Learning Ability of Deep ReLU Networks: Pairwise Tasks and Gradient Descent Methods

JUNYU ZHOU

Supervisor: Ding-Xuan Zhou and Yiming Ying

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Mathematics and Statistics
Faculty of Science
The University of Sydney
Australia

5 August 2025

Abstract

Deep neural networks (DNNs) have become a cornerstone of modern machine learning, demonstrating remarkable performance across a wide range of applications. Despite this empirical success, the theoretical understanding of DNNs, particularly in the context of generalization, remains relatively underdeveloped. ReLU (Rectified Linear Unit) is one of the most widely used activation functions in deep learning, valued for its simplicity and effectiveness in training deep networks. In this thesis, we focus on advancing the theoretical analysis of deep ReLU networks from two perspectives: pairwise learning tasks and gradient descent methods.

For pairwise learning tasks, we are concerned with the generalization performance of non-parametric estimation. Most of the existing work requires the hypothesis space to be convex or a VC-class, and the loss to be convex. However, these restrictive assumptions limit the applicability of the results in studying many popular methods, especially kernel methods and neural networks. We significantly relax these restrictive assumptions and establish a sharp oracle inequality of the empirical minimizer with a general hypothesis space for the Lipschitz continuous pairwise losses. As an application, we apply our general results to study pairwise least squares regression and derive an excess generalization bound that matches the minimax lower bound for pointwise least squares regression up to a logarithmic term. The key novelty here is to construct a structured deep ReLU neural network as an approximation of the true predictor and design the target hypothesis space consisting of the structured networks with controllable complexity. This successful application demonstrates that the obtained general results indeed help us to explore the generalization performance on a variety of problems that cannot be handled by existing approaches.

We study the generalization performance of metric and similarity learning by leveraging the *specific structure* of the true metric (the target function). Specifically, by deriving the explicit form of the true metric for metric and similarity learning with the hinge loss, we

construct a structured deep ReLU neural network as an approximation of the true metric, whose approximation ability relies on the network complexity. Here, the network complexity corresponds to the depth, the number of nonzero weights and the computation units of the network. Considering the hypothesis space which consists of the structured deep ReLU networks, we develop the excess generalization error bounds for a metric and similarity learning problem by estimating the *approximation error* and the *estimation error* carefully. An optimal excess risk rate is derived by choosing the proper capacity of the constructed hypothesis space. To the best of our knowledge, this is the *first-ever-known* generalization analysis providing the excess generalization error for metric and similarity learning. In addition, we investigate the properties of the true metric of metric and similarity learning with general losses.

For gradient descent methods, we focus on studying the statistical generalization performance of gradient descent (GD) and stochastic gradient descent (SGD) for overparameterized neural networks within the neural tangent kernel (NTK) regime. Most of the existing work on regression problems is limited to shallow network architectures, leaving a notable gap in the theory for deep neural networks. We address this gap by presenting a comprehensive generalization analysis for deep ReLU networks trained using GD and SGD. Specifically, we establish the *first known* minimax-optimal rates of excess population risk for both GD and SGD with deep ReLU networks, under the assumption that the network width scales *polynomially* with respect to the network depths and training sample size. Our results demonstrate that with sufficient widths, gradient descent methods for deep ReLU networks can achieve the optimal rates of generalization on par with kernel methods.

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Junyu Zhou

July 2025

Author Attribution Statement

Chapter 2 of the thesis is based on [Zhou et al. 2024b], which is under review at IEEE Transactions on Neural Networks and Learning Systems. My contribution involved addressing technical ideas and drafting the manuscript. My coauthor Dr. Puyu Wang helped revise the manuscript. My supervisor Prof. Ding-Xuan Zhou supervised this work and gave feedback on the final manuscript.

Chapter 3 of the thesis is available as a preprint as [Zhou et al. 2024a]. My contribution involved addressing technical ideas and drafting the manuscript. My coauthor Dr. Puyu Wang helped revise the manuscript. My supervisor Prof. Ding-Xuan Zhou supervised this work and gave feedback on the final manuscript.

Chapter 4 of the thesis is based on [Zhou et al. 2025]. My contribution involved addressing technical ideas and drafting the manuscript. My coauthor Dr. Puyu Wang helped revise the manuscript. My co-supervisor Prof. Yiming Ying supervised this work and my supervisor Prof. Ding-Xuan Zhou gave feedback on the final manuscript.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Junyu Zhou

July 2025

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Prof. Ding-Xuan Zhou

July 2025

As co-supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Prof. Yiming Ying
July 2025

GEN AI attribution statement

During the preparation of the thesis, the author used ChatGPT for the purposes of text enhancement. The use of this generative AI tool includes minor sentence restructuring and clarity improvement in Section 5.2 (Future Work). The author confirms that where text was modified by generative AI, the content was reviewed for possible errors, inaccuracies, and bias. The author takes full responsibility for the submitted thesis and ensures the work is their own and has used generative AI within the parameters of use (refer to the University of Sydney generative AI guide for researchers).

Junyu Zhou

July 2025

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Ding-Xuan Zhou, for his continuous support and guidance throughout my PhD journey. I first met Professor Zhou during a summer camp at City University of Hong Kong, and I was deeply impressed by his kindness and patience even then. Since becoming his student, I have benefited immensely from his deep insights into learning theory and his thoughtful, patient supervision. I have learned a great deal from him, and I am especially inspired by his genuine passion and unwavering dedication to research.

I am also deeply grateful to my co-supervisor, Professor Yiming Ying, who introduced me to a new research area—gradient descent methods in neural networks—during the final year of my PhD. Under his guidance, I gained important knowledge in optimization theory and techniques.

I would like to express my sincere gratitude to Dr. Puyu Wang for her invaluable support throughout my studies. Her generous guidance—from directing me to the most relevant literature, to helping me avoid common mistakes—has greatly deepened my understanding. Through her support, I have come to more fully appreciate the importance of logical structure and clarity in writing, as well as the rigor and consistency essential to academic work.

My sincere thanks go to Dr. Yunwen Lei, for his insightful suggestions when I was writing my first paper.

