$$\widehat{\mathfrak{R}}_n[\mathcal{F}] \le \widetilde{\mathcal{O}}\bigg(\frac{L\widetilde{R}}{\sqrt{n}} + \frac{L^3\widetilde{R}^{4/3}}{m^{1/6}}\bigg).$$

Further combining this bound with (5.9.7) and recaling δ completes the proof. □

5.9.3 Proof of Lemma 5.7.4

Proof of Lemma 5.7.4. Different from the proof of Lemma 5.6.1, online SGD only queries one data to update the model parameters in each iteration, i.e., $\mathbf{W}^{i+1} = \mathbf{W}^i - \eta \nabla L_{i+1}(\mathbf{W}^{(i)})$. By this update rule, we have

$$\begin{aligned}
&\|\mathbf{W}^{(i)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(i+1)} - \mathbf{W}^*\|_F^2 \\
&\quad = 2\eta \langle \mathbf{W}^{(i)} - \mathbf{W}^*, \nabla_{\mathbf{W}} L_{i+1}(\mathbf{W}^{(i)}) \rangle - \eta^2 \sum_{l=1}^{L} \|\nabla_{\mathbf{W}_l} L_{i+1}(\mathbf{W}^{(i)})\|_F^2. \qquad (5.9.9)
\end{aligned}$$

With exactly the same proof as (5.9.5) in the proof of Lemma 5.6.1, we have

$$\langle \mathbf{W}^{(t)} - \mathbf{W}^*, \nabla_{\mathbf{W}} L_i(\mathbf{W}^{(t)}) \rangle \geq (1 - 2\epsilon_{\text{app}}(\tau)) \ell\big(y_i f_{\mathbf{W}^{(t)}}(\mathbf{x}_i)\big) - \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big), \qquad (5.9.10)$$

for all $i = 0, \ldots, n' - 1$. By the fact that $-\ell'(\cdot) \leq \ell(\cdot)$ and $-\ell'(\cdot) \leq 1$, we have

$$\begin{aligned}
\sum_{l=1}^{L} \|\nabla_{\mathbf{W}_l} L_{i+1}(\mathbf{W}^{(i)})\|_F^2 &\leq \sum_{l=1}^{L} \ell\big(y_{i+1} f_{\mathbf{W}_t}(\mathbf{x}_{i+1})\big) \cdot \|\nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(i)}}(\mathbf{x}_{i+1})\|_F^2 \\
&\leq \mathcal{O}\big(LM(\tau)^2\big) \cdot L_{i+1}(\mathbf{W}^{(i)}). \qquad (5.9.11)
\end{aligned}$$

Then plugging (5.9.10) and (5.9.11) into (5.9.9) gives

$$\begin{aligned}
&\|\mathbf{W}^{(i)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(i+1)} - \mathbf{W}^*\|_F^2 \\
&\quad \geq (2 - 4\epsilon_{\text{app}}(\tau)) \eta L_{i+1}(\mathbf{W}^{(i)}) - 2\eta \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big) - \mathcal{O}\big(\eta^2 LM(\tau)^2\big) L_{i+1}(\mathbf{W}^{(i)}) \\
&\quad \geq (\frac{3}{2} - 4\epsilon_{\text{app}}(\tau)) \eta L_{i+1}(\mathbf{W}^{(i)}) - 2\eta \ell\big(y_i F_{\mathbf{W}^{(0)}, \mathbf{W}^*}(\mathbf{x}_i)\big),
\end{aligned}$$

where the last inequality is by $\eta = \mathcal{O}(L^{-1}M(\tau)^{-2})$ and merging the third term on the second line into the first term. Taking telescope sum over $i = 0, \ldots, n' - 1$, we obtain

$$\begin{aligned}
&\|\mathbf{W}^{(0)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(n')} - \mathbf{W}^*\|_F^2 \\
&\quad \geq \Big(\frac{3}{2} - 4\epsilon_{\text{app}}(\tau)\Big)\eta \sum_{i=1}^{n'} L_i(\mathbf{W}^{(i-1)}) - 2\eta \sum_{i=1}^{n'} \ell\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}^*}(\mathbf{x}_i)\big). \\
&\quad \geq \Big(\frac{3}{2} - 4\epsilon_{\text{app}}(\tau)\Big)\eta \sum_{i=1}^{n'} L_i(\mathbf{W}^{(i-1)}) - 2\eta \sum_{i=1}^{n} \ell\big(y_i F_{\mathbf{W}^{(0)},\mathbf{W}^*}(\mathbf{x}_i)\big). \\
&\quad \geq \Big(\frac{3}{2} - 4\epsilon_{\text{app}}(\tau)\Big)\eta \sum_{i=1}^{n'} L_i(\mathbf{W}^{(i-1)}) - 2n\eta\epsilon_{\text{NTRF}}.
\end{aligned}$$

This finishes the proof. □

5.10 Conclusions

In this work, we established the global convergence and generalization error bounds of GD and SGD for training deep ReLU networks for the binary classification problem. We show that a network width condition that is polylogarithmic in the sample size n and the inverse of target error ϵ^{-1} is sufficient to guarantee the learning of deep ReLU networks. Our results resolve an open question raised in [JT20].

CHAPTER 6

Generalization of Adam and SGD in Learning Neural Networks with Regularization

6.1 Introduction

Adaptive gradient methods [DHS11; HSS12; KB15; RKK18] such as Adam are very popular optimizers for training deep neural networks. By adjusting the learning rate coordinate-wisely based on historical gradient information, they are known to be able to automatically choose appropriate learning rates to achieve fast convergence in training. Because of this advantage, Adam and its variants are widely used in deep learning. Despite their fast convergence, adaptive gradient methods have been observed to achieve worse generalization performance compared with gradient descent and stochastic gradient descent (SGD) [WRS17; LXL18; CZT20; ZFM20] in many deep learning tasks such as image classification (we have done some simple deep learning experiments to justify this, the results are reported in Table 6.1). Even with explicit weight decay regularization, achieving good test error with

Models	AlexNet	VGG-16	ResNet-18
SGD	75.22	93.25	94.62
Adam	73.08	92.19	92.93

Table 6.1: Test accuracy (%) comparison between Adam and SGD on the CIFAR-10 dataset.

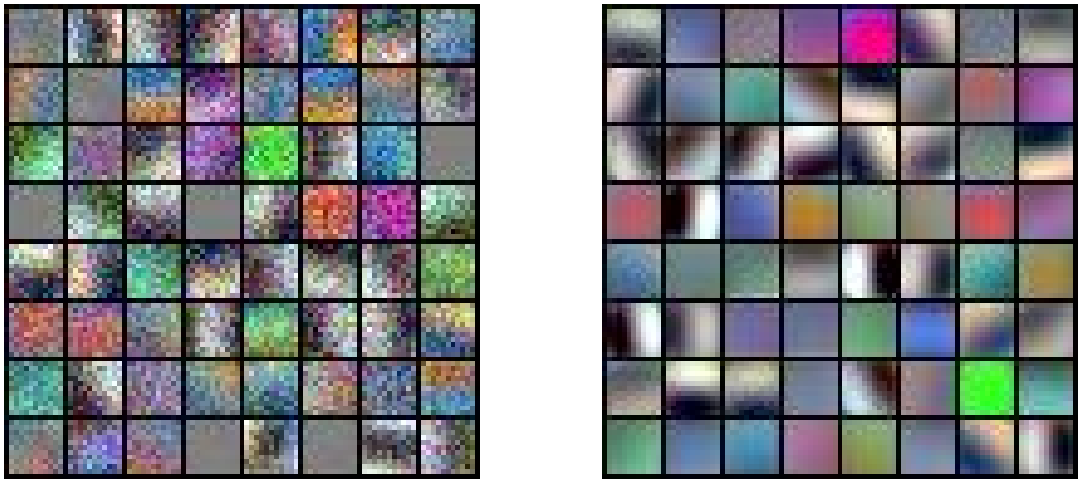
adaptive gradient methods seems to be challenging.

In this paper, we aim to provide a theoretical explanation towards the generalization gap between GD and Adam in image classification task. Specifically, we study Adam and GD for training neural networks with weight decay regularization on an image-like data model, and demonstrate the different behaviors of Adam and GD based on the notion of feature learning/noise memorization decomposition. We consider a model where the data are generated as a combination of feature and noise patches under certain sparsity conditions, and analyze the convergence and generalization of Adam and GD for training a two-layer convolutional neural network (CNN). The contributions of this paper are summarized as follows.

- We establish global convergence guarantees for Adam and GD with weight decay regularization. We show that, starting at the same random initialization, Adam and GD can both train a two-layer convolutional neural network to achieve zero training error after polynomially many iterations, despite the nonconvex optimization landscape.

- We further show that GD and Adam in fact converge to different global solutions with different generalization performance: when performed on the considered image-like data model, GD can achieve nearly zero test error, while the generalization performance of the model found by Adam is no better than a random guess. In particular, we show that the reason for this gap is due to the different training behaviors of Adam and GD: Adam is more likely to fit dense noises and output a model that is largely contributed by the noise patches; GD prefers to fit training data using their feature patch and finds a solution that is mainly composed by the true features. We also illustrate such different training processes in Figure 6.1, where it can be seen that the model trained by Adam is clearly more "noisy" than that trained by SGD.

- We also show that for convex settings with weight decay regularization, both Adam and gradient descent converge to the same solution and therefore have no test error



(a) Adam (b) SGD

Figure 6.1: Visualization of the first layer of AlexNet trained by Adam and SGD on the CIFAR-10 dataset. Both algorithms are run for 100 epochs with weight decay regularization and standard data augmentations, but without batch normalization. Clearly, the model learned by Adam is more "noisy" than that learned by SGD, implying that Adam is more likely to overfit the noise in the training data.

difference. This suggests that the difference between Adam and GD cannot be fully explained by linear models or neural networks trained in the "almost convex" neural tangent kernel (NTK) regime [JGH18; ALS19a; DLL19; ZCZ19; ADH19b; CG19; JT20; CCZ21]. It also demonstrates that the inferior generalization performance of Adam is closely tied to the nonconvex landscape of deep learning optimization, and cannot be solved by adding regularization.

6.2 Problem Setup and Preliminaries

We consider learning a CNN with Adam and GD based on n independent training examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ generated from a data model $\mathcal{D}$. In the following. we first introduce our data model $\mathcal{D}$, and then explain our neural network model and the details of the training

algorithms.

Data model. We consider a data model where the data inputs consist of feature and noise patches. Such a data model is motivated by image classification problems where the label of an image usually only depends on part of an image, and the other parts of the image showing random objects, or features that belong to other classes, can be considered as noises. When using CNN to fit the data, the convolution operation is applied to each patch of the data input separately. We claim that our data model is more practical than those considered in [WRS17; RKK18], which are handcrafted for showing the failure of Adam in term of either convergence or generalization. For simplicity, we only consider the case where the data consists of one feature patch and one noise patch. However, our result can be easily extended to cover the setting where there are multiple feature/noise patches. The detailed definition of our data model is given in Definition 6.2.1 as follows.

Definition 6.2.1. Each data $(\mathbf{x}, y)$ with $\mathbf{x} \in \mathbb{R}^{2d}$ and $y \in \{-1, 1\}$ is generated as follows,

$$\mathbf{x} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top]^\top,$$

where one of $\mathbf{x}_1$ and $\mathbf{x}_2$ denotes the feature patch that consists of a feature vector $y \cdot \mathbf{v}$, which is assumed to be 1-sparse, and the other one denotes the noise patch and consists of a noise vector $\boldsymbol{\xi}$. Without loss of generality, we assume $\mathbf{v} = [1, 0, \ldots, 0]^\top$. The noise vector $\boldsymbol{\xi}$ is generated according to the following process:

- Randomly select s coordinates from $[d]\backslash\{1\}$ with equal probabilities, which is denoted as a vector $\mathbf{s} \in \{0, 1\}^d$.
- Generate $\boldsymbol{\xi}$ from distribution $\mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{I})$, and then mask off the first coordinate and other $d - s - 1$ coordinates, i.e., $\boldsymbol{\xi} = \boldsymbol{\xi} \odot \mathbf{s}$.
- Add feature noise to $\boldsymbol{\xi}$, i.e., $\boldsymbol{\xi} = \boldsymbol{\xi} - \alpha y \mathbf{v}$, where $0 < \alpha < 1$ is the strength of the feature noise.

In particular, throughout this paper we set $d = \Omega(n^4)$, $s = \Theta\big(\frac{d^{1/2}}{n^2}\big)$, $\sigma_p^2 = \Theta\big(\frac{1}{s\cdot\mathrm{polylog}(n)}\big)$ and $\alpha = \Theta\big(\sigma_p \cdot \mathrm{polylog}(n)\big)$.

The most natural way to think of our data model is to treat $\mathbf{x}$ as the output of some intermediate layer of a CNN. In literature, [PRE17] pointed out that the outputs of an intermediate layer of a CNN are usually sparse. [Yan19] also discussed the setting where the hidden nodes in such an intermediate layer are sampled independently. This motivates us to study sparse features and entry-wisely independent noises in our model. In this paper, we focus on the case where the feature vector $\mathbf{v}$ is 1-sparse and the noise vector is s-sparse for simplicity. However, these sparsity assumptions can be generalized to the settings where the feature and the noises are denser, as long as the sparsity gap between feature and noises exists.

Note that in Definition 6.2.1, each data input consists of two patches: a feature patch $y\mathbf{v}$ that is positively correlated with the label, and a noise patch $\boldsymbol{\xi}$ which contains the "feature noise" $-\alpha y\mathbf{v}$ as well as random Gaussian noises. Importantly, the feature noise $-\alpha y\mathbf{v}$ in the noise patch plays a pivotal role in both the training and test processes, which connects the noise overfitting in the training process and the inferior generalization ability in the test process.

Moreover, we would like to clarify that the data distribution considered in our paper is an **extreme case** where we assume there is only one feature vector and all data has a feature noise, since we believe this is the simplest model that captures the fundamental difference between Adam and SGD. With this data model, we aim to show why Adam and SGD perform differently. Our theoretical results and analysis techniques can also be extended to more practical settings where there are multiple feature vectors and multiple patches, each data can either contain a single feature or multiple features, together with pure random noise or feature noise.

Two-layer CNN model. We consider a two-layer CNN model F using truncated polyno-

mial activation function $\sigma(z) = (\max\{0, z\})^q$ and fix the weights of second layer to be all 1's, where $q \geq 3$. Mathematically, given the data $(\mathbf{x}, y)$, the j-th output of the CNN can be formulated as

$$F_j(\mathbf{W}, \mathbf{x}) = \sum_{r=1}^{m} \big[\sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_1 \rangle) + \sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_2 \rangle)\big] = \sum_{r=1}^{m} \big[\sigma(\langle \mathbf{w}_{j,r}, y \cdot \mathbf{v} \rangle) + \sigma(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi} \rangle)\big], \tag{6.2.1}$$

where m is the width of the network, $\mathbf{w}_{j,r} \in \mathbb{R}^d$ denotes the weight at the r-th neuron, and $\mathbf{W}$ is the collection of model weights. We remark that we set the output layer as all 1's for the ease of analysis, our analyses and results can still be applied if using an random weights for different neurons, i.e., $F_j(\mathbf{W}, \mathbf{x}) = \sum_{r=1}^{m} v_{j,r} \big[\sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_1 \rangle) + \sigma(\langle \mathbf{w}_{j,r}, \mathbf{x}_2 \rangle)\big]$, where $v_{j,r}$ are randomly generated with a constant scaling.

Besides, the motivation of using polynomial ReLU activation function is to guarantee that the loss function is (locally) smooth and the amplification ability of pattern learning. It can be replaced by a smoothed ReLU activation function (e.g., the activation function used in [AL20]). If we assume the input data distribution is Gaussian, we can also deal with ReLU activation function [LMZ20]. Moreover, we would like to emphasize that $\mathbf{x}_1$ and $\mathbf{x}_2$ denote two data patches, which are randomly assigned with feature vector or noise vector independently for each data point. The leaner has no knowledge about which one is the feature patch (or noise patch).

In this paper we assume the width of the network is polylogarithmic in the training sample size, i.e., $m = \mathrm{polylog}(n)$. We assume $j \in \{-1, 1\}$ in order to make the logit index be consistent with the data label. Moreover, we assume that the each weight is initialized from a random draw of Gaussian random variable $\sim N(0, \sigma_0^2)$ with $\sigma_0 = \Theta\big(d^{-1/4}\big)$.

Training objective. Given the training data $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$, we consider to learn the model parameter $\mathbf{W}$ by optimizing the empirical loss function with weight decay regularization

$$L(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} L_i(\mathbf{W}) + \frac{\lambda}{2} \|\mathbf{W}\|_F^2, \tag{6.2.2}$$

where $L_i(\mathbf{W}) = -\log \frac{e^{F_{y_i}(\mathbf{W},\mathbf{x}_i)}}{\sum_{j\in\{-1,1\}} e^{F_j(\mathbf{W},\mathbf{x}_i)}}$ denotes the individual loss for the data $(\mathbf{x}_i, y_i)$ and $\lambda \geq 0$ is the regularization parameter. In particular, the regularization parameter can be arbitrary as long as it satisfies $\lambda \in \big(0, \lambda_0\big)$ with $\lambda_0 = \Theta\big(\frac{1}{d^{(q-1)/4} n \cdot \mathrm{polylog}(n)}\big)$. We claim that the λ_0 is the largest feasible regularization parameter that the training process will not stuck at the origin point (recall that $L(\mathbf{W})$ admits zero gradient at $\mathbf{W} = \mathbf{0}$.)

Training algorithms. In this paper, we consider gradient descent and Adam with full gradient[1]. In particular, starting from initialization $\mathbf{W}^{(0)} = \{\mathbf{w}_{j,r}^{(0)}, j = \{\pm 1\}, r \in [m]\}$, the gradient descent update rule is

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \eta \cdot \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)}),$$

where η is the learning rate. Meanwhile, Adam store historical gradient information in the momentum $\mathbf{m}^{(t)}$ and a vector $\mathbf{v}^{(t)}$ as follows

$$\mathbf{m}_{j,r}^{(t+1)} = \beta_1 \mathbf{m}_{j,r}^{(t)} + (1-\beta_1) \cdot \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)}), \tag{6.2.3}$$

$$\mathbf{v}_{j,r}^{(t+1)} = \beta_2 \mathbf{v}_{j,r}^{(t)} + (1-\beta_2) \cdot [\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})]^2, \tag{6.2.4}$$

and entry-wisely adjusts the learning rate:

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \eta \cdot \mathbf{m}_{j,r}^{(t)} / \sqrt{\mathbf{v}_{j,r}^{(t)}}, \tag{6.2.5}$$

where β_1, β_2 are the hyperparameters of Adam (a popular choice in practice is $\beta_1 = 0.9$, and $\beta_2 = 0.99$), and in (6.2.4) and (6.2.5), the square $(\cdot)^2$, square root $\sqrt{\cdot}$, and division $\cdot/\cdot$ all denote entry-wise calculations. We would like to clarify the original Adam paper [KB15] considers to normalize the gradient $\mathbf{m}_{j,r}^{(t)}$ via $\sqrt{\mathbf{v}_{j,r}^{(t)} + \epsilon}$, while the small bias term ϵ is ignored in our paper. In practice, tuning ϵ can help improve the generalization ability of Adam in practice [CSN19], as it allows to make a trade-off between the normalized gradient update and gradient update (i.e., GD). We also do not consider the initialization bias correction in the original Adam paper for the ease of analysis.

[1]Our theory can still hold when applying mini-batch stochastic gradients, which we will discuss in later.

6.3 Main Results

In this section we will state the main theorems in this paper. We first provide the learning guarantees of Adam and Gradient descent for training a two-layer CNN model in the following theorem. Recall that in this setting the training objective is nonconvex.

Theorem 6.3.1 (Nonconvex setting)**.** Consider a two-layer CNN defined in (6.2.1) with $d = \Omega(n^4)$ and regularized training objective (6.2.2) with a regularization parameter $\lambda > 0$, suppose the network width is $m = \mathrm{polylog}(n)$ and the data distribution follows Definition 6.2.1, then we have the following guarantees on the training and test errors for the models trained by Adam and Gradient descent:

- Suppose we run **Adam** for $T = \frac{\mathrm{poly}(n)}{\eta}$ iterations with $\eta = \frac{1}{\mathrm{poly}(n)}$, then with probability at least $1 - O(n^{-1})$, we can find a NN model $\mathbf{W}^*_{\mathrm{Adam}}$ such that $\|\nabla L(\mathbf{W}^*_{\mathrm{Adam}})\|_1 \leq \frac{1}{T\eta}$. Moreover, the model $\mathbf{W}^*_{\mathrm{Adam}}$ also satisfies:
  - Training error is zero: $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left[F_{y_i}(\mathbf{W}^*_{\mathrm{Adam}}, \mathbf{x}_i) \leq F_{-y_i}(\mathbf{W}^*_{\mathrm{Adam}}, \mathbf{x}_i)\right] = 0$.
  - Test error is high: $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[F_y(\mathbf{W}^*_{\mathrm{Adam}}, \mathbf{x}) \leq F_{-y}(\mathbf{W}^*_{\mathrm{Adam}}, \mathbf{x})\big] \geq \frac{1}{2}$.
- Suppose we run **gradient descent** for $T = \frac{\mathrm{poly}(n)}{\eta}$ iterations with learning rate $\eta = \frac{1}{\mathrm{poly}(n)}$, then with probability at least $1 - O(n^{-1})$, we can find a NN model $\mathbf{W}^*_{\mathrm{GD}}$ such that $\|\nabla L(\mathbf{W}^*_{\mathrm{GD}})\|_F^2 \leq \frac{1}{T\eta}$. Moreover, the model $\mathbf{W}^*_{\mathrm{GD}}$ also satisfies:
  - Training error is zero: $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\left[F_{y_i}(\mathbf{W}^*_{\mathrm{GD}}, \mathbf{x}_i) \leq F_{-y_i}(\mathbf{W}^*_{\mathrm{GD}}, \mathbf{x}_i)\right] = 0$.
  - Test error is nearly zero: $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[F_y(\mathbf{W}^*_{\mathrm{GD}}, \mathbf{x}) \leq F_{-y}(\mathbf{W}^*_{\mathrm{GD}}, \mathbf{x})\big] = \frac{1}{\mathrm{poly}(n)}$.

From the optimization perspective, Theorem 6.3.1 shows that both Adam and GD can be guaranteed to find a point with a very small gradient, which can also achieve zero classification error on the training data. Moreover, it can be seen that given the same iteration number T and learning rate η, Adam can be guaranteed to find a point with up to $1/(T\eta)$ gradient norm in ℓ_1 metric, while gradient descent can only be guaranteed to find a point

with up to $1/\sqrt{T\eta}$ gradient norm in ℓ_2 metric. This suggests that Adam could enjoy a faster convergence rate compared to SGD in the training process, which is consistent with the practice findings. We would also like to point out that there is no contradiction between our result and the recent work [RKK18] showing that Adam can fail to converge, as the counterexample in [RKK18] is for the online version of Adam, while we study the full batch Adam.

In terms of the test performance, their generalization abilities are largely different, even with weight decay regularization. In particular, the output of gradient descent can generalize well and achieve nearly zero test error, while the output of Adam gives nearly $1/2$ test error. In fact, this gap is due to two major aspects of the training process: (1) At the early stage of training where weight decay exhibits negligible effect, Adam and GD behave very differently. In particular, Adam prefers the denser and thus tends to fit the noise vectors $\boldsymbol{\xi}$, gradient descent prefers the data patch of larger ℓ_2 norm and thus will learn the feature patch; (2) At the late stage of training where the weight decay regularization cannot be ignored, both Adam and gradient descent will be enforced to converge to a *local minimum* of the regularized objective, which maintains the pattern learned in the early stage. Consequently, the model learned by Adam will be biased towards the noise patch to fit the feature noise vector $-\alpha y\mathbf{v}$, which is opposite in direction to the true feature vector and therefore leads to a test error no better than a random guess. More details about the training behaviors of Adam and GD are given in Section 6.4.

Theorem 6.3.1 shows that when optimizing a nonconvex training objective, Adam and gradient descent will converge to different global solutions with different generalization errors, even with weight decay regularization. In comparison, the following theorem gives the learning guarantees of Adam and gradient descent when optimizing convex and smooth training objectives (e.g., linear model $F(\mathbf{w}, \mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ with logistic loss).