I would also like to thank all the members of Professor Zhou's group, my friends at CityU, as well as my colleagues in Ngau Tau Kok, Hong Kong, for their support and companionship. I am especially thankful to Professor Dachun Yang—without his kind help, I would not have the opportunity to begin my PhD studies here.

Lastly, I want to express my heartfelt appreciation to my parents, whose unconditional love and support have always been my strongest source of motivation.

Publication

The majority of the methods, concepts, analyses and results contained in this thesis have appeared previously in pre-prints listed below:

(1) **Junyu Zhou**, Shuo Huang, Han Feng, Puyu Wang, Ding-Xuan Zhou. " Fine-grained Analysis of Non-parametric Estimation for Pairwise Learning." *Under review in a Revision at IEEE Transactions on Neural Networks and Learning Systems. arXiv preprint, arXiv:2305.19640.*

(2) **Junyu Zhou**, Puyu Wang, Ding-Xuan Zhou." Generalization Analysis with Deep ReLU Networks for Metric and Similarity Learning." *Under Review. arXiv preprint, arXiv:2405.06415.*

(3) **Junyu Zhou**, Puyu Wang, Yunwen Lei, Yiming Ying, Ding-Xuan Zhou. " Optimal Rates for Generalization of Gradient Descent Methods with Deep Neural Networks." *Under Review.*

The following publication or preprint, related to this thesis, are the results of collaborative research that I contributed to throughout my candidature.

(1) Shuo Huang, **Junyu Zhou**, Han Feng, Ding-Xuan Zhou. " Generalization Analysis of Pairwise Learning for Ranking with Deep Neural Networks." *Neural Computation*, 35(6): 1135-1158, 2023.

(2) Puyu Wang, **Junyu Zhou**, Yunwen Lei, Jun Fan, Yiming Ying. " Optimal Rates for Gradient Methods with Two-Layer ReLU Neural Networks." *Under Review.*

Contents

CHAPTER 1

Introduction

Deep neural networks (DNNs) have become a cornerstone in modern machine learning due to their impressive performance across a wide range of applications, including computer vision [Krizhevsky et al. 2012; Simonyan and Zisserman 2014; He et al. 2016], natural language processing [Sutskever et al. 2014; Vaswani et al. 2017], and speech recognition [Hinton et al. 2012; Amodei et al. 2016]. While empirical results [LeCun et al. 2015; Devlin et al. 2019] have consistently demonstrated the power of DNNs, theoretical analysis, especially generalization analysis, remains relatively limited [Zhou 2020; Zhang et al. 2021a; Jacot et al. 2018; Cao and Gu 2019]. In this thesis, we focus on exploring the theoretical behaviors of deep neural networks.

Rectified Linear Unit (ReLU), as a popular activation function, has been widely used in practice because of its high efficiency and effectiveness. It overcomes the vanishing gradient problem that occurs when the activation functions of the sigmoid and hyperbolic tangent are utilized during the training process (see [Glorot et al. 2011] and Chapter 6 in [Goodfellow et al. 2016]). In addition, it has been shown in the literature [Yarotsky 2017] that deep ReLU networks can approximate any continuous and Sobolev smooth functions to arbitrary accuracy, making them widely applicable to different learning tasks. Hence, we are interested in studying the learning ability of deep ReLU networks, measured in terms of statistical generalization performance.

In the following chapters, we will introduce two types of learning problems that are widely used in modern machine learning, namely pairwise learning task and gradient descent methods. Novel analysis will be proposed to study the statistical generalization performance of the above-mentioned two learning problems with deep ReLU networks.

1.1 Generalization of Pairwise Learning

Pairwise learning is an important learning task that has attracted much attention in the modern machine learning society. Unlike pointwise learning which studies a target model based on a pointwise loss function, pairwise learning refers to learning tasks in which the loss function takes a pair of examples as the input. Representative pairwise learning tasks include ranking, metric learning and pairwise least squares regression, which can be demonstrated as follows:

- **Ranking**: the aim is to learn a predictor $f(\cdot, \cdot)$ which is capable to predict an order of objects based on their observed features. Given observers x, x' , if $f(x, x') \geq 0$, then we predict that x has a higher rank than x' and vice versa. Let y, y' be the ranking labels associated with x, x', the performance of a predictor f is measured by $\ell(f(x, x'), y, y') = \ell(sgn(y - y')f(x, x'))$, where $sgn(t) = 1$ for $t > 0, sgn(t) = -1$ for $t < 0$ and $sgn(t) = 0$ for $t = 0$, and $\ell(\cdot)$ is non-increasing.
- **Metric learning**: the aim is to study a metric $d(\cdot, \cdot)$ that estimates the distance or the similarity between a pair of observers (x, x'). Let y, y' be the labels associated with x, x', $\tau(y, y')$ be the reducing function defined by $\tau(y, y') = 1$ if $y = y'$ and $\tau(y, y') = -1$ else. The performance of d on a pair (z, z') is measured by the loss $\ell(\tau(y, y')d(x, x'))$, where $\ell(\cdot)$ is non-decreasing.
- **Pairwise least squares regression**: analogs to pointwise least squares regression, we aim to learn a predictor $f(\cdot, \cdot)$ which measures how well the predicted value $f(x, x')$ approximates the output value difference $y - y'$. We take the loss $\ell(f(x, x'), y, y') = (f(x, x') - y + y')^2$.

Let ρ be a population distribution defined on the sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^p$ is the bounded closed input space and $\mathcal{Y} \subset \mathbb{R}$ is the output space. Let $\ell : \mathbb{R} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ be a pairwise loss function. Given a predictor $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the generalization error (population risk) $\mathcal{E}(f)$ with the loss ℓ is defined as

$$\begin{aligned}\mathcal{E}(f) :&= \int_{\mathcal{Z}\times\mathcal{Z}} \ell(f(x, x'), y, y') d\rho(z) d\rho(z') \\ &= \mathbb{E}[\ell(f(X, X'), Y, Y')].\end{aligned}$$

Given a sample $S = \{Z_i = (X_i, Y_i)\}_{i=1}^n$ independently drawn from ρ, the empirical error of f based on S is given by

$$\mathcal{E}_z(f) := \frac{1}{n(n-1)} \sum_{i \neq j=1}^{n} \ell(f(X_i, X_j), Y_i, Y_j).$$

Let $f_\rho = \arg\min_{f \in \mathcal{F}} \mathcal{E}(f)$ be the true predictor (target function) that minimizes the generalization error over the space $\mathcal{F}$ consisting of all measurable functions from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $\hat{f}_z = \arg\min_{f \in \mathcal{H}} \mathcal{E}_z(f)$ be the empirical minimizer with the hypothesis space $\mathcal{H}$. We are interested in studying the statistical generalization performance of $\hat{f}_z$, which is measured by the *excess generalization error* $\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho)$, i.e., the distance between the expected error of $\hat{f}_z$ and the least possible error $\mathcal{E}(f_\rho)$. We consider using the following error decomposition to study the excess generalization error

$$\begin{aligned} \mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho) &= \{\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_{\mathcal{H}})\} + \{\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho)\}, \\ &=: S(\mathcal{H}) + D(\mathcal{H}), \end{aligned} \tag{1.1}$$

where $f_{\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}(f)$. The terms $S(\mathcal{H})$ and $D(\mathcal{H})$ are called the estimation error and the approximation error, respectively, and will be estimated part by part.

There are amount of work on the generalization analysis of non-parametric estimation for pairwise learning [Cao et al. 2016; Clémençon et al. 2008; Rejchel 2012; Ying and Li 2012; Lei and Shi 2024]. However, most of them required restrictive assumptions on both the loss function and the hypothesis space and focused on the specific learning problems. For example, [Clémençon et al. 2008] studied ranking problems and provided an oracle inequality of the empirical minimizer in the order of $O\big((\frac{V \log(n)}{n})^{\frac{1}{2-\beta}}\big)$ with the 0-1 loss, where n is the training sample size and V is the VC-dimension of the class of the ranking rules. The requirement of the class of ranking rules to be a VC-class is quite strict since many hypothesis spaces, especially the Reproducing Kernel Hilbert Space (RKHS), do not satisfy this assumption. This renders the results of [Clémençon et al. 2008] inapplicable to the widely used kernel methods. In addition, the 0-1 loss is not usually used in practice due to the difficulty of finding the empirical minimizer through an efficient algorithm. [Rejchel 2012] presented the upper bounds of the estimation error for ranking problems with the convex and nonnegative loss

under the assumption that the hypothesis space is convex. This convexity assumption on the hypothesis space makes their results unusable for the study of neural networks whose hypothesis space is not convex in general. [Cao et al. 2016; Jin et al. 2009; Lei and Ying 2020] established the estimation error bounds $O(n^{-\frac{1}{2}})$ for the metric and similarity learning problems. However, there is no work providing generalization analysis of pairwise learning with general losses under relaxed assumptions.

In the first part of the thesis, we will investigate comprehensive generalization analysis of pairwise learning under general settings by significantly removing the restrictive assumptions on both the loss functions and the hypothesis space. The main results of this part are based on our work [Zhou et al. 2024b] whose contributions can be summarized as follows.

- We establish an oracle inequality of the empirical minimizer for general hypothesis space when the loss is Lipschitz continuous and satisfies the symmetric property. Our results extend the existing literature in the following two aspects. First, we consider the general losses that are Lipschitz continuous and symmetric. Various commonly used surrogate losses including the hinge loss, least squares loss and logistic loss satisfy these assumptions, which makes our results applicable to a wide range of pairwise learning problems including ranking, pairwise regression and metric and similarity learning. Second, we consider general hypothesis spaces satisfying general capacity assumptions of the covering number. The hypothesis space is not required to be convex or a VC-class, which makes our results useful for studying the generalization performance of neural networks and kernel methods.
- We apply our general results to study generalization bounds of pairwise least squares regression. The derived oracle inequality implies that if we can find a suitable hypothesis space with controllable capacity such that the approximation error is small enough, then desired excess generalization error bounds can be established. Our key idea for designing the targeted hypothesis space is to construct a novel structured deep ReLU network as an approximation of the true predictor f_ρ according to the specific anti-symmetric structure of f_ρ. Then, considering the hypothesis space consisting of the networks with this structure, we develop an excess generalization

error bound in the order of $O\big(n^{-\frac{2r}{2r+p}}\big)$, which matches the minimax lower rate (up to a logarithmic term) of least squares regression for pointwise learning. Here, p is the dimension of the input space and $r \in \mathbb{N}$ is the smoothness index of the target function of the pointwise least squares regression. This application shows that the general oracle inequality greatly helps us explore the generalization performance on a variety of problems that cannot be handled by existing approaches.

For metric and similarity learning, we take $\mathcal{Y} = \{y_1, \ldots, y_m\}$. There are a considerable amount of theoretical works on metric and similarity learning [Cao et al. 2016; Davis et al. 2007; Guo and Ying 2014; Jin et al. 2009; Kar and Jain 2011; Lei and Ying 2016; Maurer 2008; Ye et al. 2019], most of which focused on learning the Mahalanobis distance $d_M(x, x') = (x - x')^\top M(x - x')$ for metric learning [Cao et al. 2016; Davis et al. 2007; Jin et al. 2009; Lei and Ying 2016; Ye et al. 2019] and the pairwise similarity function $s_M(x, x') = x^\top M x'$ for similarity learning [Cao et al. 2016; Chechik et al. 2010; Guo and Ying 2014; Kar and Jain 2011; Maurer 2008; Shalit et al. 2010]. Here, $M \in \mathbb{S}_p^+ \subset \mathbb{R}^{p\times p}$ is a positive semi-definite matrix. The corresponding estimation error has been studied in [Cao et al. 2016; Guo and Ying 2014; Huai et al. 2019; Ye et al. 2019]. Specifically, [Cao et al. 2016] studied the estimation error with the hypothesis space $\mathcal{H} = \{(x - x')^\top M(x - x') - b : M \in \mathbb{S}_p^+, b \in \mathbb{R}^+\}$ for metric learning with the hinge loss. With the same hypothesis space, [Ye et al. 2019] derived fast learning rates of the estimation error for metric learning with smooth loss function and strongly convex objective. [Guo and Ying 2014] investigated the estimation error with the hypothesis space $\mathcal{H} = \{b - x^\top M x' : M \in \mathbb{S}_p^+, b \in \mathbb{R}^+\}$ for similarity learning. However, the expressive power of the studied hypothesis spaces is very limited. Indeed, functions in the Mahalanobis distance and the pairwise similarity can be viewed as a multivariate polynomial with order 2 on $\mathbb{R}^p$, then the dimension of the linear span of the hypothesis space $\dim(span(\mathcal{H}))$ is controlled by p^2. It implies that the dimension p of the input space determines the complexity of $\mathcal{H}$ and further limits the expressive ability of $\mathcal{H}$. If the expressive ability of $\mathcal{H}$ is not enough, there will be a situation where the estimation error converges to 0 fast while the approximation error is very large. To fully understand the generalization performance of f, it is necessary to study the approximation error in addition