Theorem 6.3.2 (Convex setting)**.** For any convex and smooth training objective with positive regularization parameter λ, suppose we run **Adam** and **gradient descent** for

$T = \frac{\mathrm{poly}(n)}{\eta}$ iterations, then with probability at least $1 - n^{-1}$, the obtained parameters $\mathbf{W}^*_{\mathrm{Adam}}$ and $\mathbf{W}^*_{\mathrm{GD}}$ satisfy that $\|\nabla L(\mathbf{W}^*_{\mathrm{Adam}})\|_1 \leq \frac{1}{T\eta}$ and $\|\nabla L(\mathbf{W}^*_{\mathrm{Adam}})\|_2^2 \leq \frac{1}{T\eta}$ respectively. Moreover, let $F(\mathbf{W}, \mathbf{x}) \in \mathbb{R}$ be the output of the convex model with parameter $\mathbf{W}$ and input $\mathbf{x}$, it holds that:

- Training errors are the same, $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\big[y_i F(\mathbf{W}^*_{\mathrm{Adam}}, \mathbf{x}_i) > 0\big] = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\big[y_i F(\mathbf{W}^*_{\mathrm{GD}}, \mathbf{x}_i) > 0\big]$.
- Test errors are nearly the same: $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y_i F(\mathbf{W}^*_{\mathrm{Adam}}, \mathbf{x}_i) > 0\big] = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y_i F(\mathbf{W}^*_{\mathrm{GD}}, \mathbf{x}) > 0\big] \pm 1/\mathrm{poly}(n)$.

Theorem 6.3.2 shows that when optimizing a convex and smooth training objective (e.g., a linear model with logistic loss) with weight decay regularization, both Adam and gradient can converge to almost the same solution and enjoy very similar generalization performance. The proof will be relying on the strong convexity of the training objective and the convergence (to the first-order stationary) guarantee of Adam [DBB20] and GD. Combining this result and Theorem 6.3.1, it is clear that the inferior generalization performance is closely tied to the nonconvex landscape of deep learning, and cannot be understood by standard weight decay regularization.

6.4 Proof Outline of the Main Results

In this section we provide the proof sketch of Theorem 6.3.1 and explain the different generalization abilities of the models found by gradient descent and Adam.

Before moving to the proof of main results, we first give the following lemma which shows that for data generated from the data distribution $\mathcal{D}$ in Definition 6.2.1, with high probability all noise vectors $\{\boldsymbol{\xi}_i\}_{i=1,\ldots,n}$ have nearly disjoint supports.

Lemma 6.4.1. Let $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$ be the training dataset generated by Definition 6.2.1. Moreover, recall that $\mathbf{x}_i = [y_i\mathbf{v}^\top, \boldsymbol{\xi}_i^\top]^\top$ (or $\mathbf{x}_i = [\boldsymbol{\xi}_i^\top, y_i\mathbf{v}^\top]^\top$), let $\mathcal{B}_i = \mathrm{supp}(\boldsymbol{\xi}_i)\backslash\{1\}$ be the

support of $\boldsymbol{\xi}_i$ except the first coordinate. Then with probability at least $1 - n^{-2}$, $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset$ for all $i \neq j$.

This lemma implies that the optimization of each coordinate of the model parameter $\mathbf{W}$, except for the first one, is mostly determined by only one training data. Technically, this lemma can greatly simplify the analysis for Adam so that we can better illustrate its optimization behavior and explain the generalization performance gap between Adam and gradient descent.

Proof outline. For both Adam and gradient descent, we will show that the training process can be decomposed into two stages. In the first stage, which we call *pattern learning stage*, the weight decay regularization will be less important and can be ignored, while the algorithms tend to learn the pattern from the training data. In particular, we will show that in the patter learning stage, the optimization algorithms have different *algorithmic bias*: Adam tends to fit the noise patch while gradient descent will mainly learn the feature patch. In the second stage, which we call it *regularization stage*, the effect of regularization cannot be neglected, which will regularize the algorithm to converge at some local stationary points. However, due to the nonconvex landscape of the training objective, the pattern learned in the first stage will remain unchanged, even when running an infinitely number of iterations.

6.4.1 Proof sketch for Adam

Recall that in each iteration of Adam, the model weight is updated by using a moving-averaged gradient, normalized by a moving average of the historical gradient squares. As pointed out in [BH18; BWA18], Adam behaves similarly to sign gradient descent (signGD) when using sufficiently small step size or the moving average parameters β_1, β_2 are nearly zero. This motivates us to understand the optimization behavior of signGD and then extends it to Adam using their similarities. In particular, sign gradient descent updates the model

parameter according to the following rule:

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t+1)} - \eta \cdot \mathrm{sgn}(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})).$$

Recall that each data has two patches: feature and noise patches. By Lemma 6.4.1 and the data distribution (see Definition 6.2.1), we know that all noise vectors $\{\boldsymbol{\xi}_i\}_{i=1,\ldots,n}$ are supported on disjoint coordinates, except the first one. For data point $\mathbf{x}_i$, let $\mathcal{B}_i$ denote its support, except the first coordinate. In the subsequent analysis, we will always assume that those $\mathcal{B}_i$'s are disjoint, i.e., $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset$ if $i \neq j$.

Next we will characterize two aspects of the training process: *feature learning* and *noise memorization*. Mathematically, we will focus on two quantities: $\langle \mathbf{w}_{j,r}^{(t)}, j \cdot \mathbf{v} \rangle$ and $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$. In particular, given the training data $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i = [y_i \mathbf{v}^\top, \boldsymbol{\xi}_i^\top]^\top$, larger $\langle \mathbf{w}_{y_i,r}^{(t)}, y_i \cdot \mathbf{v} \rangle$ implies better feature learning and larger $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ represents better noise memorization. Then regarding the feature vector $\mathbf{v}$ that only has nonzero entry at the first coordinate, we have the following by the update rule of signGD

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t+1)}, j\mathbf{v} \rangle &= \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle - \eta \cdot \big\langle \mathrm{sgn}\big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})\big), j\mathbf{v} \big\rangle \qquad (6.4.1) \\
&= \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle + j\eta \cdot \mathrm{sgn}\bigg(\sum_{i=1}^{n} y_i \ell_{j,i}^{(t)} \big[\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle) - \alpha \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \big] - n\lambda \mathbf{w}_{j,r}^{(t)}[1] \bigg),
\end{aligned}$$

where $\ell_{j,i}^{(t)} := \mathbb{1}_{y_i = j} - \mathrm{logit}_j(F, \mathbf{x}_i)$ and $\mathrm{logit}_j(F, \mathbf{x}_i) = \frac{e^{F_j(\mathbf{W}, \mathbf{x}_i)}}{\sum_{k \in \{-1,1\}} e^{F_k(\mathbf{W}, \mathbf{x}_i)}}$. From (6.4.1) we can observe three terms in the signed gradient. Specifically, the first term represents the gradient over the feature patch, the second term stems from the feature noise term in the noise patch (see Definition 6.2.1), and the last term is the gradient of the weight decay regularization. On the other hand, the memorization of the noise vector $\boldsymbol{\xi}_i$ can be described by the following,

$$\begin{aligned}
\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle - \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle &= -\eta \cdot \big\langle \mathrm{sgn}\big(\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^{(t)})\big), \boldsymbol{\xi}_i \big\rangle \qquad (6.4.2) \\
&= \eta \sum_{k \in \mathcal{B}_i \cup \{1\}} \mathrm{sgn}\bigg(\ell_{y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \boldsymbol{\xi}_i[k] - n\lambda \mathbf{w}_{y_i,r}^{(t)}[k] \bigg) \cdot \boldsymbol{\xi}_i[k].
\end{aligned}$$

Throughout the proof, we will show that the training process of Adam can be decomposed into two stages: *pattern learning stage* and *regularization stage*. In the first stage, the algorithm learns the pattern of training data quickly, without being affected by the regularization

term. In the second stage, the training data has already been correctly classified since the pattern has been well captured, the regularization will play an important role in the training process and guide the model to converge.

Stage I: Learning the pattern. Mathematically, the first stage is defined as the iterations that the neural network output is smaller than some constant. In this stage, all training data remains under-fitted and can provide large gradient for model training, and the effect of weight decay regularization can be ignored due to our choice of λ. We will show that in this stage the inner product $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ grows much faster than $\langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle$ since feature learning only makes use of the first coordinate of the gradient, while noise memorization could take advantage of all the coordinates in $\mathcal{B}_i$ (see (6.4.2), note that $|\mathcal{B}_i| = s \gg 1$).

Lemma 6.4.2 (General results in Stage I)**.** Suppose the training data is generated according to Definition 6.2.1, assume $\lambda = o(\sigma_0^{q-2}\sigma_p/n)$ and $\eta = 1/\mathrm{poly}(d)$, then for any $t \leq T_0$ with $T_0 = \widetilde{O}\big(\frac{1}{\eta s \sigma_p}\big)$ and any $i \in [n]$,

$$\langle \mathbf{w}_{j,r}^{(t+1)}, j \cdot \mathbf{v} \rangle \leq \langle \mathbf{w}_{j,r}^{(t)}, j \cdot \mathbf{v} \rangle + \Theta(\eta), \quad \langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle = \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle + \widetilde{\Theta}(\eta s \sigma_p).$$

Since $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ enjoys much faster increasing rate than that of $\langle \mathbf{w}_{j,r}^{(t)}, j \cdot \mathbf{v} \rangle$, after a certain number of iterations, the learning of noise patch will dominate the learning of feature patch (i.e., $\alpha\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) > \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle)$). Thus, by (6.4.1), the model will tend to fit the feature noise in the noise patch (i.e., $-\alpha y_i \mathbf{v}$), leading to a flipped feature learning phenomenon.

Lemma 6.4.3 (Flipping the feature learning)**.** Suppose the training data is generated according to Definition 6.2.1, $\alpha \geq \widetilde{\Theta}\big((s\sigma_p)^{1-q} \vee \sigma_0^{q-1}\big)$ and $\sigma_0 < \widetilde{O}((s\sigma_p)^{-1})$, then for any $t \in [T_r, T_0]$ with $T_r = \widetilde{O}\big(\frac{\sigma_0}{\eta s \sigma_p \alpha^{1/(q-1)}}\big) \leq T_0$,

$$\langle \mathbf{w}_{j,r}^{(t+1)}, j \cdot \mathbf{v} \rangle = \langle \mathbf{w}_{j,r}^{(t)}, j \cdot \mathbf{v} \rangle - \Theta(\eta).$$

Moreover, it holds that

- $\mathbf{w}_{j,r}^{(T_0)}[1] = -\mathrm{sgn}(j) \cdot \widetilde{\Omega}\big(\frac{1}{s\sigma_p}\big)$

- $\mathbf{w}_{j,r}^{(T_0)}[k] = \mathrm{sgn}(\boldsymbol{\xi}_i[k]) \cdot \widetilde{\Omega}\big(\frac{1}{s\sigma_p}\big)$ or $\mathbf{w}_{j,r}^{(T_0)}[k] = \pm\widetilde{O}(\eta)$ for $k \in \mathcal{B}_i$ with $y_i = j$
- $\mathbf{w}_{j,r}^{(T_0)}[k] = \pm\widetilde{O}(\eta)$ otherwise.

From Lemma 6.4.3 it can be observed that at the iteration T_0, the sign of the first coordinate of $\mathbf{w}_{j,r}^{(T_0)}$ is different from that of the true feature, i.e., $j \cdot \mathbf{v}$. This implies that at the end of the first training stage, the model is biased towards the noise patch to fit the feature noise.

Stage II: Regularizing the model. In this stage, as the neural network output becomes larger, part of training data starts to be well fitted and gives smaller gradient. As a consequence, the feature learning and noise memorization processes will be slowed down and the weight decay regularization term cannot be ignored. However, although weight decay regularization can prevent the model weight from being too large, it will maintain the pattern learned in Stage I and cannot push the model back to "forget" the noise and learn the feature and stops at some local stationary points. We summarize these results in the following lemma.

Lemma 6.4.4 (Maintain the pattern)**.** If $\alpha = O\big(s\sigma_p^2/n\big)$ and $\eta = o(\lambda)$, then let $r^* = \arg\max_{r\in[m]}\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle$, for any $t \geq T_0$, $i \in [n]$, $j \in [2]$ and $r \in [m]$, it holds that

$$\langle \mathbf{w}_{y_i,r^*}^{(t)}, \boldsymbol{\xi}_i\rangle = \widetilde{\Theta}(1),\ \sum_{k\in\mathcal{B}_i} |\mathbf{w}_{y_i,r^*}^{(t)}[k]| \cdot |\boldsymbol{\xi}_i[k]| = \widetilde{\Theta}(1),\ \langle \mathbf{w}_{j,r}^{(t)}, \mathrm{sgn}(j)\cdot\mathbf{v}\rangle \in [-o(1), O(\lambda^{-1}\eta)].$$

Lemma 6.4.4 shows that in the second stage, $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle$ will always be large while $\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\cdot\mathbf{v}\rangle$ is still negative, or positive but extremely small. Next we will show that within polynomial steps, the algorithm can be guaranteed to find a point with small gradient.

Lemma 6.4.5 (Convergence guarantee)**.** If $\eta = O(d^{-1/2})$, then for any t it holds that

$$L(\mathbf{W}^{(t+1)}) - L(\mathbf{W}^{(t)}) \leq -\eta\|\nabla L(\mathbf{W}^{(t)})\|_1 + \widetilde{\Theta}(\eta^2 d).$$

Lemma 6.4.5 shows that we can pick a sufficiently small η and $T = \mathrm{poly}(n)/\eta$ to ensure that the algorithm can find a point with up to $O(1/(T\eta))$ in ℓ_1 norm. Then we can show that

given the results in Lemma 6.4.4, the formula of the algorithm output $\mathbf{W}^*$ can be precisely characterized, which we can show that $\langle \mathbf{w}_{y_i,r}^*, y_i \cdot \mathbf{v} \rangle < 0$. This implies that the output model will be biased to fit the feature noise $-\alpha y \mathbf{v}$ but not the true one $\mathbf{v}$. Then when it comes to a fresh test example the model will fail to recognize its true feature. Also note that the noise in the test data is nearly independent of the noise in training data. Consequently, the model will not be able to identify the label of the test data and therefore cannot be better than a random guess.

6.4.2 Proof sketch for gradient descent

Similar to the proof for Adam, we also decompose the entire training process into two stages.

Stage I: Learning the pattern. In this stage the gradient from training loss function is large and and the effect of regularization can be ignored. Unlike Adam that is sensitive to the sparsity of the feature vector or noise vector, gradient descent is more focusing on the ℓ_2 norm of them, where the vector (which can be either feature vector or noise vector) with larger ℓ_2 norm is more likely to be discovered and learnt by GD. Note that the feature vector has a larger ℓ_2 norm than the noise, we can show that, in the following lemma, gradient descent will learn the feature vector very quickly, while barely tend to memorize the noise.

Lemma 6.4.6. Let $\Lambda_j^{(t)} = \max_{r\in[m]} \langle \mathbf{w}_{j,r}^{(t+1)}, j \cdot \mathbf{v} \rangle$, $\Gamma_{j,i}^{(t)} = \max_{r\in[m]} \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$, and $\Gamma_j^{(t)} = \max_{i:y_i=j} \Gamma_{j,i}^{(t)}$. Let T_j be the iteration number that $\Lambda_j^{(t)}$ reaches $\Theta(1/m) = \widetilde{\Theta}(1)$, then we have

$$T_j = \widetilde{\Theta}(\sigma_0^{2-q}) \quad \text{for all } j \in \{-1, 1\}.$$

Moreover, let $T_0 = \max_j \{T_j\}$, then for all $t \leq T_0$ it holds that $\Gamma_j^{(t)} = \widetilde{O}(\sigma_0)$ for all $j \in \{-1, 1\}$.

Stage II: Regularizing the model. Similar to Lemma 6.4.4, we show that in the second stage at which the impact of weight decay regularization cannot be ignored, the pattern of the training data learned in the first stage will remain unchanged.

Lemma 6.4.7. If $\eta \leq O(\sigma_0)$, it holds that $\Lambda_j^{(t)} = \widetilde{\Theta}(1)$ and $\Gamma_j^{(t)} = \widetilde{O}(\sigma_0)$ for all $t \geq \min_j T_j$.

The following lemma further shows that within polynomial steps, gradient descent is guaranteed to find a point with small gradient.

Lemma 6.4.8. If the learning rate satisfies $\eta = o(1)$, then for any $t \geq 0$ it holds that

$$L(\mathbf{W}^{(t+1)}) - L(\mathbf{W}^{(t)}) \leq -\frac{\eta}{2}\|\nabla L(\mathbf{W}^{(t)})\|_F^2.$$

Lemma 6.4.8 shows that we can pick a sufficiently small η and $T = \mathrm{poly}(n)/\eta$ to ensure that gradient descent can find a point with up to $O(1/(T\eta)^{1/2})$ in ℓ_2 norm. By Lemma 6.4.7, it is clear that the output model of GD can well learn the feature vector while memorizing nearly nothing from the noise vectors, which can therefore achieve nearly zero test error.

6.5 Experiments

In this section we perform numerical experiments on the synthetic data generated according to Definition 6.2.1 to verify our main results. In particular, we set the problem dimension $d = 1000$, the training sample size $n = 200$ (100 positive examples and 100 negative examples), feature vector $\mathbf{v} = [1, 0, \ldots, 0]^\top$, noise sparsity $s = 0.1d = 100$, standard deviation of noise $\sigma_p = 1/s^{1/2} = 0.1$, feature noise strength $\alpha = 0.2$, initialization scaling $\sigma_0 = 0.01$, regularization parameter $\lambda = 1 \times 10^{-5}$, network width $m = 20$, activation function $\sigma(z) = \max\{0, z\}^3$, total iteration number $T = 1 \times 10^4$, and the learning rate $\eta = 5 \times 10^{-5}$ for Adam (default choices of β_1 and β_2 in pytorch), $\eta = 0.02$ for GD.

We first report the training error and test error achieved by the solutions found by SGD and Adam in Table 6.2, where the test error is calculated on a test dataset of size 10^4. It is clear that both Adam and SGD can achieve zero training error, while they have entirely different results on the test data: SGD generalizes well and achieve zero test error; Adam generalizes worse than SGD and gives > 0.5 test error, which verifies our main result (Theorem 6.3.1).

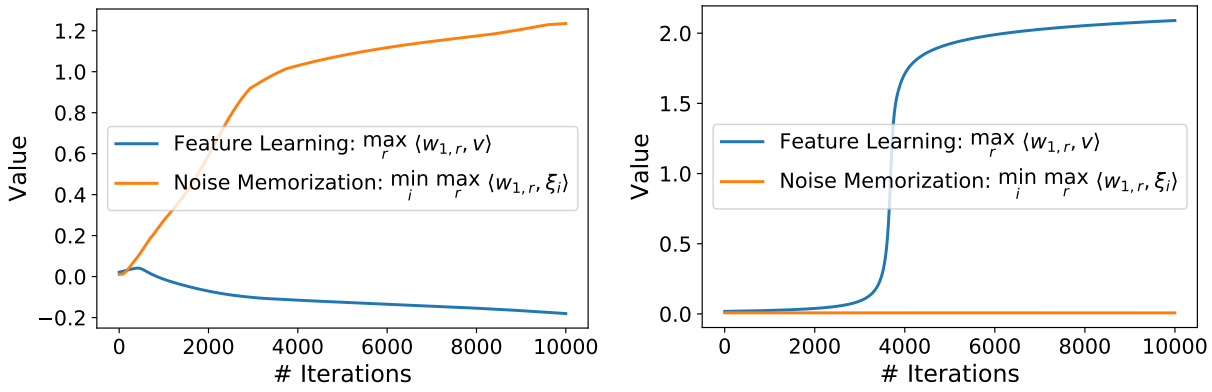
Algorithm	Adam	SGD
Training error	0	0
Test error	0.884	0

Table 6.2: Training and test errors achieved by GD and Adam.

Moreover, we also calculate the inner products: $\max_r \langle \mathbf{w}_{1,r}, \mathbf{v} \rangle$ and $\min_i \max_r \langle \mathbf{w}_{1,r}, \boldsymbol{\xi}_i \rangle$, representing feature learning and noise memorization respectively, to verify our key lemmas. Here we only consider positive examples as the results for negative examples are similar. The results are reported in Figure 6.2. For Adam, from Figure 6.2(a), it can be seen that the algorithm will perform feature learning in the first few iterations and then entirely forget the feature (but fit feature noise), i.e., the feature learning is flipped, which verifies Lemma 6.4.3. In the meanwhile, the noise memorization happens in the entire training process and enjoys much faster rate than feature learning, which verifies Lemma 6.4.2. In addition, we can also observe that there are two stages for the increasing of $\min_i \max_r \langle \mathbf{w}_{1,r}, \boldsymbol{\xi}_i \rangle$: in the first stage $\min_i \max_r \langle \mathbf{w}_{1,r}, \boldsymbol{\xi}_i \rangle$ increases linearly, and in the second stage its increasing speed gradually slows down and $\min_i \max_r \langle \mathbf{w}_{1,r}, \boldsymbol{\xi}_i \rangle$ will remain in a constant order. This verifies Lemma 6.4.2 and Lemma 6.4.4. For GD, from Figure 6.2(b), it can be seen that the feature learning will dominate the noise memorization: feature learning will increases to a constant in the first stage and then remains in a constant order in the second stage; noise memorization will keep in a low level which is nearly the same as that at the initialization. This verifies Lemmas 6.4.6 and 6.4.7.

6.6 Extensions to Mini-batch Stochastic Gradients

One natural extension of our paper is proving the separation between mini-batch SGD and mini-batch Adam, which we believe is not difficult. In particular, let $\mathcal{I}_t$ of size B be the set



(a) Adam (b) GD

Figure 6.2: Visualization of the feature learning ($\max_r \langle \mathbf{w}_{1,r}, \mathbf{v} \rangle$) and noise memorization ($\min_i \max_r \langle \mathbf{w}_{1,r}, \boldsymbol{\xi}_i \rangle$) in the training process.

of indices of the mini-batch data used in the t-th iteration, the update rule of SGD is

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \eta \cdot \frac{1}{B} \sum_{i \in \mathcal{I}_t} \nabla_{\mathbf{w}_{j,r}} L_i(\mathbf{W}^{(t)}) - \gamma \mathbf{w}_{j,r}^{(t)}.$$

The update rule of mini-batch Adam is

$$\mathbf{m}_{j,r}^{(t+1)} = \beta_1 \mathbf{m}_{j,r}^{(t)} + (1-\beta_1) \cdot \bigg[\frac{1}{B} \sum_{i \in \mathcal{I}_t} \nabla_{\mathbf{w}_{j,r}} L_i(\mathbf{W}^{(t)}) - \gamma \mathbf{w}_{j,r}^{(t)} \bigg],$$
$$\mathbf{v}_{j,r}^{(t+1)} = \beta_2 \mathbf{v}_{j,r}^{(t)} + (1-\beta_2) \cdot \bigg[\frac{1}{B} \sum_{i \in \mathcal{I}_t} \nabla_{\mathbf{w}_{j,r}} L_i(\mathbf{W}^{(t)}) - \gamma \mathbf{w}_{j,r}^{(t)} \bigg]^2,$$

and

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \eta \cdot \mathbf{m}_{j,r}^{(t)} / \sqrt{\mathbf{v}_{j,r}^{(t)}}.$$

Then we will take a deeper look at the speeds of feature learning and noise learning for mini-batch SGD and Adam, where we focus on the period that $|\langle \mathbf{w}_{j,r}^t, \mathbf{v} \rangle|, |\langle \mathbf{w}_{j,r}^t, \boldsymbol{\xi}_i \rangle| = o(1)$ for all j, i, and r (i.e., the pattern learning stage). This further implies that $|\ell_{j,i}^{(t)}| = 0.5 \pm o(1)$ for all j, i, and t. Thus in the following, we will assume that all $|\ell_{j,i}^{(t)}|$ has nearly the same quantity.

Feature Learning. First, according to Definition 6.2.1, we know that the feature vector $\mathbf{v}$ and feature noise are the same for all data, which implies that the learning pattern of the feature coordinate will be largely the same as that of full-batch algorithms. In particular, for mini-batch Adam, we can show that the update of the first coordinate (i.e., feature coordinate) is similar to sign-GD when using sufficiently small learning rate $\eta = 1/\mathrm{poly}(d)$ since all stochastic gradients $\nabla L_i(\mathbf{W}^{(t)})$ have the same component in this coordinate. Then using the fact that $|\ell_{j,i}^{(t)}|$'s are nearly the same for all i, we have

$$\langle \mathbf{w}_{j,r}^{(t+1)}, j\mathbf{v} \rangle \sim \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v} \rangle + j\eta \cdot \mathrm{sgn}\bigg(\sum_{i=1}^{n} y_i \ell_{j,i}^{(t)} \big[\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle) - \alpha \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \big] - n\lambda \mathbf{w}_{j,r}^{(t)}[1] \bigg).$$

which is the same as full-batch Adam (see (6.4.1)). For SGD, using the fact that $|\ell_{j,i}^{(t)}|$'s are nearly the same for all i, we can get that

$$\begin{aligned} \langle \mathbf{w}_{j,r}^{(t+1)}, j\mathbf{v} \rangle \sim{}& (1 - \eta\lambda) \cdot \langle \mathbf{w}_{j,r}^{(t)}, j \cdot \mathbf{v} \rangle \\ &+ \frac{\eta}{n} \cdot j \cdot \bigg(\sum_{i=1}^{n} y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle) - \alpha \sum_{i=1}^{n} y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \bigg) \end{aligned}$$

which is also the same as that of GD (see (6.7.28)).

Noise Memorization. Note that due to the normalization term $\mathbf{v}_{j,r}^{(t)}$ in the Adam update, all coordinates will be updated with nearly the same amount. Therefore, we only need to count the number of coordinates that are updated by full-batch Adam and mini-batch Adam.

Recall that we have shown that using mini-batch gradients will not affect the feature learning. However, the noise memorization will be slightly different, since in each iteration, full-batch Adam can update $\Theta(ns)$ coordinates while mini-batch Adam can only update $\widetilde{\Theta}(Bs)$ coordinates. To show this, we note that for any coordinate $k \neq 1$, the gradient momentum of full-batch Adam is

$$\mathbf{m}_{j,r}^{(t)}[k] \sim \sum_{\tau=0}^{\bar{\tau}} \beta_1^{\tau}(1 - \beta_1) \cdot \frac{1}{n} \sum_{i \in [n]} \Big[\nabla_{\mathbf{w}_{j,r}} L_i(\mathbf{W}^{(t-\tau)})[k] + \lambda \mathbf{w}_{j,r}^{(t-\tau)}[k] \Big],$$

while for mini-batch Adam,

$$\mathbf{m}_{j,r}^{(t)}[k] \sim \sum_{\tau=0}^{\bar{\tau}} \beta_1^\tau (1-\beta_1) \cdot \frac{1}{B} \sum_{i \in \mathcal{I}_{t-\tau}} \Big[\nabla_{\mathbf{w}_{j,r}} L_i(\mathbf{W}^{(t-\tau)})[k] + \lambda \mathbf{w}_{j,r}^{(t-\tau)}[k] \Big],$$

where we only maintain the recent $\bar{\tau} = \text{polylog}(n)$ gradients since for $\tau \leq t - \bar{\tau}$, the decaying terms $(\beta_1)^\tau \leq (\beta_i)^{\bar{\tau}}$ becomes negligible. Therefore, by comparing the above two equations and applying Definition 6.2.1, it is clear that for full-batch Adam can update all noise coordinates, i.e., $k \in \cup_{i \in [n]} \mathcal{B}_i$, which is of size $\Theta(ns)$. In contrast, mini-batch Adam can only update a subset of noise coordinates, i.e., $k \in \cup_{\tau \in [\bar{\tau}]} \cup_{i \in [\mathcal{I}_{t-\tau}]} \mathcal{B}_i$, which is of size $\bar{\tau} B s = \widetilde{\Theta}(Bs)$. This further implies that in each epoch (one pass of the data, $\Theta(n/B)$ steps), the noise coordinates in $\mathcal{B}_i$ will be updated by *mini-batch Adam* in at most $\bar{\tau} = \widetilde{\Theta}(1)$ steps, while within the same amount of iterations, the noise coordinates in $\mathcal{B}_i$ will be updated by *full-batch* Adam for $\Theta(n/B)$ steps, suggesting that mini-batch Adam admits a slower rate of noise memorization by a $\widetilde{\Theta}(n/B)$ factor.

For SGD, it is easy to show that the rate of noise memorization will still be nearly the same as that of GD. In particular, during each training epoch ($\Theta(n/B)$ steps), SGD will learn the noise vector $\boldsymbol{\xi}_i$ in only one step with the mini-batch gradient $\frac{1}{B} \nabla L_i(\mathbf{W}^\tau)$ for some τ in this epoch, while within the same amount of steps, GD will learn the noise vector $\boldsymbol{\xi}_i$ in all $\Theta(n/B)$ steps but with strength $\frac{1}{n} \nabla L_i(\mathbf{W}^\tau)$, giving the same total learning ability. This suggests that SGD admits a nearly the same rate of noise memorization compared to GD.

To sum up, we have shown that (1) mini-batch SGD and mini-batch Adam will not change the learning rate of feature vector $\mathbf{v}$ compared to their full-batch counterparts; (2) mini-batch Adam reduces the noise memorization rate of full-batch Adam by a $\widetilde{\Theta}(n/B)$ factor, while mini-batch SGD has nearly the same noise memorization rate compared to full-batch GD. Additionally, recall that in our paper, the separation between Adam and GD is characterized by a $\text{poly}(d)$ factor: the speed of feature learning in Adam and GD, and the rate of noise memorization in GD are both in the order of $O(\eta)$ (in each step), while the rate of noise memorization in Adam is proportional to the number of nonzero entries,

which is in the order of $\eta \cdot \mathrm{poly}(d)$. Therefore, the separation between mini-batch SGD and mini-batch Adam in terms of the generalization error can still hold under a stronger over-parameterization condition (the previous $\mathrm{poly}(d)$ separation needs to dominate the $\widetilde{\Theta}(n/B)$ improvement brought by mini-batch Adam).

6.7 Proof of Theorem 6.3.1: Nonconvex Case

In the beginning of the proof we first present the following useful lemma.

6.7.1 Preliminaries

We first recall the magnitude of all parameters:

$$d = \mathrm{poly}(n),\ \eta = \frac{1}{\mathrm{poly}(n)},\ s = \Theta\bigg(\frac{d^{1/2}}{n^2}\bigg),\ \sigma_p^2 = \Theta\bigg(\frac{1}{s \cdot \mathrm{polylog}(n)}\bigg),\ \sigma_0^2 = \Theta\bigg(\frac{1}{d^{1/2}}\bigg),$$
$$m = \mathrm{polylog}(n),\ \alpha = \Theta\big(\sigma_p \cdot \mathrm{polylog}(n)\big),\ \lambda = O\bigg(\frac{1}{d^{(q-1)/4} n \cdot \mathrm{polylog}(n)}\bigg).$$

Here $\mathrm{poly}(n)$ denotes a polynomial function of n with degree of a sufficiently large constant, $\mathrm{poly}(n)$ denotes a polynomial function of $\log(n)$ with degree of a sufficiently large constant. Based on the parameter configuration, we claim that the following equations hold, which will be frequently used in the subsequent proof.

$$\lambda = o\bigg(\frac{\sigma_0^{q-2}\sigma_p}{n}\bigg),\ \alpha = \omega\big((s\sigma_p)^{1-q}\sigma_0^{q-1}\big),\ \sigma_0 = o\bigg(\frac{1}{s\sigma_p}\bigg), \alpha = o\bigg(\frac{s\sigma_p^2}{n}\bigg),\ \eta = o\Big(\lambda\sigma_0^q\sigma_p^q\Big).$$

Lemma 6.7.1 (Non-overlapping support)**.** Let $\{(\mathbf{x}_i, y_i)\}_{i=1,\ldots,n}$ be the training dataset sampled according to Definition 6.2.1. Moreover, let $\mathcal{B}_i = \mathrm{supp}(\boldsymbol{\xi}_i)\backslash\{1\}$ be the support of $\mathbf{x}_i$ except the first coordinate[2]. Then with probability at least $1 - n^{-2}$, $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset$ for all $i, j \in [n]$.

Proof of Lemma 6.7.1. For any fixed $k \in [n]$ and $j \in \mathrm{supp}(\boldsymbol{\xi}_k)\backslash\{1\}$, by the model assumption

[2]Recall that all data inputs have nonzero first coordinate by Definition 6.2.1

we have

$$\mathbb{P}\{(\boldsymbol{\xi}_i)_j \neq 0\} = s/(d-1),$$

for all $i \in [n]\backslash\{k\}$. Therefore by the fact that the data samples are independent, we have

$$\mathbb{P}(\exists i \in [n]\backslash\{k\} : (\xi_i)_j \neq 0) = 1 - [1 - s/(d-1)]^n.$$

Applying a union bound over all $k \in [n]$ and $j \in \mathrm{supp}(\boldsymbol{\xi}_k)\backslash\{1\}$, we obtain

$$\mathbb{P}(\exists k \in [n], j \in \mathrm{supp}(\boldsymbol{\xi}_k)\backslash\{1\}, i \in [n]\backslash\{k\} : (\boldsymbol{\xi}_i)_j \neq 0) \leq n \cdot s \cdot \{1 - [1 - s/(d-1)]^n\}. \tag{6.7.1}$$

By the data distribution assumption we have $s \leq \sqrt{d}/(2n^2)$, which clearly implies $s/(d-1) \leq 1/2$. Therefore we have

$$\begin{aligned} n \cdot s \cdot [1 - (1 - s/d)^n] &= n \cdot s \cdot \{1 - \exp[n \log(1 - s/(d-1))]\} \\ &\leq n \cdot s \cdot [1 - \exp(n \cdot 2s/(d-1))] \\ &\leq n \cdot s \cdot [1 - \exp(n \cdot 4s/d)] \\ &\leq n \cdot s \cdot (4ns/d) \\ &= 4n^2 s^2/d \\ &\leq n^{-2}, \end{aligned}$$

where the first inequality follows by the inequalities $\log(1-z) \geq -2z$ for $z \in [0, 1/2]$, the second inequality follows by $s/(d-1) \geq 2s/d$, the third inequality follows by the inequality $1 - \exp(-z) \leq z$ for $z \in \mathbb{R}$, and the last inequality follows by the assumption that $s \leq \sqrt{d}/(2n^2)$. Plugging the bound above into (6.7.1) finishes the proof.

□

6.7.2 Proof for Adam

Before moving to the detailed proof, we first state the update rules of feature learning and noise memorization when the sign gradient is applied.

$$\begin{aligned}\langle \mathbf{w}_{j,r}^{(t+1)}, j\mathbf{v}\rangle &= \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle - \eta \cdot \big\langle \mathrm{sgn}\big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})\big), j\mathbf{v}\big\rangle \\ &= \langle \mathbf{w}_{j,r}^{(t)}, j\mathbf{v}\rangle + j\eta \cdot \mathrm{sgn}\bigg(\sum_{i=1}^{n} y_i \ell_{j,i}^{(t)} \big[\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{v}\rangle) - \alpha\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\big] - n\lambda \mathbf{w}_{j,r}^{(t)}[1]\bigg),\end{aligned} \tag{6.7.2}$$

where $\ell_{j,i}^{(t)} := \mathbb{1}_{y_i=j} - \mathrm{logit}_j(F, \mathbf{x}_i)$ and $\mathrm{logit}_j(F, \mathbf{x}_i) = \frac{e^{F_j(\mathbf{W}, \mathbf{x}_i)}}{\sum_{k\in\{-1,1\}} e^{F_k(\mathbf{W}, \mathbf{x}_i)}}$. From (6.7.2) we can observe three terms in the signed gradient. Specifically, the first term represents the gradient over the feature patch, the second term stems from the feature noise term in the noise patch (see Definition 6.2.1), and the last term is the gradient of the weight decay regularization. On the other hand, the memorization of the noise vector $\boldsymbol{\xi}_i$ can be described by the following update rule,

$$\begin{aligned}\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle &= \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle - \eta \cdot \big\langle \mathrm{sgn}\big(\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^{(t)})\big), \boldsymbol{\xi}_i\big\rangle \\ &= \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle + \eta \cdot \sum_{k\in\mathcal{B}_i} \bigg\langle \mathrm{sgn}\bigg(\ell_{y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\boldsymbol{\xi}_i[k] - n\lambda \mathbf{w}_{y_i,r}^{(t)}[k]\bigg), \boldsymbol{\xi}_i[k]\bigg\rangle \\ &\quad - \alpha y_i \eta \cdot \mathrm{sgn}\bigg(\sum_{i=1}^{n} y_i \ell_{y_i,i}^{(t)} \big[\sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\mathbf{v}\rangle) - \alpha\sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\big] - n\lambda \mathbf{w}_{y_i,r}^{(t)}[1]\bigg).\end{aligned} \tag{6.7.3}$$

In this subsection we first provide the following lemma that shows for most of the coordinate (with slightly large gradient), the Adam update is similar to signGD update (up to some constant factors). In the remaining proof for Adam, we will largely apply this lemma to get a signGD-like result for Adam (similar to the technical lemmas in Section 6.4). Besides, the proofs for all lemmas in Section 6.4 can be viewed as a simplified version of the proofs for technical lemmas for Adam, thus are omitted in the paper.

Lemma 6.7.2 (Closeness to SignGD)**.** Recall the update rule of Adam, let $\mathbf{W}^{(t)}$ be the t-th iterate of the Adam algorithm. Suppose that $\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle, \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle = \widetilde{\Theta}(1)$ for all $j \in \{\pm 1\}$ and $r \in [m]$. Then if $\beta_2 \geq \beta_1^2$, we have

- For all $k \in [d]$,
$$\left| \frac{\mathbf{m}_{j,r}^{(t)}[k]}{\sqrt{\mathbf{v}_{j,r}^{(t)}[k]}} \right| \leq \Theta(1).$$
- For every $k \notin \cup_{i=1}^n \mathcal{B}_i$ (including $k = 1$) we have either $|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]| \leq \widetilde{\Theta}(\eta)$ or
$$\frac{\mathbf{m}_{j,r}^{(t)}[k]}{\sqrt{\mathbf{v}_{j,r}^{(t)}[k]}} = \mathrm{sgn}\big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]\big) \cdot \Theta(1).$$
- For every $k \in \mathcal{B}_i$, we have $|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]| \leq \widetilde{\Theta}\big(\eta n^{-1} s\sigma_p |\ell_{j,i}^{(t)}|\big) \leq \widetilde{\Theta}(\eta s \sigma_p)$ or
$$\frac{\mathbf{m}_{j,r}^{(t)}[k]}{\sqrt{\mathbf{v}_{j,r}^{(t)}[k]}} = \mathrm{sgn}\big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]\big) \cdot \Theta(1).$$

Proof. First recall that the gradient $\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})$ can be calculated as

$$\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)}) = -\frac{1}{n}\bigg[\sum_{i=1}^n y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle) \cdot \mathbf{v} + \sum_{i=1}^n \ell_{j,i}^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \boldsymbol{\xi}_i \rangle) \cdot \boldsymbol{\xi}_i \bigg] + \lambda \mathbf{w}_{j,r}^{(t)}.$$

More specifically, for the first coordinate of $\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})$, we have

$$\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[1] = -\frac{1}{n}\bigg[\sum_{i=1}^n y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle) - \alpha \sum_{i=1}^n y_i \ell_{j,i}^{(t)} \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \bigg] + \lambda \mathbf{w}_{j,r}^{(t)}[1]. \tag{6.7.4}$$

For any $k \in \mathcal{B}_i$, by Lemma 6.7.1 we know that the gradient over this coordinate only depends on the training data $\boldsymbol{\xi}_i$, therefore, we have

$$\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k] = -\frac{1}{n} \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \boldsymbol{\xi}_i[k] + \lambda \mathbf{w}_{j,r}^{(t)}[k]. \tag{6.7.5}$$

For the remaining coordinates, we have

$$\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k] = \lambda \mathbf{w}_{j,r}^{(t)}[k]. \tag{6.7.6}$$

Now let us focus on the moving averaged gradient $\mathbf{m}_{j,r}^{(t)}$ and squared gradient $\mathbf{v}_{j,r}^{(t)}$. We first show that for all $k \in [d]$, it holds that

$$\frac{\big|\mathbf{m}_{j,r}^{(t)}[k]\big|}{\sqrt{\mathbf{v}_{j,r}^{(t)}[k]}} \leq \Theta(1). \tag{6.7.7}$$

By the update rule of $\mathbf{m}_{j,r}^{(t)}$, we have

$$\begin{aligned}\mathbf{m}_{j,r}^{(t)}[k] &= \beta_1 \mathbf{m}_{j,r}^{(t-1)}[k] + (1-\beta_1)\cdot \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k] \\ &= \sum_{\tau=0}^{t} \beta_1^\tau (1-\beta_1)\cdot \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k].\end{aligned}$$

Similarly, we also have

$$\mathbf{v}_{j,r}^{(t)}[k] = \sum_{\tau=0}^{t} \beta_2^\tau (1-\beta_2)\cdot \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k]^2.$$

Then by Cauchy-Schwartz inequality we have

$$\big(\mathbf{m}_{j,r}^{(t)}[k]\big)^2 \leq \bigg(\sum_{\tau=0}^{t} \frac{[\beta_1^\tau(1-\beta_1)]^2}{\alpha_\tau^2}\cdot \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k]^2\bigg)\cdot\bigg(\sum_{\tau=0}^{t}\alpha_\tau^2\bigg).$$

Let $\alpha_\tau^2 = \frac{[\beta_1^\tau(1-\beta_1)]^2}{\beta_2^\tau(1-\beta_2)}$, which forms an exponentially decaying sequence if $\beta_2 \geq \beta_1^2$. Therefore, we have $\sum_{\tau=0}^{t}\alpha_\tau^2 = \Theta(1)$ and the above inequality implies that

$$\big(\mathbf{m}_{j,r}^{(t)}[k]\big)^2 \leq \mathbf{v}_{j,r}^{(t)}[k]\cdot\Theta(1),$$

which proves (6.7.7).

Now we are going to prove the main argument of this lemma. Note that $\mathbf{m}_{j,r}^{(t)}$, which is a weighted average of all historical gradients, where the weights decay exponentially fast, then we can take on a threshold $\bar{\tau} = \mathrm{polylog}(\eta^{-1})$ such that $\sum_{\tau=\bar{\tau}}^{t}\beta_1^\tau(1-\beta_1) = \frac{1}{\mathrm{poly}(\eta^{-1})}$. Then for each $k \in [d]$ we have

$$\begin{aligned}\mathbf{m}_{j,r}^{(t)}[k] &= \sum_{\tau=0}^{\bar{\tau}} \beta_1^\tau(1-\beta_1)\cdot\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k] + \sum_{\tau=\bar{\tau}}^{t}\beta_1^\tau(1-\beta_1)\cdot\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k] \\ &= \sum_{\tau=0}^{\bar{\tau}} \beta_1^\tau(1-\beta_1)\cdot\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k] \pm \frac{1}{\mathrm{poly}(\eta^{-1})},\end{aligned}$$

where in the last inequality we use the fact that $|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k]| = \widetilde{O}(1)$ for all $k \in [d]$. Similarly, we can also have the following on $\mathbf{v}_{j,r}^{(t)}$,

$$\mathbf{v}_{j,r}^{(t)}[k] = \sum_{\tau=0}^{\bar{\tau}} \beta_2^\tau (1-\beta_2) \cdot \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k]^2 \pm \frac{1}{\mathrm{poly}(\eta^{-1})}.$$

Here we slightly abuse the notation by using the same $\bar{\tau}$. Then we have

$$\frac{\mathbf{m}_{j,r}^{(t)}[k]}{\sqrt{\mathbf{v}_{j,r}^{(t)}[k]}} = \frac{\sum_{\tau=0}^{\bar{\tau}} \beta_1^\tau (1-\beta_1) \cdot \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k] \pm \frac{1}{\mathrm{poly}(\eta^{-1})}}{\sqrt{\sum_{\tau=\bar{\tau}}^{\bar{\tau}} \beta_2^\tau (1-\beta_2) \cdot \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t-\tau)})[k]^2} \pm \frac{1}{\mathrm{poly}(\eta^{-1})}}.$$

In order to prove the main argument of this lemma, the key is to show that within $\bar{\tau}$ iterations, the gradient $\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]$ barely changes. In particular, by (6.7.7), we have the update of each coordinate in one step is at most $\Theta(\eta)$. This implies that

$$\begin{aligned}
\big|\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle - \langle \mathbf{w}_{j,r}^{(\tau)}, \mathbf{v}\rangle\big| &\le \Theta(\eta\bar{\tau}), \\
\big|\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle - \langle \mathbf{w}_{j,r}^{(\tau)}, \boldsymbol{\xi}_i\rangle\big| &\le \Theta(\eta\bar{\tau} s\sigma_p), \\
|\mathbf{w}_{j,r}^{(t)}[k] - \mathbf{w}_{j,r}^{(\tau)}[k]| &\le \Theta(\eta\bar{\tau}).
\end{aligned}$$

Then applying the fact that $|\langle \mathbf{w}_{j,r}^{(\tau)}, \mathbf{v}\rangle| \le \widetilde{\Theta}(1)$ and $|\langle \mathbf{w}_{j,r}^{(\tau)}, \boldsymbol{\xi}_i\rangle| \le \widetilde{\Theta}(1)$, we further have

$$\big|F_j(\mathbf{W}^{(\tau)}, \mathbf{x}_i) - F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)\big| \le \Theta(m\eta\bar{\tau} s\sigma_p) = \widetilde{\Theta}(\eta\bar{\tau} s\sigma_p),$$

where we use the fact that $m = \widetilde{\Theta}(1)$ and $s\sigma_p = \omega(1)$. Then it holds that

$$\begin{aligned}
\ell_{j,i}^{(\tau)} &= \frac{e^{F_j(\mathbf{W}^{(\tau)}, \mathbf{x}_i)}}{\sum_{k\in\{-1,1\}} e^{F_k(\mathbf{W}^{(\tau)}, \mathbf{x}_i)}} \\
&\le \frac{e^{F_j(\mathbf{W}^{(t)}, \mathbf{x}_i) + \widetilde{\Theta}(\eta\bar{\tau} s\sigma_p)}}{e^{F_j(\mathbf{W}^{(\tau)}, \mathbf{x}_i) + \widetilde{\Theta}(\eta\bar{\tau} s\sigma_p)} + e^{F_{-j}(\mathbf{W}^{(t)}, \mathbf{x}_i) - \widetilde{\Theta}(\eta\bar{\tau} s\sigma_p)}} \\
&= \mathrm{sgn}(\ell_{j,i}^{(t)}) \cdot \Theta(|\ell_{j,i}^{(t)}|),
\end{aligned}$$

where we use the fact that $\widetilde{\Theta}(\eta\bar{\tau} s\sigma_p) = o(1)$. Similarly, we can also show that $\ell_{j,i}^{(\tau)} \ge \mathrm{sgn}(\ell_{j,i}^{(t)}) \cdot \Theta(|\ell_{j,i}^{(t)}|)$, which further implies

$$\ell_{j,i}^{(\tau)} = \mathrm{sgn}(\ell_{j,i}^{(t)}) \cdot \Theta(|\ell_{j,i}^{(t)}|)$$

for all $\tau \in [t-\bar{\tau}, t]$. Note that $|\ell_{j,i}^{(\tau)}| \leq 1$, then it holds that

$$\begin{aligned}\ell_{j,i}^{(\tau)}\sigma'(\langle \mathbf{w}_{j,r}^{(\tau)}, \mathbf{v}\rangle) &= \mathrm{sgn}(\ell_{j,i}^{(t)}) \cdot \Theta(|\ell_{j,i}^{(t)}|) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(\tau)}, \mathbf{v}\rangle) \\ &\leq \mathrm{sgn}(\ell_{j,i}^{(t)}) \cdot \Theta(|\ell_{j,i}^{(t)}|) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle) + \Theta(|\ell_{j,i}^{(t)}|) \cdot \widetilde{\Theta}(\eta\bar{\tau}).\end{aligned}$$

We can also similarly derive the following

$$\begin{aligned}\ell_{j,i}^{(\tau)}\sigma'(\langle \mathbf{w}_{j,r}^{(\tau)}, \mathbf{v}\rangle) &\geq \mathrm{sgn}(\ell_{j,i}^{(t)}) \cdot \Theta(|\ell_{j,i}^{(t)}|) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle) - \Theta(|\ell_{j,i}^{(t)}|) \cdot \widetilde{\Theta}(\eta\bar{\tau}), \\ \ell_{j,i}^{(\tau)}\sigma'(\langle \mathbf{w}_{j,r}^{(\tau)}, \boldsymbol{\xi}_i\rangle) &\leq \mathrm{sgn}(\ell_{j,i}^{(t)}) \cdot \Theta(|\ell_{j,i}^{(t)}|) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) + \Theta(|\ell_{j,i}^{(t)}|) \cdot \widetilde{\Theta}(\eta\bar{\tau}s\sigma_p), \\ \ell_{j,i}^{(\tau)}\sigma'(\langle \mathbf{w}_{j,r}^{(\tau)}, \boldsymbol{\xi}_i\rangle) &\geq \mathrm{sgn}(\ell_{j,i}^{(t)}) \cdot \Theta(|\ell_{j,i}^{(t)}|) \cdot \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) - \Theta(|\ell_{j,i}^{(t)}|) \cdot \widetilde{\Theta}(\eta\bar{\tau}s\sigma_p).\end{aligned}$$

Combining the above results, applying (6.7.4), (6.7.5), and (6.7.6), we can show that for the first coordinate, we have

$$\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(\tau)})[1] = \Theta\Big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[1]\Big) \pm \Theta\bigg(\frac{1}{n}\sum_{i=1}^{n} |\ell_{j,i}^{(t)}|\bigg) \cdot \widetilde{O}(\eta\bar{\tau}) \pm \Theta(\lambda\eta\bar{\tau});$$

for any $k \in \mathcal{B}_j$, we have

$$\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(\tau)})[k] = \Theta\Big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]\Big) \pm \Theta\bigg(\frac{|\ell_{j,i}^{(t)}|}{n}\bigg) \cdot \widetilde{O}(\eta\bar{\tau}s\sigma_p) \pm \Theta(\lambda\eta\bar{\tau});$$

and for remaining coordinates, we have

$$\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(\tau)})[k] = \Theta\Big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]\Big) \pm \Theta(\lambda\eta\widetilde{\tau}).$$

Now we can plug the above results into the formula of $\mathbf{m}_{j,r}^{(t)}$ and $\mathbf{v}_{j,r}^{(t)}$. Using the fact that $\bar{\tau} = \widetilde{\Theta}(1)$, $\lambda = o(1)$, and $|\ell_{j,i}^{(t)}| \leq 1$, we have for all $k = 1$ or $k \notin \mathcal{B}_i$ for any i,

$$\frac{\mathbf{m}_{j,r}^{(t)}[k]}{\sqrt{\mathbf{v}_{j,r}^{(t)}[k]}} = \frac{\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k] \pm \widetilde{\Theta}(\eta)}{\Theta\Big(|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]|\Big) \pm \widetilde{\Theta}(\eta))}.$$

For $k \in \mathcal{B}_i$ we have

$$\frac{\mathbf{m}_{j,r}^{(t)}[k]}{\sqrt{\mathbf{v}_{j,r}^{(t)}[k]}} = \frac{\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k] \pm \widetilde{\Theta}\Big(\frac{\eta s\sigma_p|\ell_{j,i}^{(t)}|}{n}\Big) \pm \widetilde{\Theta}(\lambda\eta)}{\Theta\Big(|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]|\Big) \pm \widetilde{\Theta}\Big(\frac{\eta s\sigma_p|\ell_{j,i}^{(t)}|}{n}\Big) \pm \widetilde{\Theta}(\lambda\eta)}.$$

Then, we can conclude that for all $k = 1$ or $k \notin \mathcal{B}_i$ for any i, we have either $|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]| \leq \widetilde{\Theta}(\eta)$ or

$$\frac{\mathbf{m}_{j,r}^{(t)}[k]}{\sqrt{\mathbf{v}_{j,r}^{(t)}[k]}} = \operatorname{sgn}\big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]\big) \cdot \Theta(1).$$

For any $k \in \mathcal{B}_i$, we have either $|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]| \leq \widetilde{\Theta}\big(\eta n^{-1} s\sigma_p |\ell_{j,i}^{(t)}| + \lambda\eta\big)$ or

$$\frac{\mathbf{m}_{j,r}^{(t)}[k]}{\sqrt{\mathbf{v}_{j,r}^{(t)}[k]}} = \operatorname{sgn}\big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[k]\big) \cdot \Theta(1).$$

This completes the proof.