to the estimation error. However, to the best of our knowledge, there is no study that provides estimates of the approximation error for metric and similarity learning problems due to the unknown of the explicit form of the true metric.

In the second part of the thesis, we will fill the gap in studying the generalization performance of metric and similarity learning with the hinge loss by exploring the specific structure of the true metric. The main results of this part are based on our work [Zhou et al. 2024a]. The main contributions can be summarized as follows.

- We conduct comprehensive generalization analysis for metric and similarity learning with the hinge loss. Different from the previous works [Cao et al. 2016; Guo and Ying 2014; Huai et al. 2019; Ye et al. 2019] which only focus on the estimation error, we study both the estimation error and the approximation error and further derive the *first-ever-known* excess generalization error bounds for metric and similarity learning. Under some mild conditions, an optimal learning rate (up to a logarithmic term) $O(n^{-\frac{(\theta+1)r}{p+(\theta+2)r}})$ is established, where p is the dimension of the input space, θ is the parameter of a noise condition to be given later, and r is the smoothness index of the conditional probabilities.
- The technical novelty is to construct a structured deep ReLU neural network $F_a(1 - 2\sum_{i=1}^m \phi(h_i(x), h_i(x')))$ as an approximation of the true metric d_ρ by observing that $d_\rho(x, x') = sgn(1 - 2\sum_{i=1}^m p_i(x)p_i(x'))$ for $x, x' \in \mathcal{X}$ (see Theorem 5 for more details), where $p_i(x) = Prob\{Y = y_i|X = x\}$ and $a > 0$ is a free parameter. Indeed, the form of d_ρ implies that if we can design a series of sub-networks that approximate $p_i(x)$ well for each $i = 1, \ldots, m$, then a good approximation of $d_\rho(x, x')$ can be constructed by introducing a deep ReLU network $\phi(x, y)$ and a two-layer ReLU network $F_a(x)$ for approximating the multiplication function xy and the sign function $sgn(x)$, respectively. Our results (Theorems 6, 7 and 8) show that the network with the constructed form can achieve both good approximation and estimation error bounds for metric and similarity learning.
- We investigate properties of the true metric and the problem setting for metric and similarity learning with general losses. Specifically, we first show that the bias term

of the loss introduced in many works [Cao et al. 2016; Jin et al. 2009; Ye et al. 2019] can be removed, and it is reasonable to assume the output space $\mathcal{Y}$ has only finite labels rather than containing a continuous interval. Furthermore, we demonstrate that the true metric is symmetric, and the true metric between any two identical samples must be less than or equal to that between different samples under mild and intuitive conditions, which provides a rationale for the use of symmetric models like Mahalanobis distance in metric learning.

1.2 Generalization of Gradient Descent Methods with Deep ReLU Networks

DNNs trained with gradient descent methods have achieved a remarkable success across a wide range of applications, including computer vision, natural language processing, and speech recognition [Bahdanau et al. 2014; Hinton et al. 2012; Krizhevsky et al. 2017; Silver et al. 2016]. Despite their highly nonconvex and overparameterized nature, DNNs can achieve a near-zero training error while still generalizing well to unseen data [Zhang et al. 2021b]. To demystify this phenomenon, an extensive amount of work has been done to understand the generalization and optimization properties of gradient descent methods for training DNNs.

The neural tangent kernel (NTK), introduced by [Jacot et al. 2018], has emerged as a powerful framework for understanding the generalization performance of overparameterized neural networks trained using gradient descent methods. It reveals that, in the infinite-width limit, the training trajectory of a neural network with random initialization closely mirrors the behavior of its counterpart in the RKHS associated with the NTK. This connection effectively bridges the gap between learning with DNNs and classical kernel methods, allowing insights from the kernel methods to inform our understanding of DNNs.

Following this perspective, the global convergence of gradient descent methods with DNNs has been extensively studied [Allen-Zhu et al. 2019b; Du et al. 2019; Zou et al. 2018; Zou and Gu 2019], while their generalization properties have only been investigated in a few works

[Cao and Gu 2019; Cao and Gu 2020; Chen et al. 2021a; Xu and Zhu 2024]. Specifically, the appealing work [Cao and Gu 2019] and [Cao and Gu 2020] developed algorithm-dependent misclassification error bounds for deep ReLU networks trained by gradient descent (GD) and stochastic gradient descent (SGD), respectively. [Chen et al. 2021a] relaxed the requirement of their network width for both GD and SGD. However, all of these works focused on classification problems under data separation assumptions. Very recently, the work [Xu and Zhu 2024] studied one-pass SGD in the streaming (continuously coming) data setting with deep ReLU networks for regression problems and showed that the prediction error of one-pass SGD for deep ReLU networks can converge to zero in expectation, provided that the width of the network scales exponentially with the number of layers.

On another important front, it is well-established in the kernel methods literature [Dieuleveut and Bach 2016; Lin and Rosasco 2017; Yao et al. 2007; Ying and Zhou 2006] that for least squares regression, GD and SGD in the kernel setting can achieve the minimax-optimal rates of the excess population risk $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$, under standard regularity assumptions on the regression function and capacity assumptions associated with the RKHS [Caponnetto and De Vito 2007]. Here, n is the size of training data, $\beta > 0$ is the smoothness of the target function f_ρ and $\gamma \in [0, 1]$ is a parameter that measures the capacity of the hypothesis space.