□

Lemma 6.7.3 (Lemma 6.4.2, restated)**.** Suppose the training data is generated according to Definition 6.2.1, assume $\lambda = o(\sigma_0^{q-2}\sigma_p/n)$ and $\eta = 1/\mathrm{poly}(d)$, then for any $t \leq T_0$ with $T_0 = \widetilde{O}\big(\frac{1}{\eta s \sigma_p}\big)$ and any $i \in [n]$,

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t+1)}, j \cdot \mathbf{v} \rangle &\leq \langle \mathbf{w}_{j,r}^{(t)}, j \cdot \mathbf{v} \rangle + \Theta(\eta), \\
\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle &= \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle + \widetilde{\Theta}(\eta s \sigma_p).
\end{aligned}$$

Proof. At the initialization, we have

$$|\langle \mathbf{w}_{j,r}^{(0)}, \mathbf{v} \rangle| = \widetilde{\Theta}(\sigma_0), \quad |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle| = \widetilde{\Theta}(s^{1/2}\sigma_p\sigma_0 + \alpha) = \widetilde{\Theta}(s^{1/2}\sigma_p\sigma_0), \quad \mathbf{w}_{j,r}^{(0)}[k] = \widetilde{\Theta}(\sigma_0),$$

which also imply that $|\ell_{j,i}^{(0)}| = \Theta(1)$. Besides, note that $\ell_{j,i}^{(t)} = \mathbb{1}_{j=y_i} - \mathrm{logit}_j(F^{(t)}, \mathbf{x}_i)$, we have

$$\operatorname{sgn}\big(y_i \ell_{j,i}^{(t)}\big) = \operatorname{sgn}(j),$$

where we recall that $j \in \{-1, 1\}$. Therefore, given that $\lambda = o(\sigma_0^{q-1})$, $\alpha = o(1)$, $s^{1/2}\sigma_p = \widetilde{O}(1)$, and assume $\ell_{j,i}^{(t)} = \Theta(1)$ (which will be verified later),

$$\begin{aligned}
&\operatorname{sgn}\bigg(\sum_{i=1}^n y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle) - \alpha \sum_{i=1}^n y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) - n\lambda \mathbf{w}_{j,r}^{(t)}[1] \bigg) \\
&= \operatorname{sgn}\big[j \cdot \widetilde{\Theta}(n\sigma_0^{q-1}) - j \cdot \widetilde{\Theta}(\alpha n (s^{1/2}\sigma_p\sigma_0)^{q-1}) \pm o\big(\sigma_0^{q-1}\sigma_p\big) \big] \\
&= \operatorname{sgn}(j).
\end{aligned}$$

Since $\mathbf{v}$ is 1-sparse, then by Lemma 6.7.2, the following inequality naturally holds,

$$\langle \mathbf{w}_{j,r}^{(t+1)}, j\cdot\mathbf{v}\rangle \le \langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle - \eta\Big\langle \mathbf{m}_{j,r}^{(t)}/\sqrt{\mathbf{v}_{j,r}^{(t)}}, j\cdot\mathbf{v}\Big\rangle \le \langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle + \Theta(\eta).$$

Additionally, in terms of the memorization of noise, we first consider the iterate in the initialization. By the condition that $\eta = o(1/d) = o(1/(s\sigma_p))$ and note that for a sufficiently large fraction of $k \in \mathcal{B}_i$ (e.g., 0.99), we have $|\boldsymbol{\xi}_i[k]| \ge \widetilde{\Theta}(\sigma_p) \ge \widetilde{\Theta}(\eta n^{-1} s\sigma_p |\ell_{j,i}^{(0)}|)$ and thus

$$\begin{aligned}\operatorname{sgn}\big(\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^{(0)})[k]\big) &= \operatorname{sgn}\bigg(\ell_{y_i,i}^{(t)}\sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\boldsymbol{\xi}_i[k] - n\lambda \mathbf{w}_{y_i,r}^{(0)}[k]\bigg)\\ &= -\operatorname{sgn}\big[\widetilde{\Theta}\big((d^{1/2}\sigma_p\sigma_0)^{q-1}\sigma_p\cdot\operatorname{sgn}(\boldsymbol{\xi}_i[k])\big) \pm o(\sigma_0^{q-1}\sigma_p)\big] = -\operatorname{sgn}(\boldsymbol{\xi}_i[k]).\end{aligned} \tag{6.7.8}$$

Therefore, by Lemma 6.7.2 we have the following according to (6.7.3),

$$\begin{aligned}\langle \mathbf{w}_{y_i,r}^{(1)}, \boldsymbol{\xi}_i\rangle &= \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \eta\Big\langle \mathbf{m}_{j,r}^{(t)}/\sqrt{\mathbf{v}_{y_i,r}^{(t)}}, \boldsymbol{\xi}_i\Big\rangle\\ &\ge \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \Theta(\eta)\cdot\sum_{k\in\mathcal{B}_i}\langle \operatorname{sgn}(\boldsymbol{\xi}_i[k]), \boldsymbol{\xi}_i[k]\rangle - O(\eta s\sigma_p) - O(\eta\alpha)\\ &= \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \widetilde{\Theta}(\eta s\sigma_p),\end{aligned}$$

where in the first inequality the term $O(\eta s\sigma_p)$ represents the coordinates that $|\boldsymbol{\xi}_i[k]| \le O(\sigma_p)$ (so that we cannot use the sign information of $\nabla_{y_i,r}L(\mathbf{W}^{(0)})$ but directly bound it by $\Theta(1)$) and the last inequality is due to the fact that $|\mathcal{B}_i| \ge s-1$ and $\alpha = o(1)$. For general t, we will consider the following induction hypothesis:

$$\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle = \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle + \widetilde{\Theta}(\eta s\sigma_p), \tag{6.7.9}$$

which has already been verified for $t = 0$. By Hypothesis (6.7.9), the following holds at time t,

$$\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle = \langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \widetilde{\Theta}(t\eta s\sigma_p) = \widetilde{\Theta}(s^{1/2}\sigma_p\sigma_0 + t\eta s\sigma_p).$$

In the meanwhile, we have the following upper bound for $|\mathbf{w}_{j,r}^{(t)}[k]|$,

$$|\mathbf{w}_{j,r}^{(t)}[k]| \le |\mathbf{w}_{j,r}^{(t)}[k]| + \eta|\operatorname{sign}(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)}))| \le |\mathbf{w}_{j,r}^{(0)}[k]| + t\eta = \widetilde{\Theta}(\sigma_0 + t\eta). \tag{6.7.10}$$

Besides, it is also easy to verify that for any $t \leq T_0 = \widetilde{\Theta}\big(\frac{1}{s\sigma_p\eta m}\big) = \widetilde{\Theta}\big(\frac{1}{s\sigma_p\eta}\big)$, we have $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle, \langle \mathbf{w}_{y_i,r}^{(t)}, j\cdot\mathbf{v}\rangle < \Theta(1/m)$ and thus $|\ell_{j,i}^{(t)}| = \Theta(1)$. Then similar to (6.7.8), we have

$$
\begin{aligned}
&\mathrm{sgn}\big(\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^{(t)})[k]\big) \\
&= \mathrm{sgn}\bigg(\ell_{y_i,i}^{(t)}\sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\boldsymbol{\xi}_i[k] - n\lambda \mathbf{w}_{y_i,r}^{(0)}[k]\bigg) \\
&= -\mathrm{sgn}\big(\widetilde{\Theta}\big[(s^{1/2}\sigma_p\sigma_0 + t\eta s\sigma_p)^{q-1}\sigma_p\cdot\mathrm{sgn}(\boldsymbol{\xi}_i[k])\big) \pm o\big(\sigma_0^{q-2}\sigma_p\cdot(\sigma_0+t\eta)\big)\big] \\
&= -\mathrm{sgn}(\boldsymbol{\xi}_i[k]). \qquad (6.7.11)
\end{aligned}
$$

This further implies that

$$
\begin{aligned}
\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle &\geq \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle - \Theta(\eta)\cdot\sum_{k\in\mathcal{B}_i}\langle \mathrm{sgn}\big(\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^{(t)})[k]\big), \boldsymbol{\xi}_i[k]\rangle - O(\eta^2 s^2\sigma_p^2) - O(\eta\alpha) \\
&= \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle + \widetilde{\Theta}(\eta s\sigma_p),
\end{aligned}
$$

where the term $-O(\eta^2 s^2\sigma_p^2)$ is contributed by the gradient coordinates that are smaller than $\Theta(\eta s\sigma_p)$. This verifies Hypothesis (6.7.9) at time t and thus completes the proof. □

From Lemma 6.7.3, note that $s\sigma_p = \omega(1)$, then it can be seen that $\langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle$ increases much faster than $\langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle$. By looking at the update rule of $\langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle$ (see (6.7.2)), it will keeps increasing only when, roughly speaking, $\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle) > \alpha\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)$. Since $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle$ increases much faster than $\langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle$, it can be anticipated after a certain number of iterations, $\langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle$ will start to decrease. In the following lemma, we provide an upper bound on the iteration number such that this decreasing occurs.

Lemma 6.7.4 (Lemma 6.7.4, restated)**.** Suppose the training data is generated according to Definition 6.2.1, $\alpha \geq \widetilde{\Theta}\big((s\sigma_p)^{1-q} \vee \sigma_0^{q-1}\big)$ and $\sigma_0 < \widetilde{O}((s\sigma_p)^{-1})$, then for any $t \in [T_r, T_0]$ with $T_r = \widetilde{O}\big(\frac{\sigma_0}{\eta s\sigma_p\alpha^{1/(q-1)}}\big) \leq T_0$,

$$
\langle \mathbf{w}_{j,r}^{(t+1)}, j\cdot\mathbf{v}\rangle = \langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle - \Theta(\eta).
$$

Moreover, it holds that

$$\mathbf{w}_{j,r}^{(T_0)}[k] = \begin{cases} -\mathrm{sgn}(j)\cdot\widetilde{\Omega}\big(\frac{1}{s\sigma_p}\big), & k=1, \\ \mathrm{sgn}(\boldsymbol{\xi}_i[k])\cdot\widetilde{\Omega}\big(\frac{1}{s\sigma_p}\big) \text{ or } \pm\widetilde{O}(\eta), & k\in\mathcal{B}_i, \text{ with } y_i=j, \\ \pm\widetilde{O}(\eta), & \text{otherwise.} \end{cases}$$

Proof. Recall from Lemma 6.7.3 that for any $t \le T_0$ we have

$$\begin{aligned} \langle \mathbf{w}_{j,r}^{(t+1)}, j\cdot\mathbf{v}\rangle &\le \langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle + \Theta(\eta) \le \langle \mathbf{w}_{j,r}^{(0)}, j\cdot\mathbf{v}\rangle + \Theta(t\eta), \\ \langle \mathbf{w}_{y_s,r}^{(t+1)}, \boldsymbol{\xi}_s\rangle &= \langle \mathbf{w}_{y_s,r}^{(t)}, \boldsymbol{\xi}_s\rangle + \widetilde{\Theta}(\eta s\sigma_p) \le \langle \mathbf{w}_{y_s,r}^{(0)}, \boldsymbol{\xi}_s\rangle + \widetilde{\Theta}(t\eta s\sigma_p). \end{aligned}$$

Besides, by Lemma 6.7.2 we also have $|\mathbf{w}_{j,r}^{(t)}[k]| \le |\mathbf{w}_{j,r}^{(0)}[k]| + O(t\eta)$. Then it can be verified that for some $T_r = \widetilde{O}\big(\frac{\sigma_0}{\eta s\sigma_p \alpha^{1/(q-1)}}\big)$, we have for all $i\in[n]$ and $t\in[T_r, T_0]$

$$\alpha\sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle) \ge C\cdot\big[\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle) + \lambda n|\mathbf{w}_{j,r}^{(t)}[1]|\big]$$

for some constant C. This further implies that

$$\begin{aligned} &\mathrm{sgn}\big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[1]\big) \\ &= -\mathrm{sgn}\bigg(\sum_{i=1}^n y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{v}\rangle) - \alpha\sum_{i=1}^n y_i\ell_{j,i}^{(t)}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) - n\lambda\mathbf{w}_{j,r}^{(t)}[1]\bigg) \\ &= -\mathrm{sgn}\big[-\alpha\sum_{i=1}^n y_i\ell_{j,i}^{(t)}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\big] \\ &= \mathrm{sgn}(j), \end{aligned}$$

where we use the fact that $\mathrm{sgn}(y_i\ell_{j,i}^{(t)}) = \mathrm{sgn}(j)$ for all $i\in[n]$. Then by Lemma 6.7.2 and (6.7.2), we have for all $t\in[T_r, T_0]$,

$$\langle \mathbf{w}_{j,r}^{(t+1)}, j\cdot\mathbf{v}\rangle = \langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle - \Theta(\eta)\cdot\mathrm{sgn}(j)\cdot\mathrm{sgn}\big(\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[1]\big) = \langle \mathbf{w}_{j,r}^{(t)}, j\cdot\mathbf{v}\rangle - \Theta(\eta).$$

Then at iteration T_0, for the first coordinate we have

$$\mathbf{w}_{j,r}^{(T_0)}[1] = \mathbf{w}_{j,r}^{(0)}[1] + \mathrm{sgn}(j)\cdot\Theta(T_r\eta) - \mathrm{sgn}(j)\cdot\Theta((T_0-T_r)\eta) \ge -\mathrm{sgn}(j)\cdot\widetilde{\Omega}\bigg(\frac{1}{s\sigma_p}\bigg)$$

For any $k \in \mathcal{B}_i$ with $y_i = j$, we have either the coordinate will increase at a rate of $\Theta(1)$ or fall into 0. As a consequence we have either $\mathbf{w}_{j,r}^{(T_0)}[k] \in [-\widetilde{\Theta}(\eta), \widetilde{\Theta}(\eta)]$ or

$$\mathbf{w}_{j,r}^{(T_0)}[k] = \mathbf{w}_{j,r}^{(0)}[k] + \mathrm{sgn}(\boldsymbol{\xi}_i[k]) \cdot \Theta(T_0\eta) \geq \mathrm{sgn}(\boldsymbol{\xi}_i[k]) \cdot \widetilde{\Omega}\bigg(\frac{1}{s\sigma_p}\bigg).$$

For the remaining coordinate, its update will be determined by the regularization term, which will finally fall into the region around zero since we have $T_0\eta = \omega(\sigma_0)$. By Lemma 6.7.2 it is clear that $\mathbf{w}_{j,r}^{(T_0)}[k] \in [-\widetilde{\Theta}(\eta), \widetilde{\Theta}(\eta)]$. □

Lemma 6.7.5 (Lemma 6.4.4, restated)**.** If $\alpha = O\big(\frac{s\sigma_p^2}{n}\big)$ and $\eta = o(\lambda)$, then let $r^* = \arg\max_{r\in[m]} \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$, for any $t \geq T_0$, $i \in [n]$, $j \in [2]$ and $r \in [m]$, it holds that

$$\langle \mathbf{w}_{y_i,r^*}^{(t)}, \boldsymbol{\xi}_i \rangle = \widetilde{\Theta}(1), \quad \sum_{k\in\mathcal{B}_i} |\mathbf{w}_{y_i,r^*}^{(t)}[k]| \cdot |\boldsymbol{\xi}_i[k]| = \widetilde{\Theta}(1),$$
$$\forall r \in [m], \quad \langle \mathbf{w}_{j,r}^{(t)}, \mathrm{sgn}(j) \cdot \mathbf{v} \rangle \in [-\widetilde{O}\big(\frac{n\alpha}{s\sigma_p^2}\big), O(\lambda^{-1}\eta)].$$

Proof. The proof will be relying on the following three induction hypothesis:

$$\langle \mathbf{w}_{y_i,r^*}^{(t)}, \boldsymbol{\xi}_i \rangle = \widetilde{\Omega}(1), \tag{6.7.12}$$
$$\sum_{k\in\mathcal{B}_i} |\mathbf{w}_{y_i,r^*}^{(t+1)}[k]| \cdot |\boldsymbol{\xi}_i[k]| = \widetilde{\Theta}(1), \tag{6.7.13}$$
$$\forall r \in [m], \ \langle \mathbf{w}_{j,r}^{(t)}, \mathrm{sgn}(j) \cdot \mathbf{v} \rangle \in \Big[-\widetilde{O}\big(\frac{n\alpha}{s\sigma_p^2}\big), O(\lambda^{-1}\eta)\Big], \tag{6.7.14}$$

which we assume they hold for all $\tau \leq t$ and $r \in [m]$, $i \in [n]$, and $j \in [2]$. It is clear that all hypothesis hold when $t = T_0$ according to Lemma 6.7.4.

Verifying Hypothesis (6.7.12)**.** We first verify Hypothesis (6.7.12). Recall that the update rule for $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ is given as follows,

$$
\begin{aligned}
&\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle \\
&= \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle - \eta \cdot \big\langle \mathbf{m}_{y_i,r}^{(t)} / \sqrt{\mathbf{v}_{y_i,r}^{(t)}}, \boldsymbol{\xi}_i \big\rangle \\
&\geq \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle - \Theta(\eta) \cdot \big\langle \mathrm{sgn}\big(\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^{(t)})\big), \boldsymbol{\xi}_i \big\rangle - \widetilde{\Theta}(\eta^2 s^2 \sigma_p^2) \\
&= \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle + \Theta(\eta) \cdot \sum_{k \in \mathcal{B}_i} \bigg\langle \mathrm{sgn}\bigg(\ell_{y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \boldsymbol{\xi}_i[k] - n\lambda \mathbf{w}_{y_i,r}^{(t)}[k] \bigg), \boldsymbol{\xi}_i[k] \bigg\rangle \\
&\quad - \alpha y_i \Theta(\eta) \cdot \mathrm{sgn}\bigg(\sum_{i=1}^n y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle) - \alpha \sum_{i=1}^n y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) - n\lambda \mathbf{w}_{j,r}^{(t)}[1] \bigg) \\
&\quad - \widetilde{\Theta}(\eta^2 s^2 \sigma_p^2). \qquad (6.7.15)
\end{aligned}
$$

Note that for any a and b we have $\mathrm{sgn}(a-b) \cdot a \geq |a| - 2|b|$. Then it follows that

$$
\begin{aligned}
\sum_{k \in \mathcal{B}_i} \bigg\langle \mathrm{sgn}\bigg(\ell_{y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \boldsymbol{\xi}_i[k] - n\lambda \mathbf{w}_{y_i,r}^{(t)}[k] \bigg), \boldsymbol{\xi}_i[k] \bigg\rangle &\geq \sum_{k \in \mathcal{B}_i} \bigg(|\boldsymbol{\xi}_i[k]| - \frac{2n\lambda |\mathbf{w}_{y_i,r}^{(t)}[k]|}{\ell_{y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{y_i}^{(t)}, \boldsymbol{\xi}_i \rangle)} \bigg) \\
&\geq \widetilde{\Theta}(s\sigma_p) - \widetilde{\Theta}\bigg(\frac{n\lambda}{\ell_{y_i,i}^{(t)} \sigma_p} \bigg),
\end{aligned}
$$

where the last inequality follows from Hypothesis (6.7.12) and (6.7.13). Further recall that $\lambda = o(\sigma_0^{q-2} \sigma_p / n)$, plugging the above inequality to (6.7.15) gives

$$
\begin{aligned}
\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle &\geq \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle + \widetilde{\Theta}(\eta s \sigma_p) - \widetilde{\Theta}\bigg(\frac{\eta n \lambda}{\ell_{y_i,i}^{(t)} \sigma_p} \bigg) - \widetilde{\Theta}(\eta^2 s^2 \sigma_p^2) \\
&\geq \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle + \widetilde{\Theta}(\eta s \sigma_p) - \Theta(\alpha \eta) - \widetilde{\Theta}\bigg(\frac{\eta \sigma_0^{q-2}}{\ell_{y_i,i}^{(t)}} \bigg). \qquad (6.7.16)
\end{aligned}
$$

Then it is clear that $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$ will increase by $\widetilde{\Theta}(\eta s \sigma_p)$ if $\ell_{y_i,i}^{(t)}$ is larger than some constant of order $\widetilde{\Omega}(\frac{n\lambda}{s\sigma_p^2}) = \widetilde{\Omega}(\frac{\sigma_0^{q-2}}{s\sigma_p})$. We will first show that as soon as there is a iterate $\mathbf{W}^{(\tau)}$ satisfying $\ell_{y_i,i}^{(\tau)} \leq \widetilde{O}\big(\frac{n\lambda}{s\sigma_p^2}\big)$ for some $\tau \leq t$, then it must hold that $\ell_{y_i,i}^{(\tau')}$ will also be smaller than some constant in the order of $\widetilde{O}\big(\frac{n\lambda}{s\sigma_p^2}\big)$ for all $\tau' \in [\tau, t+1]$. To prove this, we first note that if $\ell_{y_i,i}^{(t)}$

reaches some constant in the order of $\widetilde{O}\big(\frac{n\lambda}{s\sigma_p^2}\big)$, we have for all $r \in [m]$ by (6.7.16)

$$\begin{aligned}
\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle &\geq \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle + \widetilde{\Theta}(\eta s \sigma_p), \\
\langle \mathbf{w}_{-y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle &\leq \langle \mathbf{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle + O(\alpha\eta), \\
|\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{v} \rangle| &\leq |\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle| + O(\eta).
\end{aligned} \tag{6.7.17}$$

Therefore, we have

$$\begin{aligned}
\ell_{y_i,i}^{(t+1)} &= \frac{e^{F_{-y_i}(\mathbf{W}^{(t+1)}, \mathbf{x}_i)}}{\sum_{j \in \{-1,1\}} e^{F_j(\mathbf{W}^{(t+1)}, \mathbf{x}_i)}} \\
&= \frac{1}{1 + \exp\big[\sum_{r=1}^m \big[\sigma(\langle \mathbf{w}_{y_i,r}^{(t+1)}, \mathbf{v} \rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle) - \sigma(\langle \mathbf{w}_{-y_i,r}^{(t+1)}, \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{-y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle)\big]\big]} \\
&\leq \frac{1}{1 + \exp\big[\sum_{r=1}^m \big[\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \mathbf{v} \rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) - \sigma(\langle \mathbf{w}_{-y_i,r}^{(t)}, \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle)\big] + \widetilde{\Theta}(\eta s \sigma_p^2)\big]} \\
&\leq \frac{1}{1 + \exp\big[\sum_{r=1}^m \big[\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \mathbf{v} \rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle)\big] - \sigma(\langle \mathbf{w}_{-y_i,r}^{(t)}, \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle)\big]\big]} \\
&= \ell_{y_i,i}^{(t)},
\end{aligned}$$

where inequality follows from (6.7.17). Therefore, this implies that as long as $\ell_{y_i,i}^{(t)}$ is larger than some constant $b = \widetilde{O}\big(\frac{n\lambda}{s\sigma_p^2}\big)$, then the adam algorithm will prevent it from further increasing. Besides, since $m\eta\sigma_p^2 = o(1)$, then we must have $\ell_{y_i,i}^{(t+1)} \in [0.5\ell_{y_i,i}^{(t)}, 2\ell_{y_i,i}^{(t)}]$. As a consequence, we can deduce that $\ell_{y_i,i}^{(t)}$ cannot be larger than $2b$, since otherwise there must exists a iterate $\mathbf{W}^{(\tau)}$ with $\tau \leq t$ such that $\ell_{y_i,i}^{(\tau)} \in [b, 2b]$ and $\ell_{y_i,i}^{(\tau+1)} \geq \ell_{y_i,i}^{(\tau)}$, which contradicts the fact that $\ell_{y_i,i}^{(\tau)}$ should decreases if $\ell_{y_i,i}^{(\tau)} \geq b$. Therefore, we can claim that if $\ell_{y_i,i}^{(\tau)} \leq b = \widetilde{O}\big(\frac{n\lambda}{s\sigma_p^2}\big)$ for some $\tau \leq t$, then we have

$$\ell_{y_i,i}^{(\tau')} \leq \widetilde{O}\bigg(\frac{n\lambda}{s\sigma_p^2}\bigg) \tag{6.7.18}$$

for all $\tau' \in [\tau, t+1]$. Then further note that

$$2\ell_{y_i,i}^{(t+1)} \geq \ell_{y_i,i}^{(t)} = \frac{e^{F_{-y_i}(\mathbf{W}^{(t)},\mathbf{x}_i)}}{\sum_{j\in\{-1,1\}} e^{F_j(\mathbf{W}^{(t)},\mathbf{x}_i)}}$$
$$\geq \exp\bigg(-\sum_{r=1}^{m} \big[\sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, y_i\mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\big]\bigg)$$
$$\geq \exp\big(-\Theta\big(m \max_{r\in[m]} \sigma(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\big)\big), \tag{6.7.19}$$

where in the last inequality we use Hypothesis (6.7.14). Then by the fact that $\ell_{y_i,i}^{(t+1)} \leq \widetilde{O}\big(\frac{n\lambda}{s\sigma_p^2}\big) = o(1)$ and $m = \widetilde{\Theta}(1)$, it is clear that $\exp\big(-\Theta\big(m \max_{r\in[m]} \sigma(\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle)\big)\big) = o(1)$ so that $\max_{r\in[m]}\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle = \widetilde{\Omega}(1)$. This verifies Hypothesis (6.7.12).

Verifying Hypothesis (6.7.13)**.** Now we will verify Hypothesis (6.7.13). First, note that we have already shown that $\langle \mathbf{w}_{y_i,r^*}^{(t+1)}, \boldsymbol{\xi}_i\rangle = \widetilde{\Omega}(1)$ so it holds that

$$\sum_{k\in\mathcal{B}_i} |\mathbf{w}_{y_i,r^*}^{(t+1)}[k]| \cdot |\boldsymbol{\xi}_i[k]| + \alpha|\mathbf{w}_{y_i,r^*}^{(t+1)}[1]| \geq \langle \mathbf{w}_{y_i,r^*}^{(t+1)}, \boldsymbol{\xi}_i\rangle = \widetilde{\Omega}(1).$$

By Hypothesis (6.7.14), we have $|\mathbf{w}_{y_i,r^*}^{(t+1)}[1]| \leq |\mathbf{w}_{y_i,r^*}^{(t)}[1]| + \eta = o(1)$. Besides, since each coordinate in $\boldsymbol{\xi}_i$ is a Gaussian random variable, then $\max_{k\in\mathcal{B}_i} |\boldsymbol{\xi}_i[k]| = \widetilde{O}(\sigma_p)$. This immediately implies that

$$\sum_{k\in\mathcal{B}_i} |\mathbf{w}_{y_i,r^*}^{(t+1)}[k]| \cdot |\boldsymbol{\xi}_i[k]| = \widetilde{\Omega}(1).$$

Then we will prove the upper bound of $\sum_{k\in\mathcal{B}_i} |\mathbf{w}_{y_i,r}^{(t+1)}[k]| \cdot |\boldsymbol{\xi}_i[k]|$. Recall that by Lemma 6.7.2, for any $k \in \mathcal{B}_i$ such that $\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^{(t)})[k] \geq \widetilde{\Theta}(n^{-1}\eta s\sigma_p \ell_{y_i,i}^{(t)})$, we have

$$\mathbf{w}_{y_i,r}^{(t+1)}[k] = \mathbf{w}_{y_i,r}^{(t)}[k] + \Theta(\eta) \cdot \mathrm{sgn}\bigg(\ell_{y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\boldsymbol{\xi}_i[k] - n\lambda \mathbf{w}_{y_i,r}^{(t)}[k]\bigg).$$

Note that by Lemma 6.7.4, for every $k \in \mathcal{B}_i$, we have either $\mathbf{w}_{y_i,r}^{(T_0)}[k] = \mathrm{sgn}(\boldsymbol{\xi}_i[k]) \cdot \widetilde{\Theta}\big(\frac{1}{s\sigma_p}\big)$ or $|\mathbf{w}_{y_i,r}^{(T_0)}[k]| \leq \eta$. Then during the training process after T_0, we have either $\mathrm{sgn}(\mathbf{w}_{y_i,r}^{(t)}[k]) = \mathrm{sgn}(\boldsymbol{\xi}_i[k])$ or $\mathrm{sgn}(\boldsymbol{\xi}_i[k]) \cdot \mathbf{w}_{y_i,r}^{(t)} \geq -\widetilde{O}(\eta)$ since if for some iteration number t' that we have $\mathrm{sgn}(\mathbf{w}_{y_i,r}^{(t')}[k]) = -\mathrm{sgn}(\boldsymbol{\xi}_i[k])$ but $\mathrm{sgn}(\mathbf{w}_{y_i,r}^{(t'-1)}[k]) = \mathrm{sgn}(\boldsymbol{\xi}_i[k])$, then after $\bar{\tau} = \widetilde{O}(1)$ steps (see

the proof of Lemma 6.7.2 for the definition of $\bar{\tau}$) in the constant number of steps the gradient will must be in the same direction of $\boldsymbol{\xi}_i[k]$, which will push $\mathbf{w}_{y_i,r}[k]$ back to zero or become positive along the direction of $\boldsymbol{\xi}_i[k]$. Therefore, based on this property we have the following regarding the inner product $\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle$,

$$\begin{aligned}
\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle &= \sum_{k \in \mathcal{B}_i \cup \{1\}} \mathbf{w}_{y_i,r}^{(t)}[k] \cdot \boldsymbol{\xi}_i[k] \\
&\geq \sum_{k \in \mathcal{B}_i \cup \{1\}} |\mathbf{w}_{y_i,r}^{(t)}[k]| \cdot |\boldsymbol{\xi}_i[k]| - \widetilde{O}(\eta) \cdot \sum_{k \in \mathcal{B}_i \cup \{1\}} |\boldsymbol{\xi}_i[k]| \\
&= \sum_{k \in \mathcal{B}_i \cup \{1\}} |\mathbf{w}_{y_i,r}^{(t)}[k]| \cdot |\boldsymbol{\xi}_i[k]| - \widetilde{O}(\eta s \sigma_p),
\end{aligned}$$

where the second inequality follows from the fact that the entry $\mathbf{w}_{y_i,r}^{(t)}[k]$ that has different sign of $\boldsymbol{\xi}_i[k]$ satisfies $|\mathbf{w}_{y_i,r}^{(t)}[k]| \leq \widetilde{O}(\eta)$. Then let $B_i^{(t)} = \sum_{j \in \mathcal{B}_i \cup \{1\}} \big| \mathbf{w}_{y_i,r}^{(t)}[k] \cdot \mathbb{1}(|\mathbf{w}_{y_i,r}^{(t)}[k]| \geq \widetilde{O}(\eta)) \big| \cdot |\boldsymbol{\xi}_i[k]|$, which satisfies $B_i^{(T_0)} = \widetilde{\Theta}(1)$ by Lemma 6.7.4. Then assume $B_i^{(t)}$ keeps increasing and reaches some value in the order of $\Theta\big(\log(dn\eta^{-1})\big)$, it holds that according to the inequality above

$$\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle = \Theta\big(\log(dn\eta^{-1})\big) - \widetilde{\Theta}(\eta s \sigma_p) = \Theta\big(\log(dn\eta^{-1})\big),$$

where we use the condition that $\eta = O\big((s\sigma_p)^{-1}\big)$. Then by Hypothesis (6.7.12) and (6.7.14) we know that $|\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle| = o(1)$, $\langle \mathbf{w}_{y_i,r^*}^{(t)}, \boldsymbol{\xi}_i \rangle = \widetilde{\Omega}(1)$, and $|\langle \mathbf{w}_{-y_i,r^*}^{(t)}, \boldsymbol{\xi}_i \rangle| = \widetilde{O}(d\eta) + \alpha |\langle \mathbf{w}_{-y_i,r^*}^{(t)}, \mathbf{v} \rangle| = o(1)$ then similar to (6.7.19), it holds that

$$\ell_{y_i,i}^{(t)} = \frac{e^{F_{-y_i}(\mathbf{W}^{(t)}, \mathbf{x}_i)}}{\sum_{j \in \{-1,1\}} e^{F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)}} \leq \exp\big(- \Theta\big(\sigma(\langle \mathbf{w}_{y_i,r^*}^{(t)}, \boldsymbol{\xi}_i \rangle)\big)\big) \leq \mathrm{poly}(d^{-1}, n^{-1}, \eta).$$

Therefore, at this time we have for all $k \in \mathcal{B}_i$,

$$\ell_{y_i,i}^{(t)} \sigma\langle (\mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \boldsymbol{\xi}_i[k] \leq \mathrm{poly}(d^{-1}, n^{-1}, \eta) \cdot \Theta\big(\log^{q-1}(dn\eta^{-1})\big) \cdot \widetilde{\Theta}(\sigma_p) \leq n\lambda\eta.$$

Then for all $|\mathbf{w}_{y_i,r}^{(t)}[k]| \geq \widetilde{O}(\eta)$, the sign of the gradient satisfies

$$\begin{aligned}
\mathrm{sgn}\big(\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^{(t)})[k]\big) &= -\mathrm{sgn}\bigg(\ell_{y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \boldsymbol{\xi}_i[k] - n\lambda \mathbf{w}_{y_i,r}^{(t)}[k] \bigg) \\
&= \mathrm{sgn}(n\lambda\eta - \mathbf{w}_{y_i,r}^{(t)}[k]) \\
&= \mathrm{sgn}(\mathbf{w}_{y_i,r}^{(t)}[k]).
\end{aligned}$$

Then note that $|\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^{(t)})[k]| = \Theta(|\lambda \mathbf{w}_{y_i,r}^{(t)}[k]|) \geq \Theta\big(n^{-1}\eta s\sigma_p \ell_{y_i,i}^{(t)} + \lambda\eta\big)$, by the update rule of $\mathbf{w}_{y_i,r}^{(t)}[k]$ and Lemma 6.7.2, we know the sign gradient will dominate the update process. Then we have $|\mathbf{w}_{y_i,r}^{(t+1)}[k]| = |\mathbf{w}_{y_i,r}^{(t)}[k] - \Theta(\eta) \cdot \mathrm{sgn}(\mathbf{w}_{y_i,r}^{(t)}[k])| \leq |\mathbf{w}_{y_i,r}^{(t)}[k]|$, which implies that $\big|\mathbf{w}_{y_i,r}^{(t)}[k] \cdot \mathbb{1}(|\mathbf{w}_{y_i,r}^{(t)}[k]| \geq \widetilde{O}(\eta))\big|$ decreases so that $B_i^{(t)}$ also decreases. Therefore, we can conclude that $B_i^{(t)}$ will not exceed $\Theta\big(\log(dn\eta^{-1})\big)$. Then combining the results for all $i \in [n]$ gives

$$\sum_{k\in\mathcal{B}_i} |\mathbf{w}_{y_i,r^*}^{(t)}[k]| \cdot |\boldsymbol{\xi}_i[k]| \leq B_i^{(t)} + \widetilde{O}(s\eta\sigma_p) \leq \Theta\big(\log(dn\eta^{-1})\big) + O(1) = \widetilde{\Theta}(1),$$

where in the first inequality we again use the condition that $\eta = o(1/d) = o\big((s\sigma_p)^{-1}\big)$. This verifies Hypothesis (6.7.13). Notably, this also implies that $\langle \mathbf{w}_{y_i,r^*}^{(t)}, \boldsymbol{\xi}_i\rangle = \max_{r\in[m]} \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle \leq \widetilde{\Theta}(1)$.

Verifying Hypothesis (6.7.14)**.** In order to verify Hypothesis (6.7.14), let us first recall the update rule of $\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle$:

$$\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{v}\rangle = \langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle - \eta \bigg\langle \frac{\mathbf{m}_{j,r}^{(t)}}{\sqrt{\mathbf{v}_{j,r}^{(t)}}}, \mathbf{v} \bigg\rangle.$$

Then by Lemma 6.7.2, we know that if $|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})[1]| \leq \widetilde{\Theta}(\eta)$, then $|\mathbf{m}_{j,r}^{(t)}/\sqrt{\mathbf{v}_{j,r}^{(t)}}| \leq \Theta(1)$ and otherwise

$$\bigg\langle \frac{\mathbf{m}_{j,r}^{(t)}}{\sqrt{\mathbf{v}_{j,r}^{(t)}}}, \mathbf{v} \bigg\rangle = -\mathrm{sgn}\bigg(\sum_{i=1}^n y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v}\rangle) - \alpha \sum_{i=1}^n y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) - n\lambda \mathbf{w}_{j,r}^{(t)}[1] \bigg) \cdot \Theta(1).$$

Without loss of generality we assume $j = 1$, then by Lemma 6.7.4 we know that $\mathbf{w}_{1,r}^{(T_0)}[1] = -\widetilde{\Omega}\big(\frac{1}{s\sigma_p}\big)$. In the remaining proof, we will show that either $\mathbf{w}_{1,r}^{(t+1)}[1] \in [0, \widetilde{\Theta}(\lambda^{-1}\eta)]$ or $\mathbf{w}_{1,r}^{(t+1)}[1] \in \big[-\widetilde{O}\big(\frac{n\alpha}{s\sigma_p^2}\big), 0\big)$.

First we will show that $\mathbf{w}_{1,r}^{(t+1)}[1] \in [0, \widetilde{\Theta}(\lambda^{-1}\eta)]$ for all r. Note that in the beginning of this stage, we have $\mathbf{w}_{1,r}^{(T_0)}[1] < 0$. In order to make the sign of $\mathbf{w}_{1,r}^{(t')}[1]$ flip, we must have, in

some iteration $t' \le t$ that satisfies $\mathbf{w}_{1,r}^{(t')}[1] \in [0, \widetilde{\Theta}(\lambda^{-1}\eta)]$, therefore

$$\begin{aligned}
-n\nabla_{\mathbf{w}_{1,r}} L(\mathbf{W}^{(t')})[1] &= \sum_{i=1}^{n} y_i \ell_{j,i}^{(t')} \sigma'(\langle \mathbf{w}_{j,r}^{(t')}, y_i \mathbf{v} \rangle) - \alpha \sum_{i=1}^{n} y_i \ell_{j,i}^{(t')} \sigma'(\langle \mathbf{w}_{j,r}^{(t')}, \boldsymbol{\xi}_i \rangle) - n\lambda \mathbf{w}_{j,r}^{(t')}[1] \\
&\le n\big[\big(\mathbf{w}_{j,r}^{(t')}[1]\big)^{q-2} - \lambda\big] \cdot \mathbf{w}_{j,r}^{(t')}[1] \le -\widetilde{\Theta}(n\eta) \le 0,
\end{aligned}$$

where the second inequality holds since $\eta = o(\lambda^{(q-1)/(q-2)})$. Note that $|\nabla_{\mathbf{w}_{1,r}} L(\mathbf{W}^{(t')})[1]| \ge \widetilde{\Theta}(\eta)$, then by Lemma 6.7.2 we know that Adam is similar to sign gradient descent and thus $\mathbf{w}_{1,r}^{(t'+1)}[1] = \mathbf{w}_{1,r}^{(t')}[1] - \Theta(\eta)$ which starts to decrease. This implies that if $\mathbf{w}_{1,r}^{(t+1)}[1]$ is positive, then it cannot exceed $\widetilde{\Theta}(\lambda^{-1}\eta) = o(1)$.

Then we can prove that if $\mathbf{w}_{1,r}^{(t+1)}[1]$ is negative, then $|\mathbf{w}_{1,r}^{(t+1)}[1]| = \widetilde{O}\big(\frac{n\alpha}{s\sigma_p^2}\big)$. In this case we have for all $t' \le t$,

$$\begin{aligned}
-n\nabla_{\mathbf{w}_{1,r}^{(t)}} L(\mathbf{W}^{(t')})[1] &= \sum_{i=1}^{n} y_i \ell_{1,i}^{(t')} \sigma'(\langle \mathbf{w}_{1,r}^{(t')}, y_i \mathbf{v} \rangle) - \alpha \sum_{i=1}^{n} y_i \ell_{1,i}^{(t')} \sigma'(\langle \mathbf{w}_{1,r}^{(t')}, \boldsymbol{\xi}_i \rangle) - n\lambda \mathbf{w}_{1,r}^{(t')}[1] \\
&\ge -\sum_{i:y_i=1} |\ell_{1,i}^{(t')}| \cdot \widetilde{\Theta}(\alpha) + n\lambda |\mathbf{w}_{1,r}^{(t')}[1]| + \sum_{i:y_i=-1} |\ell_{1,i}^{(t')}| \cdot |\mathbf{w}_{1,r}^{(t')}[1]|^{q-1}, \\
&\ge -\sum_{i:y_i=1} |\ell_{1,i}^{(t')}| \cdot \widetilde{\Theta}(\alpha) + n\lambda |\mathbf{w}_{1,r}^{(t')}[1]|,
\end{aligned}$$

where in the inequality we use Hypothesis (6.7.13) and (6.7.14) to get that

$$\langle \mathbf{w}_{y_i,r}^{(t')}, \boldsymbol{\xi}_i \rangle \le \sum_{k \in \mathcal{B}_i} |\mathbf{w}_{y_i,r}^{(t')}[k]| \cdot \max_{k \in \mathcal{B}_i} |\boldsymbol{\xi}_i[k]| + \alpha |\langle \mathbf{w}_{y_i,r}^{(t')}, \mathbf{v} \rangle| = \widetilde{\Theta}(1).$$

Recall from (6.7.18) that we have $|\ell_{j,i}^{(t')}| = \widetilde{O}\big(\frac{n\lambda}{s\sigma_p^2}\big)$, therefore we have if $\mathbf{w}_{j,r}^{(t')}[1]$ is smaller than some value in the order of $-\widetilde{\Theta}\big(\frac{n\alpha}{s\sigma_p^2}\big) \cdot \mathrm{polylog}(d)$, then

$$-n\nabla_{\mathbf{w}_{1,r}^{(t)}} L(\mathbf{W}^{(t')})[1] \ge -\widetilde{\Theta}\bigg(\frac{\alpha n^2 \lambda}{s\sigma_p^2}\bigg) + \widetilde{\Theta}\bigg(\frac{n\lambda \cdot n\alpha}{s\sigma_p^2}\bigg) \cdot \mathrm{polylog}(d) \ge \widetilde{\Theta}(n\eta),$$

which by Lemma 6.7.2 implies that $\mathbf{w}_{j,r}^{(t')}[1]$ will increase. Therefore, we can conclude that $\mathbf{w}^{(t+1)} \in \big[-\widetilde{O}\big(\frac{n\alpha}{s\sigma_p^2}\big), 0\big)$ in this case, which verifies Hypothesis (6.7.14). □

Lemma 6.7.6 (Lemma 6.4.5, restated)**.** If the step size satisfies $\eta = O(d^{-1/2})$, then for any t it holds that

$$L(\mathbf{W}^{(t+1)}) - L(\mathbf{W}^{(t)}) \le -\eta \|\nabla L(\mathbf{W}^{(t)})\|_1 + \widetilde{\Theta}(\eta^2 d).$$

Proof. Let $\Delta F_{j,i} = F_j(\mathbf{W}^{(t+1)}, \mathbf{x}_i) - F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)$. Then regarding the loss function

$$L_i(\mathbf{W}) = -\log \frac{e^{F_{y_i}(\mathbf{W}, \mathbf{x}_i)}}{\sum_j e^{F_j(\mathbf{W}, \mathbf{x}_i)}} = -F_{y_i}(\mathbf{W}, \mathbf{x}_i) + \log\Big(\sum_j e^{F_j(\mathbf{W}, \mathbf{x}_i)}\Big).$$

It is clear that the function $L_i(\mathbf{W})$ is 1-smooth with respect to the vector $[F_{-1}(\mathbf{W}, \mathbf{x}_i), F_1(\mathbf{W}, \mathbf{x}_i)]$. Then based on the definition of $\Delta F_{j,i}$, we have

$$L_i(\mathbf{W}^{(t+1)}) - L_i(\mathbf{W}^{(t)}) \leq \sum_j \frac{\partial L_i(\mathbf{W}^{(t)})}{\partial F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)} \cdot \Delta F_{j,i} + \sum_j (\Delta F_{j,i})^2. \tag{6.7.20}$$

Moreover, note that

$$F_j(\mathbf{W}^{(t)}, \mathbf{x}_i) = \sum_{r=1}^{m} \big[\sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\big].$$

By the results that $\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle \leq \widetilde{\Theta}(1)$ and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}\rangle \leq \widetilde{\Theta}(1)$, for any $\eta = O(d^{-1/2})$, we have

$$\langle \mathbf{w}_{j,r}^{(t+1)}, \mathbf{v}\rangle \leq \langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v}\rangle + \eta \leq \widetilde{\Theta}(1), \quad \langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle \leq \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle + \widetilde{\Theta}(\eta s^{1/2}) \leq \widetilde{\Theta}(1),$$

which implies that the smoothness parameter of the functions $\sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v}\rangle)$ and $\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)$ are at most $\widetilde{\Theta}(1)$ for any $\mathbf{w}$ in the path between $\mathbf{w}_{j,r}^{(t)}$ and $\mathbf{w}_{j,r}^{(t+1)}$. Then we can apply first Taylor expansion on $\sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v}\rangle)$ and $\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)$ and bound the second-order error as follows,

$$\begin{aligned}
&\big|\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, y_i \mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v}\rangle) - \big\langle \nabla_{\mathbf{w}_{j,r}} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v}\rangle), \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\big\rangle\big| \\
&\leq \widetilde{\Theta}\big(\|\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\|_2^2\big) = \widetilde{\Theta}(\eta^2 d),
\end{aligned} \tag{6.7.21}$$

where the last inequality is due to Lemma 6.7.2 that

$$[\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}]^2 = \eta^2 \bigg\| \frac{\mathbf{m}_{j,r}^{(t)}}{\sqrt{\mathbf{v}_{j,r}^{(t)}}} \bigg\|_2^2 \leq \Theta(\eta^2 d).$$

Similarly, we can also show that

$$\big|\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) - \big\langle \nabla_{\mathbf{w}_{j,r}} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle), \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\big\rangle\big| \leq \Theta(\eta^2 d). \tag{6.7.22}$$

Combining the above bounds on the second-order errors, we have

$$\big|\Delta F_{j,i} - \langle \nabla_{\mathbf{W}} F_j(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)} \rangle\big| \leq \widetilde{\Theta}(m\eta^2 d) = \widetilde{\Theta}(\eta^2 d), \tag{6.7.23}$$

where the last equation is due to our assumption that $m = \widetilde{\Theta}(1)$. Besides, by (6.7.21) and (6.7.22) the convexity property of the function $\sigma(x)$, we also have

$$\begin{aligned}
\big|\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, y_i \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle)\big| &\leq |\big\langle \nabla_{\mathbf{w}_{j,r}} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle), \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)} \rangle| + \widetilde{\Theta}(\eta^2 d) \\
&= \widetilde{\Theta}\big(\eta |\sigma'(\langle \mathbf{w}_{j,r}^{(t+1)}, y_i \mathbf{v} \rangle)| \cdot \|\mathbf{v}\|_1\big) + \widetilde{\Theta}(\eta^2 d) \\
&= \widetilde{\Theta}(\eta + \eta^2 d); \\
\big|\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle)\big| &\leq |\big\langle \nabla_{\mathbf{w}_{j,r}} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle), \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)} \rangle| + \widetilde{\Theta}(\eta^2 d) \\
&= \widetilde{\Theta}\big(\eta |\sigma'(\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle)| \cdot \|\boldsymbol{\xi}\|_1\big) + \widetilde{\Theta}(\eta^2 d) \\
&= \widetilde{\Theta}(\eta s \sigma_p + \eta^2 d).
\end{aligned}$$

These bounds further imply that

$$|\Delta F_{j,i}| \leq \widetilde{\Theta}\big(m \cdot (\eta s \sigma_p + \eta^2 d)\big) = \widetilde{\Theta}\big(\eta s \sigma_p + \eta^2 d\big). \tag{6.7.24}$$

Now we can plug (6.7.23) and (6.7.24) into (6.7.20) and get

$$\begin{aligned}
L_i(\mathbf{W}^{(t+1)}) - L_i(\mathbf{W}^{(t)}) &\leq \sum_j \frac{\partial L_i(\mathbf{W}^{(t)})}{\partial F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)} \cdot \Delta F_{j,i} + \sum_j (\Delta F_{j,i})^2 \\
&\leq \sum_j \frac{\partial L_i(\mathbf{W}^{(t)})}{\partial F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)} \cdot \langle \nabla_{\mathbf{W}} F_j(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)} \rangle \\
&\quad + \widetilde{\Theta}(\eta^2 d) + \widetilde{\Theta}\big((\eta s \sigma_p + \eta^2 d)^2\big) \\
&= \langle \nabla L_i(\mathbf{W}^{(t)}), \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)} \rangle + \widetilde{\Theta}(\eta^2 d),
\end{aligned} \tag{6.7.25}$$

where in the second inequality we use the fact that $L_i(\mathbf{W})$ is 1-Lipschitz with respect to $F_j(\mathbf{W}, \mathbf{x}_i)$ and the last equation is due to our assumption that $\sigma_p = O(s^{-1/2})$ so that $\widetilde{\Theta}((\eta s \sigma_p + \eta^2 d)^2) = \widetilde{O}(\eta^2 d)$.