Since the NTK perspective provides a close connection between the two learning processes of neural networks and kernel methods trained by gradient descent methods, it is natural to expect that neural networks trained by GD and SGD exhibit generalization performance (measured by excess population risk) comparable to their kernel-based counterparts. This conjecture has been partially validated for shallow neural networks when the width is large enough. In particular, [Nguyen and Mücke 2024; Nitanda and Taiji 2021; Wang et al. 2025b] demonstrated that GD and SGD for two-layer neural networks can replicate the classical results in the kernel setting, achieving the excess risk rates $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$. However, a critical open question remains:

Can DNNs trained by GD and SGD achieve minimax-optimal excess risk rates on par with their kernel-based counterparts?

In the third part of the thesis, we provide an affirmative answer to this question, significantly advancing the theoretical understanding of generalization of GD and SGD from shallow to deep neural networks.

We extend the results for GD and SGD from shallow to deep networks while maintaining minimax-optimal excess risk rates under mild overparameterization condition. The main results of this part are based on our work [Zhou et al. 2025]. The main contributions can be summarized as follows.

- We provide comprehensive generalization analysis for deep ReLU networks trained with gradient descent methods for regression problems. For an L-layer ReLU network with a sufficiently large width m, we show that both GD and SGD can replicate the classical results in the kernel setting with the same gradient complexity under similar assumptions. Here, gradient complexity means the number of times the algorithm calculates the gradient. Specifically, we develop the minimax-optimal excess population risk rates $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$ for both GD and SGD with deep ReLU networks when m depends polynomially on L, n and p without imposing the commonly used assumptions on the Gram matrix of the NTK, where p is the dimension of the data.
- We improve the requirement of the network width in [Xu and Zhu 2024] from exponential scaling in the number of layers L to polynomial scaling. This relaxation has been achieved in [Zou et al. 2020; Chen et al. 2021b] for the classification setting by establishing favorable properties of the network at initialization. However, their methods cannot be directly applied to our setting, as we require these properties to hold uniformly over the entire input space, while their results are typically limited to the finite training dataset.
- In particular, this is the first work to overcome the technical challenges of achieving the optimal excess risk rates for deep neural networks within the NTK regime, demonstrating that deep ReLU networks trained with GD and SGD can achieve generalization performance on par with their kernel-based counterparts. Table 4.1 summarizes the related results of GD and SGD for regression.

Technical Novelty. The minimax-optimal excess risk rates for shallow neural networks trained with gradient descent methods have been established in [Nitanda and Taiji 2021; Nguyen and Mücke 2024; Wang et al. 2025b]. However, their approaches cannot extend directly to deep neural networks. To analyze the excess population risk of $f_{\mathbf{W}(T)}$, the performance of a network $f_{\mathbf{W}}$ at the output of GD/SGD with T iterations, [Nitanda and Taiji 2021; Nguyen and Mücke 2024; Wang et al. 2025b] employ the error decomposition

$$\|f_{\mathbf{W}(T)} - f_\rho\|_\rho^2 \lesssim \|f_{\mathbf{W}(T)} - f_{\mathbf{W}(T)}^{\text{lin}}\|_\rho^2 + \|f_{\mathbf{W}(T)}^{\text{lin}} - g_T^m\|_\rho^2 + \|g_T^m - h^m\|_\rho^2 + \|h^m - f_\rho\|_\rho^2,$$

where $f_{\mathbf{W}(T)}^{\text{lin}}$ is the linear approximation of $f_{\mathbf{W}}$ at the Gaussian initialization, g_T^m is GD/SGD associated with the finite-width NTK K^m within the RKHS $\mathcal{H}_m$, h^m is either the minimizer of the regularized population risk over $\mathcal{H}_m$ [Nitanda and Taiji 2021] or GD for the population risk in $\mathcal{H}_m$ [Nguyen and Mücke 2024], and f_ρ is the target function. A critical step in their analysis is to control the term $\|g_T^m - h_m\|_\rho^2$ by $n^{-\frac{2\beta}{2\beta+\tilde{\gamma}}}$, where $\tilde{\gamma}$ is the effective dimension of $\mathcal{H}_m$. To achieve minimax-optimal rates, it is essential to demonstrate that the effective dimension of $\mathcal{H}_m$ matches that of $\mathcal{H}_K$, i.e., $\tilde{\gamma} = \gamma$. For $\gamma = 1$, this equivalence naturally holds since the integral operator associated with K^m is a trace-class operator. For $\gamma < 1$, the argument is established by treating K^m as a sum of i.i.d. random kernels with mean K (see Proposition B in [Nitanda and Taiji 2021] and Proposition A.18 in [Nguyen and Mücke 2023]). Specifically, for a two-layer ReLU network $f_{\mathbf{W}}(x) = \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \sigma(\mathbf{w}_r^\top x)$, a kernel has the structure $K^m(x, x') = \sum_{r=1}^m \langle \partial_{\mathbf{w}_r} f_{\mathbf{W}(0)}(x), \partial_{\mathbf{w}_r} f_{\mathbf{W}(0)}(x') \rangle_2$. Since $\partial_{\mathbf{w}_r} f_{\mathbf{W}(0)}(x) = m^{-\frac{1}{2}} a_r \sigma'(\mathbf{w}_r(0)^\top x) x$ depends only on the initial i.i.d. Gaussian weight $\mathbf{w}_r(0)$, K^m is a sum of $i.i.d.$ random kernels. However, this structure is not valid for deep ReLU networks. In deeper architectures, the gradient $\partial_{\mathbf{w}_r^l} f_{\mathbf{W}(0)}(x)$ is influenced not only by the weight $\mathbf{w}_r^\ell(0)$ of the l-th layer but also by the weights of all preceding and former layers. This interdependence makes a direct extension significantly challenging.

To overcome this challenge and establish minimax-optimal rates for any $\gamma \in [0, 1]$, we adopt a refined error decomposition by introducing a new GD iterate g_T associated with the infinite-width NTK K in the RKHS $\mathcal{H}_K$. Specifically, we decompose the error as

$$\|f_{\mathbf{W}(T)} - f_\rho\|_\rho^2 \lesssim \|f_{\mathbf{W}(T)} - f_{\mathbf{W}(T)}^{\text{lin}}\|_\rho^2 + \|f_{\mathbf{W}(T)}^{\text{lin}} - g_T^m\|_\rho^2 + \|g_T^m - g_T\|_\rho^2 + \|g_T - f_\rho\|_\rho^2.$$

First term: The first term $\|f_{\mathbf{W}(T)} - f^{\text{lin}}_{\mathbf{W}(T)}\|^2_\rho$ depends critically on forward and backward propagation estimates at random initialization. The work [Xu and Zhu 2024] provided such forward and backward propagation estimates with upper bounds that scale exponentially with the network depth L for deep ReLU networks. Applying their results leads to an unexpected bound $\|f_{\mathbf{W}(T)} - f^{\text{lin}}_{\mathbf{W}(T)}\|^2_\rho \lesssim C^L m^{-\frac{1}{3}}\text{Poly}(\eta T)$ valid when $m \gtrsim C^L\text{Poly}(\eta T, p)$. Here, $C > 1$ is an absolute constant and $\eta > 0$ is the step size. By extending the results of [Zou et al. 2020; Chen et al. 2021b] from the finite training set S to the full input space $\mathcal{X}$, we obtain the improved bound $\|f_{\mathbf{W}(T)} - f^{\text{lin}}_{\mathbf{W}(T)}\|^2_\rho \lesssim m^{-\frac{1}{3}}\text{Poly}(L, \eta T)$ with a relaxed requirement $m \gtrsim \text{Poly}(L, \eta T, p)$, reducing the dependence on the network depth from exponential to polynomial.

Second and fourth terms: The second term $\|f^{\text{lin}}_{\mathbf{W}(T)} - g^m_T\|^2_\rho$ can be controlled by the first term and the gap between the gradients of $f_{\mathbf{W}}$ at initialization $\mathbf{W}(0)$ and at $\mathbf{W}(T)$, while the final term, $\|g_T - f_\rho\|^2_\rho$, is bounded using standard results for kernel methods [Lin and Rosasco 2017].

Third term: The most challenging term, $\|g^m_T - g_T\|^2_\rho$, requires more nuanced analysis. A key insight here is that the infinity norm between g^m_T and g_T can be controlled by the distance between the corresponding kernels K^m and K, yielding $\|g^m_T - g_T\|^2_\rho \lesssim \|g^m_T - g_T\|^2_\infty \lesssim (\eta T)^4\|K^m - K\|^2_\infty$. [Xu and Zhu 2024] proved that $\|K^m - K\|_\infty \lesssim C^L m^{-\frac{1}{6}}\sqrt{p}$ assuming exponential scaling of m with L. By applying a more refined analysis (see Lemma 20), we establish $\|K^m - K\|_\infty \lesssim m^{-\frac{1}{6}}\sqrt{pL}$ under a relaxed condition $m \gtrsim pL^3$. This improvement significantly relaxes the overparameterization requirements, completing the analysis. Further details can be found in Proposition 9. Note that if $f_{\mathbf{W}(T)}$ is produced by SGD, the estimate strategy for the other three terms remains unchanged. There will be an additional error in the third due to the discrepancy between the SGD and GD iterates in the RKHS $\mathcal{H}_m$, which can be estimated using the results in [Lin and Rosasco 2016].

The combination of the refined error decomposition and the key insights effectively extend the analysis from shallow to deep neural networks.

CHAPTER 2

Generalization of Non-parametric Estimation for Pairwise Learning

---

In this chapter, we are concerned with the generalization performance of non-parametric estimation for pairwise learning. The main results in this chapter are based on [Zhou et al. 2024b]. The rest of the chapter is organized as follows. Section 2.1 presents the main results of the chapter. Section 2.2 applies the main results to study the generalization performance of pairwise learning with structured deep ReLU networks. The proofs are given in Section 2.3.

2.1 Main Results on Pairwise Learning

Throughout this chapter, we denote by $C_{\alpha,\beta,\gamma}$ a constant depends on parameters α, β and γ, and denote C an absolute constant. These constants may differ from line to line and are always assumed to be greater than or equal to 1. We denote $\mathbb{I}\{A\}$ the indicator function that takes value 1 if the event A happens and 0 otherwise. Let $\lceil C \rceil$ denote the least integer number greater than or equal to C. For a measurable set $A \subset \mathcal{X}$, let $\rho_{\mathbf{x}}(A) := Prob\{A\} = \mathbb{E}_X[\mathbb{I}\{A\}]$ and $\rho(A|x) := Prob\{A|X = x\} = \mathbb{E}[\mathbb{I}\{A\}|X = x]$.

2.1.1 Oracle inequalities with general losses

Our analysis requires the following assumptions on the true predictor f_ρ and the loss ℓ.

ASSUMPTION 1. *There exists a constant $\eta > 0$ such that*

$$\|f_\rho\|_{L^\infty(\mathcal{X}\times\mathcal{X})} := \sup_{x,x'\in\mathcal{X}} |f_\rho(x, x')| \leq \eta.$$

Both the true predictors of pairwise least squares regression and ranking problems satisfy Assumption 1. For the pairwise least squares regression with loss $\ell(f(x,x'),y,y') = (f(x,x')-y+y')^2$, it is known that the true predictor has the form $f_\rho(x,x') = \tilde{f}_\rho(x) - \tilde{f}_\rho(x')$, where $\tilde{f}_\rho(x) := \mathbb{E}[Y|X=x]$ is the Bayes rule for pointwise least squares problem [Ying and Zhou 2016]. One often assumes that the distribution of Y is bounded by a constant $B > 0$, i.e., $Prob\{|Y| \leq B\} = 1$, then we can show that Assumption 1 holds with $\eta = 2B$ by noting $\|f_\rho\|_{L^\infty(\mathcal{X}\times\mathcal{X})} \leq 2\|\tilde{f}_\rho\|_{L^\infty(\mathcal{X})} \leq 2B$. For ranking problems with the hinge loss $\ell(f(x,x'),y,y') = (1 - sgn(y-y')f(x,x'))_+$, the work [Huang et al. 2023] proved that the true predictor has the form $f_\rho(x,x') = sgn(T(x,x') - T(x',x))$, where $T(x,x') = Prob\{Y > Y'|X=x, X'=x'\}$. Then it is easy to show that $\eta = 1$ in this case since $|sgn(t)| \leq 1$ for any $t \in \mathbb{R}$.