Now we are ready to characterize the behavior on the entire training objective $L(\mathbf{W}) = n^{-1} \sum_{i=1}^n L_i(\mathbf{W}) + \lambda \|\mathbf{W}\|_F^2$. Note that $\lambda \|\mathbf{W}\|_F^2$ is 2λ-smoothness, where $\lambda = o(1)$. Then

applying (6.7.25) for all $i \in [n]$ gives

$$\begin{aligned} L(\mathbf{W}^{(t+1)}) - L(\mathbf{W}^{(t)}) &= \frac{1}{n}\sum_{i=1}^{n}\big[L_i(\mathbf{W}^{(t+1)}) - L_i(\mathbf{W}^{(t)})\big] + \lambda\big(\|\mathbf{W}^{(t+1)}\|_F^2 - \|\mathbf{W}^{(t)}\|_F^2\big) \\ &\leq \langle \nabla L(\mathbf{W}^{(t)}), \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)}\rangle + \widetilde{\Theta}(\eta^2 d), \end{aligned}$$

where the second equation uses the fact that $\|\mathbf{W}^{(t+1)} - \mathbf{W}^{(t)}\|_F^2 = \widetilde{\Theta}(\eta^2 d)$. Recall that we have

$$\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)} = -\eta \cdot \frac{\mathbf{m}_{j,r}^{(t)}}{\sqrt{\mathbf{v}_{j,r}^{(t)}}}.$$

Then by Lemma 6.7.2, we know that $\mathbf{m}_{j,r}^{(t)}[k]/\sqrt{\mathbf{v}_{j,r}^{(t)}}[k]$ is close to sign gradient if $\nabla L(\mathbf{w}^{(t)})[k]$ is large. Then we have

$$\begin{aligned} \bigg\langle \nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)}), \frac{\mathbf{m}_{j,r}^{(t)}}{\sqrt{\mathbf{v}_{j,r}^{(t)}}} \bigg\rangle &\geq \Theta\big(\big\|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})\big\|_1\big) - \widetilde{\Theta}\big(d \cdot \eta\big) - \widetilde{\Theta}(ns \cdot \eta s \sigma_p) \\ &\geq \Theta\big(\big\|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})\big\|_1\big) - \widetilde{\Theta}(d\eta), \end{aligned}$$

where the second and last terms on the R.H.S. of the first inequality are contributed by the small gradient coordinates $k \notin \cup_{i=1}^n \mathcal{B}_i$ and $k \in \cup_{i=1}^n \mathcal{B}_i$ respectively, and the last inequality is by the fact that $ns^2\sigma_p = O(d)$. Therefore, based on this fact (6.7.25) further leads to

$$L(\mathbf{W}^{(t+1)}) - L(\mathbf{W}^{(t)}) \leq -\eta \|\nabla L(\mathbf{W}^{(t)})\|_1 + \widetilde{\Theta}(\eta^2 d),$$

which completes the proof.

□

Lemma 6.7.7 (Generalization Performance of Adam)**.** Let

$$\mathbf{W}^* = \underset{\mathbf{W} \in \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(T)}\}}{\operatorname{argmin}} \|\nabla L(\mathbf{W})\|_1.$$

Then for all training data, we have

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[F_{y_i}(\mathbf{W}^*, \mathbf{x}_i) \leq F_{-y_i}(\mathbf{W}^*, \mathbf{x}_i)\right] = 0.$$

Moreover, in terms of the test data $(\mathbf{x}, y) \sim \mathcal{D}$, we have

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[F_y(\mathbf{W}^*, \mathbf{x}) \le F_{-y}(\mathbf{W}^*, \mathbf{x})\big] \ge \frac{1}{2}.$$

Proof. By Lemma 6.7.6, we know that the algorithm will converge to a point with very small gradient (up to $O(\eta d)$ in ℓ_1 norm). Then in terms of a noise vector $\boldsymbol{\xi}_i$, we have

$$\sum_{k\in\mathcal{B}_i} \big|\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^*)[k]\big| \le O(\eta d). \tag{6.7.26}$$

Note that

$$n\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^*)[k] = \ell^*_{y_i,i}\sigma'(\langle \mathbf{w}^*_{y_i,r}, \boldsymbol{\xi}_i\rangle)\boldsymbol{\xi}_i[k] - n\lambda \mathbf{w}^*_{y_i,r}[k],$$

where $\ell^*_{y_i,i} = 1 - \mathrm{logit}_{y_i}(F^*, \mathbf{x}_i)$. Then by triangle inequality and (6.7.26), we have for any $r \in [m]$,

$$\bigg|\sum_{k\in\mathcal{B}_i} |\ell^*_{y_i,i}|\sigma'(\langle \mathbf{w}^*_{y_i,r}, \boldsymbol{\xi}_i\rangle)|\boldsymbol{\xi}_i[k]| - n\lambda \sum_{k\in\mathcal{B}_i} |\mathbf{w}^*_{y_i,r}[k]|\bigg| \le n\sum_{k\in\mathcal{B}_i} \big|\nabla_{\mathbf{w}_{y_i,r}} L(\mathbf{W}^*)[k]\big| \le O(n\eta d).$$

Then by Lemma 6.7.5, let $r^* = \arg\max_{r\in[m]} \langle \mathbf{w}^*_{y_i,r}, \boldsymbol{\xi}_i\rangle$, we have $\langle \mathbf{w}_{y_i,r^*}, \boldsymbol{\xi}_i\rangle = \widetilde{\Theta}(1)$ and $\sum_{k\in\mathcal{B}_i} |\mathbf{w}^*_{y_i,r^*}[k]| \cdot |\boldsymbol{\xi}_i[k]| = \widetilde{\Theta}(1)$. Note that $|\boldsymbol{\xi}_i[k]| = \widetilde{O}(\sigma_p)$, we have $\sum_{k\in\mathcal{B}_i} |\mathbf{w}^*_{y_i,r^*}[k]| \ge \widetilde{\Theta}(1/\sigma_p)$. Then according to the inequality above, it holds that

$$|\ell^*_{y_i,i}| \cdot \widetilde{\Theta}(s\sigma_p) \ge \widetilde{\Theta}\bigg(n\lambda \sum_{k\in\mathcal{B}_i} |\mathbf{w}^*_{y_i,r}[k]| - n\eta d\bigg) \ge \widetilde{\Theta}\bigg(\frac{n\lambda}{\sigma_p}\bigg),$$

where the second inequality is due to our choice of η. This further implies that $|\ell^*_{y_i,i}| = |\ell^*_{-y_i,i}| = \widetilde{\Theta}\big(\frac{n\lambda}{s\sigma_p^2}\big)$ by combining the above results with (6.7.18). Then let us move to the gradient with respect to the first coordinate. In particular, since $\|\nabla L(\mathbf{W}^*)\|_1 \le O(\eta d)$, we have

$$\begin{aligned} |n\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^*)[1]| &= \bigg|\sum_{i=1}^n y_i \ell^*_{j,i}\sigma'(\langle \mathbf{w}^*_{j,r}, y_i\mathbf{v}\rangle) - \alpha\sum_{i=1}^n y_i\ell^*_{j,i}\sigma'(\langle \mathbf{w}^*_{j,r}, \boldsymbol{\xi}_i\rangle) - n\lambda \mathbf{w}^*_{j,r}[1]\bigg| \\ &\le O(n\eta d). \end{aligned} \tag{6.7.27}$$

Then note that $\mathrm{sgn}(y_i\ell^*_{j,i}) = \mathrm{sgn}(j)$, it is clear that $\mathbf{w}^*_{j,r^*}[1] \cdot j \leq 0$ since otherwise

$$|n\nabla_{\mathbf{w}_{j,r^*}} L(\mathbf{W}^*)[1]| \geq \bigg|\alpha \sum_{i=1}^{n} y_i \ell^*_{j,i} \big[\sigma'(\langle \mathbf{w}^*_{j,r^*}, \boldsymbol{\xi}_i\rangle) - \sigma'(\langle \mathbf{w}^*_{j,r^*}, y_i\mathbf{v}\rangle)\big]\bigg| \geq \widetilde{\Theta}\bigg(\frac{\alpha n^2 \lambda}{s\sigma_p^2}\bigg) \geq \widetilde{\Omega}(n\eta d),$$

which contradicts (6.7.27). Therefore, using the fact that $\mathbf{w}^*_{j,r^*}[1] \cdot j \leq 0$, we have

$$|n\nabla_{\mathbf{w}_{j,r^*}} L(\mathbf{W}^*)[1]| = \bigg|\alpha \sum_{i:y_i=j}^{n} y_i \ell^*_{j,i} \sigma'(\langle \mathbf{w}^*_{j,r^*}, \boldsymbol{\xi}_i\rangle) - \sum_{i:y_i=-j}^{n} y_i \ell^*_{j,i} \sigma'(|\mathbf{w}^*_{j,r^*}[1]|)\big] - n\lambda |\mathbf{w}^*_{j,r^*}[1]|\bigg|.$$

Then applying (6.7.27)and using the fact that $|\ell^*_{y_i,i}| = |\ell^*_{-y_i,i}| = \widetilde{\Theta}\big(\frac{n\lambda}{s\sigma_p^2}\big)$ for all $i \in [n]$, it is clear that

$$|\mathbf{w}^*_{j,r^*}[1]| \geq \widetilde{\Theta}\bigg(\alpha^{1/(q-1)} \wedge \frac{n\alpha}{s\sigma_p^2}\bigg) \geq \widetilde{\Theta}\bigg(\frac{n\alpha}{s\sigma_p^2}\bigg),$$

where the second equality is due to our choice of σ_p and α. Then combining with Lemma 6.7.5 and the fact that $\mathbf{w}^*_{j,r^*}[1] \cdot j < 0$, we have

$$\mathbf{w}^*_{j,r^*}[1] \cdot j \leq -\widetilde{\Theta}\bigg(\frac{n\alpha}{s\sigma_p^2}\bigg).$$

Now we are ready to evaluate the training error and test error. In terms of training error, it is clear that by Lemma 6.7.5, we have $\langle \mathbf{w}^*_{y_i,r^*}, \boldsymbol{\xi}_i\rangle \geq \widetilde{\Theta}(1)$, $\langle \mathbf{w}^*_{y_i,r}, \boldsymbol{\xi}_i\rangle \geq -o(1)$, and $|\langle \mathbf{w}^*_{y_i,r}, \mathbf{v}\rangle| = o(1)$, $|\langle \mathbf{w}^*_{-y_i,r}, \boldsymbol{\xi}_i\rangle| = o(1)$. Then we have for any training data $(\mathbf{x}_i, y_i)$,

$$F_{y_i}(\mathbf{W}^*, \mathbf{x}_i) = \sum_{r=1}^{m} \big[\sigma(\langle \mathbf{w}^*_{y_i,r}, y_i\mathbf{v}\rangle) + \sigma(\langle \mathbf{w}^*_{y_i,r}, \boldsymbol{\xi}_i\rangle)\big] = \widetilde{\Theta}(1),$$
$$F_{-y_i}(\mathbf{W}^*, \mathbf{x}_i) = \sum_{r=1}^{m} \big[\sigma(\langle \mathbf{w}^*_{-y_i,r}, -y_i\mathbf{v}\rangle) + \sigma(\langle \mathbf{w}^*_{-y_i,r}, \boldsymbol{\xi}_i\rangle)\big] = o(1),$$

which directly implies that the NN model $\mathbf{W}^*$ can correctly classify all training data and thus achieve zero training error.

In terms of the test data $(\mathbf{x}, y)$ where $\mathbf{x} = [y\mathbf{v}, \boldsymbol{\xi}]$, which is generated according to Definition 6.2.1. Note that for each neural, its weight $\mathbf{w}^*_{j,r}$ can be decomposed into two parts: the first coordinate and the rest $d-1$ coordinates. As previously discussed, for any $j \in [2]$ and $r = r^*$, we have $\mathrm{sgn}(j) \cdot \mathbf{w}^*_{j,r}[1] \leq -\widetilde{\Theta}\big(n\alpha/(s\sigma_p^2)\big)$ and $\mathrm{sgn}(j) \cdot \mathbf{w}^*_{j,r}[1] \leq \widetilde{\Theta}(\lambda^{-1}\eta)$

for $r \neq r^*$. Therefore, using the fact that $\widetilde{\Theta}\big(n\alpha/(s\sigma_p^2)\big) = \omega(\lambda^{-1}\eta)$ and Lemma 6.7.5, given the test data $(\mathbf{x}, y)$, we have

$$\begin{aligned}
F_y(\mathbf{W}^*, \mathbf{x}) &= \sum_{r=1}^{m} \big[\sigma(\langle \mathbf{w}_{y,r}^*, y\mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_{y,r}^*, \boldsymbol{\xi}\rangle)\big] \\
&\leq \sum_{r=1}^{m} \widetilde{\Theta}\bigg(\bigg[\alpha \cdot \frac{n\alpha}{s\sigma_p^2} + \zeta_{y,r}\bigg]_+^q\bigg), \\
F_{-y}(\mathbf{W}^*, \mathbf{x})) &= \sum_{r=1}^{m} \big[\sigma(\langle \mathbf{w}_{-y,r}^*, y\mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_{-y,r}^*, \boldsymbol{\xi}\rangle)\big] \\
&\geq \widetilde{\Theta}\big[|\mathbf{w}_{-y,r^*}^*[1]|^q + [\zeta_{-y,r^*}]_+^q\big] \\
&\geq \Theta\bigg(\bigg[\frac{n\alpha}{s\sigma_p^2}\bigg]_+^q + [\zeta_{-y,r^*}]_+^q\bigg),
\end{aligned}$$

where the random variables $\zeta_{y,r}$ and $\zeta_{y,r}$ are symmetric and independent of $\mathbf{v}$. Besides, note that $\alpha = o(1)$, it can be clearly shown that $\alpha \cdot n\alpha/(s\sigma_p^2) \ll n\alpha/(s\sigma_p^2)$. Therefore, if the random noise $\zeta_{y,r}$ and $\zeta_{-y,r}$ are dominated by the feature noise term $\langle \mathbf{w}_{-y,r^*}^*, y\mathbf{v}\rangle$, we can directly get that $F_y(\mathbf{W}^*, \mathbf{x}) \leq F_{-y}(\mathbf{W}^*, \mathbf{x}))$ (recall that $m = \widetilde{\Theta}(1)$), which implies that the model has been biased by the feature noise and the true feature vector in the test dataset will not give any "positive" effect to the classification. Also note that ζ_y and ζ_{-y} are also independent of $\mathbf{v}$, which implies that if the random noise dominates the feature noise term, the model $\mathbf{W}^*$ will give at least 0.5 error on test data. In sum, we can conclude that with probability at least $1/2$ it holds that $F_y(\mathbf{W}^*, \mathbf{x}) \leq F_{-y}(\mathbf{W}^*, \mathbf{x})$, which implies that the output of Adam achieves $1/2$ test error. □

6.7.3 Proof for Gradient Descent

Recall the feature learning and noise memorization of gradient descent can be formulated by

$$\begin{aligned}
\langle \mathbf{w}_{j,r}^{(t+1)}, j \cdot \mathbf{v} \rangle &= (1 - \eta\lambda) \cdot \langle \mathbf{w}_{j,r}^{(t)}, j \cdot \mathbf{v} \rangle \\
&\quad + \frac{\eta}{n} \cdot j \cdot \bigg(\sum_{i=1}^{n} y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \mathbf{v} \rangle) - \alpha \sum_{i=1}^{n} y_i \ell_{j,i}^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \bigg), \\
\langle \mathbf{w}_{y_i,r}^{(t+1)}, \boldsymbol{\xi}_i \rangle &= (1 - \eta\lambda) \cdot \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle + \frac{\eta}{n} \cdot \sum_{k \in \mathcal{B}_i} \ell_{y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \boldsymbol{\xi}_i[k]^2 \\
&\quad + \frac{\eta\alpha}{n} \cdot \bigg(\alpha \sum_{s=1}^{n} \ell_{y_i,s}^{(t)} \sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_s \rangle) - \sum_{s=1}^{n} y_s \ell_{y_i,s}^{(t)} \sigma'(\langle \mathbf{w}_{y_i,r}^{(t)}, y_s \mathbf{v} \rangle) \bigg). \quad (6.7.28)
\end{aligned}$$

Then similar to the analysis for Adam, we decompose the gradient descent process into multiple stages and characterize the algorithmic behaviors separately. The following lemma characterizes the first training stage, i.e., the stage where all outputs $F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)$ remain in the constant level for all j and i.

Lemma 6.7.8. [Lemma 6.4.6, restated] Suppose the training data is generated according to Definition 6.2.1 and $\lambda = o(\sigma_0^{q-2}\sigma_p/n)$. Let $\Lambda_j^{(t)} = \max_{r\in[m]} \langle \mathbf{w}_{j,r}^{(t+1)}, j \cdot \mathbf{v} \rangle$, $\Gamma_{j,i}^{(t)} = \max_{r\in[m]} \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle$, and $\Gamma_j^{(t)} = \max_{i:y_i=j} \Gamma_{j,i}^{(t)}$. Then let T_j be the iteration number that $\Lambda_j^{(t)}$ reaches $\Theta(1/m)$, we have

$$T_j = \widetilde{\Theta}(\sigma_0^{2-q}/\eta) \quad \text{for all } j \in \{-1, 1\}.$$

Moreover, let $T_0 = \min_{j\in\{\pm 1\}}\{T_j\}$, then for all $t \le T_0$ it holds that $\Gamma_j^{(t)} = \widetilde{O}(\sigma_0)$ for all $j \in \{-1, 1\}$.

We first provide the following useful lemma.

Lemma 6.7.9. Let $\{x_t, y_t\}_{t=1,\ldots}$ be two positive sequences that satisfy

$$x_{t+1} \ge x_t + \eta \cdot A x_t^{q-1},$$
$$y_{t+1} \le y_t + \eta \cdot B y_t^{q-1},$$

for some $A = \Theta(1)$ and $B = o(1)$. Then for any $q \geq 3$ and suppose $y_0 = O(x_0)$ and $\eta < O(x_0)$, we have for every $C \in [x_0, O(1)]$, let T_x be the first iteration such that $x_t \geq C$, then we have $T_x\eta = \Theta(x_0^{2-q})$ and

$$y_{T_x} \leq O(x_0).$$

Proof. By Claim C.20 in [AL20], we have $T_x\eta = \Theta(x_0^{2-q})$. Then we will show

$$y_t \leq 2x_0$$

for all $t \leq T_x$. In particular, let $T_x\eta = C'x_0^{2-q}$ for some absolute constant C' and assume $C'B2^{q-1} < 1$ (this is true since $B = o(1)$), we first made the following induction hypothesis on y_t for all $t \leq T_a$,

$$y_t \leq y_0 + t\eta B'(2x_0)^{q-1}.$$

Note that for any $t \leq T_0$, this hypothesis clearly implies that

$$y_t \leq y_0 + T_x\eta B'2^{q-1}x_0^{q-1} \leq x_0 + CB2^{q-1}x_0^{2-q} \cdot x_0^{q-1} \leq 2x_0.$$

Then we are able to verify the hypothesis at time $t+1$ based on the recursive upper bound of y_t, i.e.,

$$\begin{aligned} y_{t+1} &\leq y_t + \eta \cdot By_t^{q-1} \\ &\leq y_0 + t\eta B(2x_0)^{q-1} + \eta \cdot By_t^{q-1} \\ &\leq y_0 + (t+1)\eta B(2x_0)^{q-1}. \end{aligned}$$

Therefore, we can conclude that $y_t \leq 2x_0$ for all $t \leq T_x$. This completes the proof. □

Now we are ready to complete the proof of Lemma 6.7.8.

Proof of Lemma 6.7.8. Note that at the initialization, we have $|\langle \mathbf{w}_{j,r}^{(0)}, \mathbf{v} \rangle| = \widetilde{\Theta}(\sigma_0)$ and $|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle| = \widetilde{\Theta}(s^{1/2}\sigma_p\sigma_0)$. Then based on the parameter scaling summarized in Section 6.7.1,

we have

$$F_j(\mathbf{W}^{(0)}, \mathbf{x}_i) = \sum_{r=1}^{m} \big[\sigma(\langle \mathbf{w}_{j,r}^{(0)}, y_i \mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle)\big] = o(1)$$

for all $j \in \{-1, 1\}$. Then we have

$$|\ell_{j,i}^{(0)}| \geq \min\left\{ \frac{e^{F_j(\mathbf{W}^{(0)}, \mathbf{x}_i)}}{\sum_j e^{F_{+1}(\mathbf{W}^{(0)}, \mathbf{x}_i)}}, \frac{e^{F_{-1}(\mathbf{W}^{(0)}, \mathbf{x}_i)}}{\sum_j e^{F_j(\mathbf{W}^{(0)}, \mathbf{x}_i)}} \right\} = \Theta(1).$$

Then we will consider the training period where $|\ell_{j,i}^{(t)}| = \Theta(1)$ for all j, i, and t. Besides, note that $\mathrm{sgn}(y_i \ell_{j,i}^{(t)}) = j$. Therefore, let $r^* = \arg\max_r \langle \mathbf{w}_{j,r}^{(t-1)}, j \cdot \mathbf{v}\rangle$, (6.7.28) implies that

$$\begin{aligned}
\Lambda_j^{(t)} &= \langle \mathbf{w}_{j,r^*}^{(t-1)}, j \cdot \mathbf{v}\rangle \\
&= (1 - \eta\lambda) \cdot \langle \mathbf{w}_{j,r^*}^{(t-1)}, j \cdot \mathbf{v}\rangle \\
&\quad + \frac{\eta}{n} \cdot \bigg(\sum_{i=1}^{n} |\ell_{j,i}^{(t-1)}| \sigma'(\langle \mathbf{w}_{j,r^*}^{(t-1)}, y_i \mathbf{v}\rangle) - \alpha \sum_{i=1}^{n} |\ell_{j,i}^{(t-1)}| \sigma'(\langle \mathbf{w}_{j,r^*}^{(t-1)}, \boldsymbol{\xi}_i\rangle) \bigg) \\
&\geq (1 - \eta\lambda) \cdot \langle \mathbf{w}_{j,r^*}^{(t-1)}, j \cdot \mathbf{v}\rangle + \Theta(\eta) \cdot \big[\sigma'(\langle \mathbf{w}_{j,r^*}^{(t-1)}, j \cdot \mathbf{v}\rangle) - \alpha \sigma'(\Gamma_j^{(t-1)})\big] \\
&\geq (1 - \eta\lambda)\Lambda_j^{(t-1)} + \eta \cdot \Theta\big((\Lambda_j^{(t-1)})^{q-1}\big) - \eta \cdot \Theta\big(\alpha (\Gamma_j^{(t-1)})^{q-1}\big).
\end{aligned} \tag{6.7.29}$$

Similarly, let $r^* = \arg\max_r \langle \mathbf{w}_{y_i,r}^{(t)}, \boldsymbol{\xi}_i\rangle$, we also have the following according to (6.7.28)

$$\begin{aligned}
\Gamma_{y_i,i}^{(t)} &= \langle \mathbf{w}_{y_i,r^*}^{(t)}, \boldsymbol{\xi}_i\rangle \\
&\leq (1 - \eta\lambda)\langle \mathbf{w}_{y_i,r^*}^{(t-1)}, \boldsymbol{\xi}_i\rangle + \widetilde{\Theta}\bigg(\frac{\eta s \sigma_p^2}{n}\bigg) \cdot \sigma'(\langle \mathbf{w}_{y_i,r^*}^{(t-1)}, \boldsymbol{\xi}_i\rangle) + \Theta\bigg(\frac{\eta\alpha^2}{n}\bigg) \cdot \sum_{s=1}^{n} \sigma'(\langle \mathbf{w}_{y_i,r^*}^{(t-1)}, \boldsymbol{\xi}_s\rangle) \\
&\leq \Gamma_{y_i,i}^{(t-1)} + \widetilde{\Theta}\bigg(\frac{\eta s \sigma_p^2 \big(\Gamma_{y_i,i}^{(t-1)}\big)^{q-1}}{n}\bigg) + \Theta\bigg(\frac{\eta\alpha^2}{n} \cdot \sum_{s=1}^{n} \big(\Gamma_{y_i,s}^{(t-1)}\big)^{q-1}\bigg).
\end{aligned}$$

Then by our definition of $\Gamma_j^{(t)} = \max_{i\in[n]} \Gamma_{j,i}^{(t)}$, we further get the following for all $j \in \{-1, 1\}$,

$$\Gamma_j^{(t)} \leq \Gamma_j^{(t-1)} + \widetilde{\Theta}\bigg(\frac{\eta s \sigma_p^2 + n\eta\alpha^2}{n} \cdot \big(\Gamma_j^{(t-1)}\big)^{q-1}\bigg) = \Gamma_j^{(t-1)} + \Theta\bigg(\frac{\eta s \sigma_p^2}{n} \cdot \big(\Gamma_j^{(t-1)}\big)^{q-1}\bigg), \tag{6.7.30}$$

where the last equation is by our assumption that $\alpha = \widetilde{O}(s\sigma_p^2/n)$.