ASSUMPTION 2. *There exists a constant $K > 0$, such that for any $y, y' \in \mathcal{Y}$ and $t_1, t_2 \in [-\|f_\rho\|_{L^\infty(\mathcal{X}\times\mathcal{X})}, \|f_\rho\|_{L^\infty(\mathcal{X}\times\mathcal{X})}]$, there holds*

$$|\ell(t_1,y,y') - \ell(t_2,y,y')| \leq K\,|t_1 - t_2|\,.$$

Here, we only consider the Lipschitz property of ℓ over $\big[-\|f_\rho\|_{L^\infty(\mathcal{X}\times\mathcal{X})}, \|f_\rho\|_{L^\infty(\mathcal{X}\times\mathcal{X})}\big]$ since the values predicted by f_ρ always lie on this interval. Assumption 2 holds for the hinge loss with $K = 1$ and the pairwise least squares loss with $K = 8B$ if the distribution of Y is bounded by $B > 0$.

With a little abuse of notation, for any probability measures ρ, we denote L^2_ρ as the metric induced by the norm $\|\cdot\|_{L^2_\rho}$. Here, $\|\cdot\|_{L^2_\rho}$ is L^2 norm defined by $\|f\|_{L^2_\rho} = (\int |f|^2 d\rho)^{\frac{1}{2}}$. Recall that $\rho_{\mathbf{x}}$ is the marginal distribution of ρ on $\mathcal{X}$, we define two empirical probability measures based on the observed sample S as $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\nu_n := \frac{1}{n(n-1)}\sum_{i\neq j=1}^n \delta_{(X_i,X_j)}$, where $\delta_{(\cdot)}$ is the counting measure. For any $f \in \mathcal{H}$, we define the norms

$$\big\|f\big\|_{L^2_{\rho_{\mathbf{x}}\times\mu_n}} := \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[f(X,X_i)\big|X_i\big]^2\right)^{\frac{1}{2}}, \tag{2.1}$$

$$\big\|f\big\|_{L^2_{\nu_n}} := \left(\frac{1}{n(n-1)}\sum_{i\neq j=1}^n f(X_i,X_j)^2\right)^{\frac{1}{2}}. \tag{2.2}$$

Our analysis also needs the assumption on the capacity of the hypothesis space $\mathcal{H}$.

DEFINITION 1. Let (T, d) be a metric space. Consider a subset $A \subset T$ and let $\epsilon > 0$. A subset $\mathcal{N} \subset A$ is called an ϵ-net of A if every point in A is within a distance ϵ of some point of $\mathcal{N}$, i.e.,

$$\forall x \in A,\ \exists x_0 \in \mathcal{N} : d(x, x_0) \leq \epsilon.$$

The smallest possible cardinality of an ϵ-net of A is called the covering number of A and is denoted by $\mathcal{N}(A, d, \epsilon)$.

ASSUMPTION 3. *There exist constants $s_1, s_2 \geq e$ and $V_1, V_2 > 0$ independent of n such that for any $\epsilon \in (0, 1]$,*

$$\mathcal{N}\left(\mathcal{H}, L^2_{\rho_{\mathbf{x}} \times \mu_n}, \epsilon\right) \leq s_1 \left(\frac{1}{\epsilon}\right)^{V_1},$$
$$\mathcal{N}\left(\mathcal{H}, L^2_{\nu_n}, \epsilon\right) \leq s_2 \left(\frac{1}{\epsilon}\right)^{V_2}.$$

Any bounded subset of a finite-dimensional space [Wainwright 2019] and the space consisting of bounded deep neural networks [Schmidt-Hieber 2020; Bartlett et al. 2019] satisfy Assumption 3. However, many important spaces do not satisfy this assumption, e.g., RKHS. We introduce the following assumption as an alternative to Assumption 3 to handle this setting.

ASSUMPTION 4. *There exist constants $s_1', s_2' \geq 1$ and $V_1', V_2' \in (0, 1)$ independent of n such that for any $\epsilon \in (0, 1)$,*

$$\log\left(\mathcal{N}\left(\mathcal{H}, L^2_{\rho_{\mathbf{x}} \times \mu_n}, \epsilon\right)\right) \leq s_1' \left(\frac{1}{\epsilon}\right)^{V_1'},$$
$$\log\left(\mathcal{N}\left(\mathcal{H}, L^2_{\nu_n}, \epsilon\right)\right) \leq s_2' \left(\frac{1}{\epsilon}\right)^{V_2'}.$$

It was shown in [Bartlett et al. 2005; Clémençon et al. 2008] that the convergence of the excess risk for a learning problem is of the order $O(\frac{1}{\sqrt{n}})$. However, the convergence rate of this order is often quite slow since we only utilize the first order property (the uniform boundness) of the function class. To obtain a better convergence rate, we introduce the variance condition (the second order property) of a function class, which has been extensively studied on establishing

tight concentration inequality and deriving fast convergence of the excess risk [Bartlett et al. 2005; Bartlett and Mendelson 2006; Bousquet 2002].

DEFINITION 2 (Variance-expectation bound). Let $M > 0$ and $\beta \in [0, 1]$. Let $\mathcal{F}$ be a function class consisting of functions $f : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ with nonnegative first moment, i.e., $\mathbb{E}[f] \geq 0$ for any $f \in \mathcal{F}$. We say that $\mathcal{F}$ has a variance-expectation bound with parameter pair (β, M), if for any $f \in \mathcal{F}$,

$$\mathbb{E}[f^2] \leq M(\mathbb{E}[f])^\beta. \tag{2.3}$$

REMARK 1. In statistical learning theory, we often set $\mathcal{F}$ to be the shifted hypothesis space $\{\ell(f) - \ell(f^*) : f \in \mathcal{H}\}$, where f^* is either the true predictor f_ρ or the oracle $f_\mathcal{H}$ (best predictor in $\mathcal{H}$). The inequality (2.3) holds with $\beta = 0$ for any uniformly bounded function class. There are numerous cases in which (2.3) also holds with $\beta \in (0, 1]$. For instance, (2.3) is satisfied for ranking [Clémençon et al. 2008] and for binary classification [Boucheron et al. 2005] if the distribution ρ over $\mathcal{Z}$ satisfies a low-noise condition. [Bartlett et al. 2006] showed that (2.3) holds with $\beta = \min\{1, 2/r\}$ if the hypothesis space $\mathcal{H}$ is convex and the modulus of convexity δ of the loss satisfies $\delta(\epsilon) \geq c\epsilon^r$. Further, we will prove later (2.3) is satisfied with $\beta = 1$ for pairwise least squares regression when the distribution ρ is bounded (see Lemma 11 for more details).