Then we will prove the main argument for general t, which is based on the following two induction hypotheses

$$\Lambda_j^{(t)} \geq \Lambda_j^{(t-1)} + \eta \cdot \Theta\big((\Lambda_j^{(t-1)})^{q-1}\big), \tag{6.7.31}$$

$$\Gamma_j^{(t)} \leq \Gamma_j^{(t-1)} + \Theta\bigg(\frac{\eta s \sigma_p^2}{n} \cdot \big(\Gamma_j^{(t-1)}\big)^{q-1}\bigg). \tag{6.7.32}$$

Note that when $t = 0$, we have already verified these two hypotheses in (6.7.29) and (6.7.30), where we use the fact that $\lambda = o(\sigma_0^{q-2}\sigma_p/n) \leq \big(\Lambda_j^{(0)}\big)^{q-2}$ and $\alpha = o(1)$. Suppose that (6.7.29) and (6.7.30) hold for iterations $\tau \leq t$. At time $t+1$, for all $\tau \leq t$, we have

$$\Gamma_j^{(\tau)} \leq O(\Lambda_j^{(\tau)}),$$

as $s\sigma^2/n = o(1)$ and $\Lambda_j^{(t)}$ increases faster than $\Gamma_j^{(t)}$. Besides, we can also show that $\lambda\Gamma_j^{(t)} \leq \big(\Gamma_j^{(t)}\big)^{q-1}$, which has been verified at time $t = 0$, since $\Gamma_j^{(t)}$ keeps increasing. Therefore, we have

$$\lambda\Gamma_j^{(t)} \leq \big(\Gamma_j^{(t)}\big)^{q-1} \leq O\big(\big(\Lambda_j^{(t)}\big)^{q-1}\big),$$

and hence (6.7.29) implies

$$\begin{aligned}\Lambda_j^{(t+1)} &\geq (1-\eta\lambda)\Lambda_j^{(t)} + \eta \cdot \Theta\big((\Lambda_j^{(t)})^{q-1}\big) - \eta \cdot \Theta\big(\alpha(\Gamma_j^{(t)})^{q-1}\big) \\ &\geq \Lambda_j^{(t)} + \eta \cdot \Theta\big((\Lambda_j^{(t)})^{q-1}\big),\end{aligned}$$

which verifies Hypothesis (6.7.31) at $t+1$. Additionally, (6.7.30) implies

$$\Gamma_j^{(t+1)} \leq \Gamma_j^{(t)} + \Theta\bigg(\frac{\eta s \sigma_p^2}{n} \cdot \big(\Gamma_j^{(t)}\big)^{q-1}\bigg),$$

which verifies Hypothesis (6.7.32) at $t+1$. Then by Lemma 6.7.9, we have that $\Lambda_j^{(t)} = \widetilde{O}(1)$ for all $t \leq T_0 = \widetilde{\Theta}\big((\Lambda_j^{(0)})^{2-q}/\eta\big) = \widetilde{\Theta}(\sigma_0^{2-q}/\eta)$. Moreover, Lemma 6.7.9 also shows that $\Gamma_j^{(t+1)} = O(\Lambda_j^{(0)}) = \widetilde{O}(\sigma_0)$. This completes the proof. □

Lemma 6.7.10. For all $i \in [n]$ and $t \leq T_{-y_i}$, it holds that $\langle \mathbf{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle \leq \widetilde{\Theta}(\alpha)$.

Proof. First of all, for $j \in \{\pm 1\}$, by the definition of T_j, we have

$$\langle \mathbf{w}_{j,r}^{(t)}, j \cdot \mathbf{v} \rangle \leq \widetilde{\Theta}(1).$$

Moreover, with the same proof as Lemma 6.7.8, it is clear that $-\langle \mathbf{w}_{j,r}^{(t)}, j \cdot \mathbf{v} \rangle$ is decreasing in t for $t \leq T_j$. Therefore, by the fact that $|\langle \mathbf{w}_{j,r}^{(0)}, \mathbf{v} \rangle| \leq \widetilde{\Theta}(1)$, we have

$$|\langle \mathbf{w}_{j,r}^{(t)}, \mathbf{v} \rangle| \leq \widetilde{\Theta}(1) \tag{6.7.33}$$

for all $t \leq T_j$.

Now by the update form of GD, we have for any $k \in \mathcal{B}_i$,

$$\mathbf{w}_{-y_i,r}^{(t+1)}[k] \cdot \boldsymbol{\xi}_i[k] = (1 - \eta\lambda) \cdot \mathbf{w}_{-y_i,r}^{(t)}[k] \cdot \boldsymbol{\xi}_i[k] + \frac{\eta}{n} \cdot \sum_{k \in \mathcal{B}_i} \ell_{-y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \cdot \boldsymbol{\xi}_i[k]^2.$$

Note that $\ell_{-y_i,i}^{(t)} \sigma'(\langle \mathbf{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle) < 0$, which implies that $\mathbf{w}_{-y_i,r}^{(t)}[k] \cdot \boldsymbol{\xi}_i[k]$ is decreasing in t. Therefore, for all r and i, we have

$$\begin{aligned}
\langle \mathbf{w}_{-y_i,r}^{(t)}, \boldsymbol{\xi}_i \rangle &= \mathbf{w}_{-y_i,r}^{(t)}[1] \cdot \boldsymbol{\xi}_i[1] + \sum_{k \in \mathcal{B}_i} \mathbf{w}_{-y_i,r}^{(t)}[k] \boldsymbol{\xi}_i[k] \\
&\leq \mathbf{w}_{-y_i,r}^{(t)}[1] \cdot \boldsymbol{\xi}_i[1] + \sum_{k \in \mathcal{B}_i} \mathbf{w}_{-y_i,r}^{(0)}[k] \boldsymbol{\xi}_i[k] \\
&\leq |\mathbf{w}_{-y_i,r}^{(t)}[1] \cdot \boldsymbol{\xi}_i[1]| + \bigg| \sum_{k \in \mathcal{B}_i} \mathbf{w}_{-y_i,r}^{(0)}[k] \boldsymbol{\xi}_i[k] \bigg| \\
&\leq \widetilde{\Theta}(\alpha) + \widetilde{\Theta}(\sigma_0 \sigma_p s^{1/2}) \\
&= \widetilde{\Theta}(\alpha),
\end{aligned}$$

where the third inequality follows by (6.7.33). This completes the proof. □

Note that for different j, the iteration numbers when $\Lambda_j^{(t)}$ reaches $\widetilde{\Theta}(1/m)$ are different. Without loss of generality, we can assume $T_1 \leq T_{-1}$. Lemma 6.7.8 has provided a clear understanding about how $\Lambda_j^{(t)}$ varies within the iteration range $[0, T_j]$. However, it remains unclear how $\Gamma_1^{(t)}$ varies within the iteration range $[T_1, T_{-1}]$ since in this period we no longer have $|\ell_{j,i}^{(t)}| = \Theta(1)$ and the effect of gradient descent on the feature learning (i.e., increase of

$\langle \mathbf{w}_{j,r}, j \cdot \mathbf{v} \rangle$) becomes weaker. In the following lemma we give a characterization of $\Lambda_1^{(t)}$ for every $t \in [T_1, T_{-1}]$.

Lemma 6.7.11 (Stage I of GD: part II)**.** Without loss of generality assuming $T_1 < T_{-1}$. Then it holds that $\Lambda_1^{(t)} = \widetilde{\Theta}(1)$ for all $t \in [T_1, T_{-1}]$.

Proof. Recall from (6.7.29) that we have the following general lower bound for the increase of $\Lambda_j^{(t)}$

$$\begin{aligned}\Lambda_j^{(t+1)} &\geq (1-\eta\lambda) \cdot \langle \mathbf{w}_{j,r^*}^{(t)}, j \cdot \mathbf{v} \rangle + \frac{\eta}{n} \cdot \bigg(\sum_{i=1}^n |\ell_{j,i}^{(t)}| \sigma'(\langle \mathbf{w}_{j,r^*}^{(t)}, y_i \mathbf{v} \rangle) - \alpha \sum_{i=1}^n |\ell_{j,i}^{(t)}| \sigma'(\langle \mathbf{w}_{j,r^*}^{(t)}, \boldsymbol{\xi}_i \rangle) \bigg) \\ &\geq (1-\eta\lambda)\Lambda_j^{(t)} + \Theta\bigg(\frac{\eta}{n}\bigg) \cdot \sum_{i:y_i=j} |\ell_{j,i}^{(t)}| \cdot \big(\Lambda_j^{(t)}\big)^{q-1} - \Theta(\alpha\eta) \cdot \big(\Gamma_j^{(t)} \vee \widetilde{\Theta}(\alpha)\big)^{q-1}, \qquad (6.7.34)\end{aligned}$$

where the last inequality is by Lemma 6.7.10. Note that by Lemma 6.7.8, we have $\Gamma_j^{(t)} = \widetilde{O}(\sigma_0)$ for all $t \leq T_{-1}$ and . Then the above inequality leads to

$$\Lambda_j^{(t+1)} \geq (1-\eta\lambda)\Lambda_j^{(t)} + \Theta\bigg(\frac{\eta}{n}\bigg) \cdot \sum_{i:y_i=j} |\ell_{j,i}^{(t)}| \cdot \big(\Lambda_j^{(t)}\big)^{q-1} - \Theta(\alpha^q \eta), \qquad (6.7.35)$$

where we use the fact that $\alpha = \omega(\sigma_0)$. The the remaining proof consists of two parts: (1) proving $\Lambda_j^{(t)} \geq \Theta(1/m) = \widetilde{\Theta}(1)$ and (2) $\Lambda_j^{(t)} \leq \Theta(\log(1/\lambda))$.

Without loss of generality we consider $j = 1$. Regarding the first part, we first note that Lemma 6.7.8 implies that $\Lambda_1^{(T_1)} \geq \Theta(1/m)$. Then we consider the case when $\Lambda_1^{(t)} \leq \Theta(\log(1/\alpha)/m)$, it holds that for all $y_i = 1$,

$$\begin{aligned}\ell_{1,i}^{(t)} &= \frac{e^{F_{-1}(\mathbf{W}^{(t)}, \mathbf{x}_i)}}{\sum_{j \in \{-1,1\}} e^{F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)}} \\ &= \exp\bigg(\Theta\bigg(\sum_{r=1}^m \big[\sigma(\langle \mathbf{w}_{-1,r}^{(t)}, y_i \mathbf{v} \rangle) + \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}_i \rangle)\big] - \sum_{r=1}^m \big[\sigma(\langle \mathbf{w}_{1,r}^{(t)}, y_i \mathbf{v} \rangle) + \sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i \rangle)\big] \bigg) \bigg) \\ &\geq \exp\big(-\Theta(m\Lambda_1^{(t)}) \big) \\ &\geq \exp(-\Theta(\log(1/\alpha))) \\ &= \widetilde{\Theta}(\alpha).\end{aligned}$$

Then (6.7.35) implies that if $\Gamma_1^{(t)} \leq \Theta(\log(1/\sigma_0)/m)$, we have

$$\Lambda_1^{(t+1)} \geq (1-\eta\lambda)\Lambda_1^{(t)} + \Theta(\eta\alpha)\cdot\Lambda_1^{(t)} - \Theta(\alpha^q\eta) \geq \Lambda_1^{(t)} + \Theta(\eta\alpha)\cdot\Lambda_1^{(t)} \geq \Lambda_1^{(t)},$$

where the second inequality is due to $\lambda = o(\alpha)$. This implies that $\Lambda_1^{(t)}$ will keep increases in this case so that it is impossible that $\Lambda_1^{(t)} \leq \Theta(1/m)$, which completes the proof of the first part.

For the second part, (6.7.28) implies that

$$\Lambda_1^{(t+1)} \leq (1-\eta\lambda)\Lambda_1^{(t)} + \Theta\bigg(\frac{\eta}{n}\bigg)\cdot\sum_{i:y_i=1}|\ell_{1,i}^{(t)}|\cdot\big(\Lambda_1^{(t)}\big)^{q-1}. \tag{6.7.36}$$

Consider the case when $\Gamma_1^{(t)} \geq \Theta(\log(d))$, then for all $y_i = 1$,

$$\begin{aligned}
\ell_{1,i}^{(t)} &= \frac{e^{F_{-1}(\mathbf{W}^{(t)},\mathbf{x}_i)}}{\sum_{j\in\{-1,1\}}e^{F_j(\mathbf{W}^{(t)},\mathbf{x}_i)}} \\
&= \exp\bigg(\Theta\bigg(\sum_{r=1}^m\big[\sigma(\langle\mathbf{w}_{-1,r}^{(t)}, y_i\mathbf{v}\rangle) + \sigma(\langle\mathbf{w}_{-1,r}^{(t)},\boldsymbol{\xi}_i\rangle)\big] - \sum_{r=1}^m\big[\sigma(\langle\mathbf{w}_{1,r}^{(t)}, y_i\mathbf{v}\rangle) + \sigma(\langle\mathbf{w}_{1,r}^{(t)},\boldsymbol{\xi}_i\rangle)\big]\bigg)\bigg) \\
&\leq \exp\big(-\Theta(\Lambda_1^{(t)})\big) \\
&\leq \exp(-\Theta(\log(1/\lambda)) \\
&= \widetilde{\Theta}(\mathrm{poly}(\lambda)).
\end{aligned}$$

Then (6.7.36) further implies that

$$\begin{aligned}
\Lambda_1^{(t+1)} &\leq (1-\eta\lambda)\Lambda_1^{(t)} + \Theta\bigg(\frac{\eta}{\mathrm{poly}(d)}\bigg)\cdot\big(\Lambda_1^{(t)}\big)^{q-1} \\
&\leq \Lambda_1^{(t)} - \Theta\big(\eta\Lambda_1^{(t)}\big)\cdot\bigg(\lambda - \mathrm{poly}(\lambda)\cdot\big(\Lambda_1^{(t)}\big)^{q-2}\bigg) \leq \Lambda_1^{(t)},
\end{aligned}$$

which implies that $\Lambda_1^{(t)}$ will decrease. As a result, we can conclude that $\lambda_1^{(t)}$ will not exceed $\Theta(\log(1/\lambda))$, this completes the proof of the second part.

□

Lemma 6.7.12 (Lemma 6.4.7, restated)**.** If $\eta \leq O(\sigma_0)$, it holds that $\Lambda_j^{(t)} = \widetilde{\Theta}(1)$ and $\Gamma_j^{(t)} = \widetilde{O}(\sigma_0)$ for all $t \in [T_{-1}, T]$.

Proof. We will prove the desired argument based on the following three induction hypothesis:

$$\Lambda_j^{(t+1)} \geq (1-\lambda\eta)\Lambda_j^{(t)} + \widetilde{\Theta}\left(\frac{\eta}{n}\right)\sum_{i:y_i=j}|\ell_{j,i}^{(t)}| - \widetilde{\Theta}(\alpha^q\eta)\cdot\frac{1}{n}\sum_{i=1}^{n}|\ell_{j,r}^{(t)}|, \tag{6.7.37}$$

$$\Gamma_j^{(t)} = \widetilde{O}(\sigma_0), \tag{6.7.38}$$

$$\Lambda_j^{(t)} = \widetilde{\Theta}(1). \tag{6.7.39}$$

In terms of Hypothesis (6.7.37), we can apply Hypothesis (6.7.38) and (6.7.39) to (6.7.34) and get that

$$\begin{aligned}\Lambda_j^{(t+1)} &\geq (1-\eta\lambda)\Lambda_j^{(t)} + \Theta\left(\frac{\eta}{n}\right)\cdot\sum_{i:y_i=j}|\ell_{j,i}^{(t)}|\cdot\left(\Lambda_j^{(t)}\right)^{q-1} - \Theta(\alpha\eta)\cdot\left(\Gamma_j^{(t)}\vee\widetilde{\Theta}(\alpha)\right)^{q-1}\cdot\frac{1}{n}\sum_{i=1}^{n}|\ell_{j,r}^{(t)}| \\ &\geq (1-\lambda\eta)\Lambda_j^{(t)} + \widetilde{\Theta}\left(\frac{\eta}{n}\right)\sum_{i:y_i=j}|\ell_{j,i}^{(t)}| - \widetilde{\Theta}(\alpha^q\eta)\cdot\frac{1}{n}\sum_{i=1}^{n}|\ell_{j,r}^{(t)}|.\end{aligned}$$

where the last inequality we use the fact that $\alpha \geq \sigma_0$. This verifies Hypothesis (6.7.37).

In order to verify Hypothesis (6.7.38), we have the following according to (6.7.37),

$$\begin{aligned}\sum_{j\in\{-1,1\}}\Lambda_j^{(t+1)} &\geq (1-\lambda\eta)\sum_{j\in\{-1,1\}}\left[\Lambda_j^{(t)} + \widetilde{\Theta}\left(\frac{\eta}{n}\right)\sum_{i=1}^{n}|\ell_{j,i}^{(t)}| - \widetilde{\Theta}(\alpha^q\eta)\cdot\frac{1}{n}\sum_{i=1}^{n}|\ell_{j,r}^{(t)}|\right] \\ &= (1-\lambda\eta)\sum_{j\in\{-1,1\}}\left[\Lambda_j^{(t)} + \widetilde{\Theta}\left(\frac{\eta}{n}\right)\sum_{i=1}^{n}|\ell_{j,i}^{(t)}|\right],\end{aligned}$$

where the last equality holds since $\alpha = o(1)$. Recursively applying the above inequality from T_{-1} to t gives

$$\sum_{j\in\{-1,1\}}\Lambda_j^{(t)} \geq (1-\lambda\eta)^{t-T_{-1}}\sum_{j\in\{-1,1\}}\left[\Lambda_j^{(T_{-1})} + \widetilde{\Theta}\left(\frac{\eta}{n}\right)\cdot\sum_{\tau=0}^{t-T_{-1}-1}(1-\lambda\eta)^{\tau}\sum_{i=1}^{n}|\ell_{j,i}^{(t-1-\tau)}|\right].$$

Then by Hypothesis (6.7.39) we have

$$\widetilde{\Theta}\left(\frac{\eta}{n}\right)\cdot\sum_{\tau=0}^{t-T_{-1}-1}(1-\lambda\eta)^{\tau}\sum_{i=1}^{n}|\ell_{j,i}^{(t-1-\tau)}| \leq \widetilde{\Theta}(1).$$

Now let us look at the rate of memorizing noises. By (6.7.28) and use the fact that $\alpha^2 \leq O(s\sigma_p^2/n)$, we have

$$\begin{aligned}
\Gamma_j^{(t)} &\leq (1-\eta\lambda)\Gamma_j^{(t-1)} + \widetilde{\Theta}\bigg(\frac{\eta s\sigma_p^2}{n}\bigg) \cdot \sum_{i=1} |\ell_{j,i}| \cdot \big(\Gamma_j^{(t-1)}\big)^{q-1} \\
&\leq (1-\eta\lambda)\Gamma_j^{(t-1)} + \widetilde{\Theta}\bigg(\frac{\eta s\sigma_p^2\sigma_0^{q-1}}{n}\bigg) \cdot \sum_{i=1} |\ell_{j,i}| \\
&\leq \Gamma_j^{(T_{-1})} + \widetilde{\Theta}\bigg(\frac{\eta s\sigma_p^2\sigma_0^{q-1}}{n}\bigg) \cdot \sum_{\tau=0}^{t-T_{-1}-1} (1-\lambda\eta)^\tau \sum_{i=1}^{n} |\ell_{j,i}^{(t-1-\tau)}| \\
&\leq \widetilde{\Theta}\big(\sigma_0 + s\sigma_p^2\sigma_0^{q-1}\big) \\
&\leq \widetilde{\Theta}(\sigma_0),
\end{aligned}$$

which verifies Hypothesis (6.7.38).

Given Hypothesis (6.7.37) and (6.7.38), the verification of (6.7.39) is straightforward by applying the same proof technique of Lemma 6.7.11 and thus we omit it here. □

Lemma 6.7.13 (Lemma 6.4.8, restated)**.** If the step size satisfies, then for any $t \geq T_{-1}$ it holds that

$$L(\mathbf{W}^{(t+1)}) - L(\mathbf{W}^{(t)}) \leq -\frac{\eta}{2}\|\nabla L(\mathbf{W}^{(t)})\|_F^2.$$

Proof. The proof of this lemma is similar to that of Lemma 6.7.6, which is basically relying the smoothness property of the loss function $L(\mathbf{W})$ given certain constraints on the inner products $\langle \mathbf{w}_{j,r}, \mathbf{v}\rangle$ and $\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i\rangle$.

Let $\Delta F_{j,i} = F_j(\mathbf{W}^{(t+1)}, \mathbf{x}_i) - F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)$, we can get the following Taylor expansion on the loss function $L_i(\mathbf{W}^{(t+1)})$,

$$L_i(\mathbf{W}^{(t+1)}) - L_i(\mathbf{W}^{(t)}) \leq \sum_j \frac{\partial L_i(\mathbf{W}^{(t)})}{\partial F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)} \cdot \Delta F_{j,i} + \sum_j (\Delta F_{j,i})^2. \tag{6.7.40}$$

In particular, by Lemma 6.7.12, we know that $\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{v}\rangle \leq \widetilde{\Theta}(1)$ and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle \leq \widetilde{\Theta}(\sigma_0) \leq \widetilde{\Theta}(1)$. Then similar to (6.7.21), we can apply first-order Taylor expansion to $F_j(\mathbf{W}^{(t+1)}, \mathbf{x}_i)$,

which requires to characterize the second-order error of the Taylor expansions on $\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, y_i\mathbf{v}\rangle)$ and $\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle)$,

$$
\begin{aligned}
&\big|\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, y_i\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{v}\rangle) - \big\langle \nabla_{\mathbf{w}_{j,r}}\sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{v}\rangle), \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\big\rangle\big| \\
&\leq \widetilde{\Theta}\big(\|\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\|_2^2\big) = \widetilde{\Theta}(\eta^2\|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})\|_2^2), \\
&\big|\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle) - \big\langle \nabla_{\mathbf{w}_{j,r}}\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle), \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\big\rangle\big| \\
&\leq \widetilde{\Theta}\big(\|\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\|_2^2\big) = \widetilde{\Theta}(\eta^2\|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})\|_2^2).
\end{aligned} \tag{6.7.41}
$$

Then combining the above bounds for every $r \in [m]$, we can get the following bound for $\Delta F_{j,i}$

$$
\begin{aligned}
\big|\Delta F_{j,i} - \langle \nabla_{\mathbf{W}} F_j(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)}\rangle\big| &\leq \widetilde{\Theta}\bigg(\eta^2 \sum_{r\in[m]} \|\nabla_{\mathbf{w}_{j,r}} L(\mathbf{W}^{(t)})\|_2^2\bigg) \\
&= \widetilde{\Theta}\big(\eta^2\|\nabla L(\mathbf{W}^{(t)})\|_F^2\big).
\end{aligned} \tag{6.7.42}
$$

Moreover, since $\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{v}\rangle \leq \widetilde{\Theta}(1)$ and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle \leq \widetilde{\Theta}(1)$ and $\sigma(\cdot)$ is convex, then we have

$$
\begin{aligned}
|\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, y_i\mathbf{v}\rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{v}\rangle)| &\leq \max\big\{|\sigma'(\langle \mathbf{w}_{j,r}^{(t+1)}, y_i\mathbf{v}\rangle)|, |\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i\mathbf{v}\rangle)|\big\} \cdot |\langle \mathbf{v}, \mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\rangle| \\
&\leq \widetilde{\Theta}\big(\|\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\|_2\big).
\end{aligned}
$$

Similarly we also have

$$
|\sigma(\langle \mathbf{w}_{j,r}^{(t+1)}, \boldsymbol{\xi}_i\rangle) - \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)| \leq \widetilde{\Theta}\big(\|\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\|_2\big).
$$

Combining the above inequalities for every $r \in [m]$, we have

$$
\big|\Delta F_{j,i}\big|^2 \leq \widetilde{\Theta}\bigg(\bigg[\sum_{r\in[m]} \|\mathbf{w}_{j,r}^{(t+1)} - \mathbf{w}_{j,r}^{(t)}\|_2\bigg]^2\bigg) \leq \widetilde{\Theta}\big(m\eta^2\|\nabla L(\mathbf{W}^{(t)})\|_F^2\big) = \widetilde{\Theta}\big(\eta^2\|\nabla L(\mathbf{W}^{(t)})\|_F^2\big). \tag{6.7.43}
$$

Now we can plug (6.7.42) and (6.7.43) into (6.7.40), which gives

$$
\begin{aligned}
L_i(\mathbf{W}^{(t+1)}) - L_i(\mathbf{W}^{(t)}) &\leq \sum_j \frac{\partial L_i(\mathbf{W}^{(t)})}{\partial F_j(\mathbf{W}^{(t)}, \mathbf{x}_i)} \cdot \Delta F_{j,i} + \sum_j (\Delta F_{j,i})^2 \\
&= \langle \nabla L_i(\mathbf{W}^{(t)}), \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)}\rangle + \widetilde{\Theta}(\eta^2\|\nabla L(\mathbf{W}^{(t)})\|_F^2).
\end{aligned} \tag{6.7.44}
$$

Taking sum over $i \in [n]$ and applying the smoothness property of the regularization function $\lambda\|\mathbf{W}\|_F^2$, we can get

$$\begin{aligned}
L(\mathbf{W}^{(t+1)}) - L(\mathbf{W}^{(t)}) &= \frac{1}{n}\sum_{i=1}^{n}\big[L_i(\mathbf{W}^{(t+1)}) - L_i(\mathbf{W}^{(t)})\big] + \lambda\big(\|\mathbf{W}^{(t+1)}\|_F^2 - \|\mathbf{W}^{(t)}\|_F^2\big) \\
&\leq \langle \nabla L(\mathbf{W}^{(t)}), \mathbf{W}^{(t+1)} - \mathbf{W}^{(t)}\rangle + \widetilde{\Theta}(\eta^2\|\nabla L(\mathbf{W}^{(t)})\|_F^2) \\
&= -\big(\eta - \widetilde{\Theta}(\eta^2)\big)\cdot\|\nabla L(\mathbf{W}^{(t)})\|_F^2 \\
&\leq -\frac{\eta}{2}\|\nabla L(\mathbf{W}^{(t)})\|_F^2,
\end{aligned}$$

where the last inequality is due to our choice of step size $\eta = o(1)$ so that gives $\eta - \widetilde{\Theta}(\eta^2) \geq \eta/2$. This completes the proof. □

Lemma 6.7.14 (Generalization Performance of GD)**.** Let

$$\mathbf{W}^* = \arg\min_{\{\mathbf{W}^{(1)},\ldots,\mathbf{W}^{(T)}\}} \|\nabla L(\mathbf{W}^{(t)})\|_F.$$

Then for all training data, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\big[F_{y_i}(\mathbf{W}^*, \mathbf{x}_i) \leq F_{-y_i}(\mathbf{W}^*, \mathbf{x}_i)\big] = 0.$$

Moreover, in terms of the test data $(\mathbf{x}, y) \sim \mathcal{D}$, we have

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[F_y(\mathbf{W}^*, \mathbf{x}) \leq F_{-y}(\mathbf{W}^*, \mathbf{x})\big] = o(1).$$

Proof. By Lemma 6.7.12 it is clear that all training data can be correctly classified so that the training error is zero. Besides, for test data $(\mathbf{x}, y)$ with $\mathbf{x} = [y\mathbf{v}^\top, \boldsymbol{\xi}^\top]^\top$, it is clear that with high probability $\langle \mathbf{w}_{y,r}^*, y\mathbf{v}\rangle = \widetilde{\Theta}(1)$ and $[\langle \mathbf{w}_{y,r}^*, \boldsymbol{\xi}\rangle]_+ \leq \widetilde{O}(\sigma_0)$, then

$$F_y(\mathbf{W}^*, \mathbf{x}) = \sum_{r=1}^{m}\big[\sigma(\langle \mathbf{w}_{y,r}^*, y\mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_{y,r}^*, \boldsymbol{\xi}\rangle)\big] \geq \widetilde{\Omega}(1).$$

If $j = -y$, we have with probability at least $1 - 1/\mathrm{poly}(n)$, $\langle \mathbf{w}_{-y,r}^*, y\mathbf{v}\rangle \leq 0$ and $[\mathbf{w}_{-y,r}^*, \boldsymbol{\xi}\rangle]_+ \leq \widetilde{O}(\alpha)$, which leads to

$$F_{-y}(\mathbf{W}^*, \mathbf{x}) = \sum_{r=1}^{m}\big[\sigma(\langle \mathbf{w}_{-y,r}^*, y\mathbf{v}\rangle) + \sigma(\langle \mathbf{w}_{-y,r}^*, \boldsymbol{\xi}\rangle)\big] \leq \widetilde{O}(m\alpha^q) = \widetilde{O}(\alpha^q) = o(1).$$

This implies that GD can also achieve nearly at most $1/\mathrm{poly}(n)$ test error. This completes the proof. □

6.8 Proof of Theorem 6.3.2: Convex Case

Theorem 6.8.1 (Convex setting, restated)**.** Assume the model is over-parameterized. Then for any convex and smooth training objective with positive regularization parameter λ, suppose we run **Adam** and **gradient descent** for $T = \frac{\mathrm{poly}(n)}{\eta}$ iterations, then with probability at least $1-n^{-1}$, the obtained parameters $\mathbf{W}^*_{\mathrm{Adam}}$ and $\mathbf{W}^*_{\mathrm{GD}}$ satisfy that $\|\nabla L(\mathbf{W}^*_{\mathrm{Adam}})\|_1 \leq \frac{1}{T\eta}$ and $\|\nabla L(\mathbf{W}^*_{\mathrm{Adam}})\|_2^2 \leq \frac{1}{T\eta}$ respectively. Moreover, it holds that:

- Training errors are the same:
$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big[\mathrm{sgn}\big(F(\mathbf{W}^*_{\mathrm{Adam}}, \mathbf{x}_i)\big) \neq y_i\big] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big[\mathrm{sgn}\big(F(\mathbf{W}^*_{\mathrm{GD}}, \mathbf{x}_i)\big) \neq y_i\big].$$
- Test errors are nearly the same:
$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\mathrm{sgn}\big(F(\mathbf{W}^*_{\mathrm{Adam}}, \mathbf{x}_i)\big) \neq y\big] = \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[\mathrm{sgn}\big(F(\mathbf{W}^*_{\mathrm{GD}}, \mathbf{x})\big) \neq y\big] \pm o(1).$$

Proof. The proof is straightforward by applying the same proof technique used for Lemmas 6.7.6 and 6.7.13, where we only need to use the smoothness property of the loss function. Then it is clear that both Adam and GD can provably find a point with sufficiently small gradient. Note that the training objective becomes strongly convex when adding weight decay regularization, implying that the entire training objective only has one stationary point, i.e., point with sufficiently small gradient. This further imply that the points found by Adam and GD must be exactly same and thus GD and Adam must have nearly same training and test performance.

Besides, when the problem is sufficiently over-parameterized, with proper regularization (feasibly small), we can still guarantee zero training errors. □

6.9 Conclusions

In this paper, we study the generalization of Adam and compare it with gradient descent. We show that when training neural networks, Adam and GD starting from the same initialization

can converge to different global solutions of the training objective with significantly different generalization errors, even with proper regularization. Our analysis reveals the fundamental difference between Adam and GD in learning features or noise, and demonstrates that this difference is closely tied to the nonconvex landscape of neural networks.

We would also like to remark several important research directions. First, our current result is for two-layer networks. Extending the results to deep networks could be an important next step, where we will not only look at the input data but also consider the output of each intermediate layer as "input". Second, our current data model is motivated by the image data (i.e., sparse feature and denser noise), where Adam has been observed to perform worse than SGD in terms of generalization. In fact, our theoretical analysis can lead to an opposite conclusion on the generalization comparison between Adam and GD if the noise is sparse and feature is denser. Therefore, it would also be interesting to explore whether this is the case in other machine learning tasks such as natural language processing, where Adam is often observed to perform better than SGD.

CHAPTER 7

Conclusions

This dissertation provided a theoretical analysis towards the role of optimization algorithms in learning over-parameterized models. In the first part, we developed a novel analysis to characterize the generalization ability of SGD for learning over-parameterized linear regression problems. We provided sharp problem-dependent upper bounds on the excess risk and demonstrated its tightness by proving a matching lower bound. Based on the developed bounds, we are able to indicate the problem instance than can be well learned by SGD. By comparing the generalization error of SGD to that achieved by ridge regression, we also partially explains the implicit regularization effect of SGD.

In the second part, we investigated the optimization and generalization for neural network models. We developed the state-of-the-art optimization guarantees under general conditions on the data distribution and established the generalization guarantees for GD and SGD under certain separation conditions on the data distribution. Lastly, we studied the generalization performance achieved by different optimization algorithms and provided a theoretical explanation towards the empirical generalization gap between Adam and GD in image classification problem.

Bibliography

[ACH18] Sanjeev Arora, Nadav Cohen, and Elad Hazan. "On the optimization of deep networks: Implicit acceleration by overparameterization." In *International Conference on Machine Learning*, pp. 244–253. PMLR, 2018.

[ADH19a] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks." In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

[ADH19b] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. "On exact computation with an infinitely wide neural net." In *Advances in Neural Information Processing Systems*, 2019.

[ADT20] Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. "The implicit regularization of stochastic gradient flow for least squares." In *International Conference on Machine Learning*, pp. 233–244. PMLR, 2020.

[AGC19] Sanjeev Arora, Nadav Golowich, Noah Cohen, and Wei Hu. "A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks." In *International Conference on Learning Representations*, 2019.

[AKT19] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. "A continuous-time view of early stopping for least squares regression." In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1370–1378. PMLR, 2019.

[AL20] Zeyuan Allen-Zhu and Yuanzhi Li. "Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning." *arXiv preprint arXiv:2012.09816*, 2020.

[ALL19] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. "Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers." In *Advances in Neural Information Processing Systems*, 2019.

[ALS19a] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

[ALS19b] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "On the convergence rate of training recurrent neural networks." *Advances in neural information processing systems*, **32**, 2019.

[Bah67] R. R. Bahadur. "Rates of Convergence of Estimates and Test Statistics." *Annals of Mathematical Statistics*, **38**:303–324, 1967.

[Bah71] R. R. Bahadur. *Some Limit Theorems in Statistics*. Society for Industrial and Applied Mathematics, 1971.

[BBG20] Raphaël Berthier, Francis Bach, and Pierre Gaillard. "Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model." *Advances in Neural Information Processing Systems*, **33**:2576–2586, 2020.

[BFT17] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. "Spectrally-normalized margin bounds for neural networks." In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

[BG17] Alon Brutzkus and Amir Globerson. "Globally optimal gradient descent for a convnet with gaussian inputs." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 605–614. JMLR. org, 2017.

[BGM18] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. "SGD

Learns Over-parameterized Networks that Provably Generalize on Linearly Separable Data." In *International Conference on Learning Representations*, 2018.

[BH18] Lukas Balles and Philipp Hennig. "Dissecting adam: The sign, magnitude and variance of stochastic gradients." In *International Conference on Machine Learning*, pp. 404–413. PMLR, 2018.

[BHL18] Peter Bartlett, Dave Helmbold, and Phil Long. "Gradient descent with identity initialization efficiently learns positive definite linear transformations." In *International Conference on Machine Learning*, pp. 520–529, 2018.

[BHX20] Mikhail Belkin, Daniel Hsu, and Ji Xu. "Two models of double descent for weak features." *SIAM Journal on Mathematics of Data Science*, **2**(4):1167–1180, 2020.

[BL19] Yu Bai and Jason D Lee. "Beyond Linearization: On Quadratic and Higher-Order Approximation of Wide Neural Networks." In *International Conference on Learning Representations*, 2019.

[BLL20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression." *Proceedings of the National Academy of Sciences*, 2020.

[BM02] Peter L Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results." *Journal of Machine Learning Research*, **3**(Nov):463–482, 2002.

[BM13] Francis Bach and Eric Moulines. "Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$." *Advances in neural information processing systems*, **26**:773–781, 2013.

[BR89] Avrim Blum and Ronald L Rivest. "Training a 3-node neural network is NP-

complete." In *Advances in neural information processing systems*, pp. 494–501, 1989.

[BWA18] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. "signSGD: Compressed optimisation for non-convex problems." In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018.

[CB18] Lenaic Chizat and Francis Bach. "A note on lazy training in supervised differentiable programming." *arXiv preprint arXiv:1812.07956*, 2018.

[CCG04] Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. "On the generalization ability of on-line learning algorithms." *IEEE Transactions on Information Theory*, **50**(9):2050–2057, 2004.

[CCZ21] Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. "How Much Over-parameterization Is Sufficient to Learn Deep ReLU Networks?" In *International Conference on Learning Representations*, 2021.

[CG19] Yuan Cao and Quanquan Gu. "Generalization Bounds of Stochastic Gradient Descent for Wide and Deep Neural Networks." In *Advances in Neural Information Processing Systems*, 2019.

[CG20] Yuan Cao and Quanquan Gu. "Generalization error bounds of gradient descent for learning over-parameterized deep relu networks." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3349–3356, 2020.

[CHM15] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. "The loss surfaces of multilayer networks." In *Artificial Intelligence and Statistics*, pp. 192–204, 2015.

[CL21] Niladri S Chatterji and Philip M Long. "Finite-sample Analysis of Interpolat-

ing Linear Classifiers in the Overparameterized Regime." *Journal of Machine Learning Research*, **22**:1–30, 2021.

[CLT20] Xi Chen, Qiang Liu, and Xin T Tong. "Dimension Independent Generalization Error with Regularized Online Optimization." *arXiv preprint arXiv:2003.11196*, 2020.

[CSN19] Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. "On empirical comparisons of optimizers for deep learning." *arXiv preprint arXiv:1910.05446*, 2019.

[CZT20] Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. "Closing the Generalization Gap of Adaptive Gradient Methods in Training Deep Neural Networks." In *International Joint Conferences on Artificial Intelligence*, 2020.

[DB15a] Alexandre Défossez and Francis Bach. "Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions." In *Artificial Intelligence and Statistics*, pp. 205–213, 2015.

[DB15b] Aymeric Dieuleveut and Francis R. Bach. "Non-parametric Stochastic Approximation with Large Step sizes." *The Annals of Statistics*, 2015.

[DBB20] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. "A simple convergence proof of adam and adagrad." *arXiv preprint arXiv:2003.02395*, 2020.

[DFB17] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. "Harder, better, faster, stronger convergence rates for least-squares regression." *The Journal of Machine Learning Research*, **18**(1):3520–3570, 2017.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods

for online learning and stochastic optimization." *Journal of Machine Learning Research*, **12**(Jul):2121–2159, 2011.

[DLL19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. "Gradient descent finds global minima of deep neural networks." In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

[DLT18] Simon S Du, Jason D Lee, and Yuandong Tian. "When is a Convolutional Filter Easy to Learn?" In *International Conference on Learning Representations*, 2018.

[DZP18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. "Gradient Descent Provably Optimizes Over-parameterized Neural Networks." In *International Conference on Learning Representations*, 2018.

[FCG19] Spencer Frei, Yuan Cao, and Quanquan Gu. "Algorithm-Dependent Generalization Bounds for Overparameterized Deep Residual Networks." In *Advances in Neural Information Processing Systems*, pp. 14769–14779, 2019.

[FHT01] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[GLS18a] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. "Characterizing implicit bias in terms of optimization geometry." In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

[GLS18b] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. "Implicit bias of gradient descent on linear convolutional networks." *Advances in Neural Information Processing Systems*, **31**, 2018.

[HDY12] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. "Deep neural networks for acoustic modeling in speech recognition:

The shared views of four research groups." *IEEE Signal Processing Magazine*, **29**(6):82–97, 2012.

[HKZ12] Daniel Hsu, Sham M Kakade, and Tong Zhang. "Random design analysis of ridge regression." In *Conference on learning theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012.

[HKZ14] Daniel J. Hsu, Sham M. Kakade, and Tong Zhang. "Random Design Analysis of Ridge Regression." *Foundations of Computational Mathematics*, **14**(3):569–600, 2014.

[HM16] Moritz Hardt and Tengyu Ma. "Identity matters in deep learning." *arXiv preprint arXiv:1611.04231*, 2016.

[HMR22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. "Surprises in high-dimensional ridgeless least squares interpolation." *The Annals of Statistics*, **50**(2):949–986, 2022.

[Hor91] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks." *Neural networks*, **4**(2):251–257, 1991.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory." *Neural computation*, **9**(8):1735–1780, 1997.

[HSS12] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.", 2012.

[HZR15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of CVPR*, pp. 770–778, 2016.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks." In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

[JKK18a] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford. "A Markov Chain Theory Approach to Characterizing the Minimax Optimality of Stochastic Gradient Descent (for Least Squares)." In *37th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, 2018.

[JKK18b] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. "Accelerating Stochastic Gradient Descent for Least Squares Regression." In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*. PMLR, 2018.

[JNK17] Prateek Jain, Praneeth Netrapalli, Sham M Kakade, Rahul Kidambi, and Aaron Sidford. "Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification." *The Journal of Machine Learning Research*, **18**(1):8258–8299, 2017.

[JT19] Ziwei Ji and Matus Telgarsky. "The implicit bias of gradient descent on nonseparable data." In *Conference on Learning Theory*, pp. 1772–1798. PMLR, 2019.

[JT20] Ziwei Ji and Matus Telgarsky. "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks." In *International Conference on Learning Representations*, 2020.

[Kaw16] Kenji Kawaguchi. "Deep learning without poor local minima." In *Advances in Neural Information Processing Systems*, pp. 586–594, 2016.

[KB15] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." *International Conference on Learning Representations*, 2015.

[KH19] Kenji Kawaguchi and Jiaoyang Huang. "Gradient descent finds global minima for generalizable deep neural networks of practical sizes." In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 92–99. IEEE, 2019.

[KLS20] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. "The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization." *Journal of Machine Learning Research*, **21**(169):1–16, 2020.

[KMN16] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. "On large-batch training for deep learning: Generalization gap and sharp minima." *arXiv preprint arXiv:1609.04836*, 2016.

[Kri09] Alex Krizhevsky. "Learning multiple layers of features from tiny images." Technical report, Citeseer, 2009.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[LBB98] 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.

[LJ18] Hongzhou Lin and Stefanie Jegelka. "ResNet with one-neuron hidden layers is a Universal Approximator." In *Advances in Neural Information Processing Systems*, pp. 6172–6181, 2018.

[LL18] Yuanzhi Li and Yingyu Liang. "Learning overparameterized neural networks via stochastic gradient descent on structured data." *Advances in Neural Information Processing Systems*, **31**, 2018.

[LMZ20] Yuanzhi Li, Tengyu Ma, and Hongyang R Zhang. "Learning over-parametrized two-layer neural networks beyond ntk." In *Conference on Learning Theory*, pp. 2613–2682. PMLR, 2020.

[LPA20] Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. "Bad global minima exist and sgd can reach them." *Advances in Neural Information Processing Systems*, **33**:8543–8552, 2020.

[LS18] Chandrashekar Lakshminarayanan and Csaba Szepesvari. "Linear stochastic approximation: How far does constant step-size and iterate averaging go?" In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355, 2018.

[LXL18] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. "Adaptive Gradient Methods with Dynamic Bound of Learning Rate." In *International Conference on Learning Representations*, 2018.

[LXS19] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. "Wide neural networks of any depth evolve as linear models under gradient descent." In *Advances in Neural Information Processing Systems*, 2019.

[LY17] Yuanzhi Li and Yang Yuan. "Convergence analysis of two-layer neural networks with relu activation." *Advances in neural information processing systems*, **30**, 2017.

[MNS21] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. "Classification vs regression in overparameterized

regimes: Does the loss function matter?" *Journal of Machine Learning Research*, **22**(222):1–69, 2021.

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[NKB21] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. "Deep double descent: Where bigger models and more data hurt." *Journal of Statistical Mechanics: Theory and Experiment*, **2021**(12):124003, 2021.

[NS19] Atsushi Nitanda and Taiji Suzuki. "Refined Generalization Analysis of Gradient Descent for Over-parameterized Two-layer Neural Networks with Smooth Activations on Classification Problems." *arXiv preprint arXiv:1905.09870*, 2019.

[NTS14] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. "In search of the real inductive bias: On the role of implicit regularization in deep learning." *arXiv preprint arXiv:1412.6614*, 2014.

[NVK20] Preetum Nakkiran, Prayaag Venkat, Sham M Kakade, and Tengyu Ma. "Optimal Regularization can Mitigate Double Descent." In *International Conference on Learning Representations*, 2020.

[OS19] Samet Oymak and Mahdi Soltanolkotabi. "Towards moderate overparameterization: global convergence guarantees for training shallow neural networks." *arXiv preprint arXiv:1902.04674*, 2019.

[PJ92] Boris T Polyak and Anatoli B Juditsky. "Acceleration of stochastic approximation by averaging." *SIAM journal on control and optimization*, **30**(4):838–855, 1992.

[PRE17] Vardan Papyan, Yaniv Romano, and Michael Elad. "Convolutional neural net-

works analyzed via convolutional sparse coding." *The Journal of Machine Learning Research*, **18**(1):2887–2938, 2017.

[RKK18] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. "On the convergence of adam and beyond." In *International Conference on Learning Representations*, 2018.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[Sha21] Ohad Shamir. "Gradient methods never overfit on separable data." *Journal of Machine Learning Research*, **22**(85):1–20, 2021.

[SHM16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. "Mastering the game of Go with deep neural networks and tree search." *Nature*, **529**(7587):484–489, 2016.

[SPR18] Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. "Connecting optimization and regularization paths." *Advances in Neural Information Processing Systems*, **31**:10608–10619, 2018.

[SSB02] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[TB20] Alexander Tsigler and Peter L Bartlett. "Benign overfitting in ridge regression." *arXiv preprint arXiv:2009.14286*, 2020.

[Tel16] Matus Telgarsky. "Benefits of depth in neural networks." In *Conference on learning theory*, pp. 1517–1539. PMLR, 2016.

[Tia17] Yuandong Tian. "An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis." In *International conference on machine learning*, pp. 3404–3413. PMLR, 2017.

[Tih63] Andrei Nikolajevits Tihonov. "Solution of incorrectly formulated problems and the regularization method." *Soviet Math.*, **4**:1035–1038, 1963.

[Ver10] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices." *arXiv preprint arXiv:1011.3027*, 2010.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[WRS17] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. "The marginal value of adaptive gradient methods in machine learning." In *Advances in Neural Information Processing Systems*, pp. 4151–4161, 2017.

[Yan19] Greg Yang. "Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation." *arXiv preprint arXiv:1902.04760*, 2019.

[ZBH16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization." *arXiv preprint arXiv:1611.03530*, 2016.

[ZCZ18] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. "Stochastic gradient descent optimizes over-parameterized deep relu networks." *arXiv preprint arXiv:1811.08888*, 2018.

[ZCZ19] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. "Gradient descent optimizes over-parameterized deep ReLU networks." *Machine Learning*, Oct 2019.

[ZFM20] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. "Towards Theoretically Understanding Why Sgd Generalizes Better Than Adam in Deep Learning." *Advances in Neural Information Processing Systems*, **33**, 2020.

[ZG19] Difan Zou and Quanquan Gu. "An Improved Analysis of Training Over-parameterized Deep Neural Networks." In *Advances in Neural Information Processing Systems*, 2019.

[Zho19] Ding-Xuan Zhou. "Universality of deep convolutional neural networks." *Applied and Computational Harmonic Analysis*, 2019.

[ZSD17] Kai Zhong, Zhao Song, and Inderjit S Dhillon. "Learning Non-overlapping Convolutional Neural Networks with Multiple Kernels." *arXiv preprint arXiv:1711.03440*, 2017.

[ZWB21] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. "Benign overfitting of constant-stepsize sgd for linear regression." In *Conference on Learning Theory*, pp. 4633–4635. PMLR, 2021.

[ZYW19] Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. "Learning one-hidden-layer relu networks via gradient descent." In *The 22nd international conference on artificial intelligence and statistics*, pp. 1524–1534. PMLR, 2019.