Now, we present our main result on an oracle inequality of $\hat{f}_z$ as follows, whose proof is given in Section 2.3.1.

THEOREM 1. *Suppose Assumptions* 1 *and* 2 *hold. Let $\mathcal{H}$ be a hypothesis space with uniform bound $\eta > 0$ such that for any $f \in \mathcal{H} \cup \{f_\rho\}$ and almost $z, z' \in \mathcal{Z}$, there holds*

$$\ell(f(x, x'), y, y') = \ell(f(x', x), y', y). \tag{2.4}$$

Let $M > 0$ and $\beta \in [0, 1]$, suppose that the shifted hypothesis space $\mathcal{F} := \{\ell(f(x, x'), y, y') - \ell(f_\rho(x, x'), y, y') : f \in \mathcal{H}\}$ has a variance-expectation bound with parameter pair (β, M). The following statements hold true.

(a) If the capacity of $\mathcal{H}$ satisfies Assumption 3, then for any $\delta \in (0, 1/2)$, with probability at least $1-\delta$, there holds

$$\mathcal{E}(\hat{f}_z)-\mathcal{E}(f_\rho) \leq C_{\eta,K,M,\beta}\Big(\frac{\max\{V_1, \log(s_1)\}\log(n)}{n}\Big)^{\frac{1}{2-\beta}}\log(\frac{4}{\delta}) + C_{\eta,K}\frac{\log^2(\frac{\delta}{2})\max\{V_1, V_2, \log(s_1), \log(s_2)\}}{n} + \big(\beta+2\big)\big(\mathcal{E}(f_{\mathcal{H}})-\mathcal{E}(f_\rho)\big).$$

(b) If the capacity of $\mathcal{H}$ satisfies Assumption 4, then for any $\delta \in (0, 1/2)$, with probability at least $1-\delta$, there holds

$$\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho) \leq C_{\eta,K,M,\beta}\max\left\{\sqrt{s_1'}\Big(\frac{1}{n}\Big)^{\frac{2}{(2+V_1')(2-\beta)}}, \Big(\frac{\log(n)}{n}\Big)^{\frac{1}{2-\beta}}\log(\frac{4}{\delta})\right\} + C_{\eta,K}\frac{\max\{s_1', s_2'\}\log^2(\frac{\delta}{2})}{n\big(1-\max\{V_1', V_2'\}\big)} + \big(\beta+2\big)\big(\mathcal{E}(f_{\mathcal{H}})-\mathcal{E}(f_\rho)\big).$$

REMARK 2. Besides ranking and pairwise least squares regression, our results can also be applied to learning problems whose predictor is independent of the order of input sample pairs, e.g., the metric learning and the similarity learning problems. Let us show how to apply Theorem 1 to metric learning problems. In distance metric learning, we aim to learn a Mahalanobis distance $f(x, x') = (x - x')^\top M(x - x')$, where $M \in \mathbb{S}_+^p$ is a positive semi-definite matrix. Let $\tau(y, y')$ be a function of labels such that $\tau(y, y') = 1$ if $y = y'$ and $\tau(y, y') = -1$ else. The performance of f on a sample pair (z, z') is measured by a loss function $\ell\big(\tau(y, y)(f(x, x') - b)\big)$, where $b > 0$ is a bias term and ℓ is a convex and Lipschitz loss. The hypothesis space here is typically set to be $\mathcal{H} = \{(x - x')^\top M(x - x') : M \in \mathbb{S}_+^p, \|M\| \leq \eta\}$, where $\|\cdot\|$ denotes a regularization norm. Since the Mahalanobis distance and the function τ are both symmetric, (2.4) holds. In addition, it's easy to verify that Assumptions 1 and 2 hold by noting that y is assumed to be bounded, and $\beta = 0$ due to the uniform boundedness of the hypothesis space. Note $\mathcal{H}$ is a bounded subset of a finite-dimensional space, Theorem 1 provides an oracle inequality for metric learning problems under Assumption 3.

Comparison with related works. We give a comparison of our work with the existing results here. The most related works are [Clémençon et al. 2008; Rejchel 2012], which both studied the generalization performance of ranking problems. [Clémençon et al. 2008] derived an

oracle inequality of order $O\big((\frac{V\log(n)}{n})^{\frac{1}{2-\beta}}\big)$ when the 0-1 loss is considered, where V is the VC-dimension of the class of the ranking rules. However, the 0-1 loss is not usually used in practice since the problem considered here is NP-hard and the empirical minimizer $\hat{f}_z$ is not easy to find through an efficient algorithm. Instead, we provide an oracle inequality with the same order (see part (a) of Theorem 1) that holds for various common-used surrogate losses including the hinge loss, least squares loss and exponential loss. This makes our results more broadly applicable. In addition, [Clémençon et al. 2008] required that the class of ranking rules is a VC-class, while many hypothesis spaces (e.g., RKHSs [Lei and Shi 2024]) do not satisfy this assumption. We extend the desired result to a more general setting (see part (b) of Theorem 1, where the hypothesis space is not assumed to be a VC-class), making our results tractable for a wider range of problems. [Rejchel 2012] established upper bounds for the estimation error with $\beta = 1$ under the assumption that the hypothesis space is convex. We investigate the excess generalization error with parameter $\beta \in [0, 1]$. Moreover, the convexity assumption of the hypothesis space is not required here, which allows our results applicable to study the generalization performance of the neural networks where the hypothesis space is not convex in general. [Cao et al. 2016; Jin et al. 2009; Lei and Ying 2020] provided the estimation error bounds of order $O(n^{-\frac{1}{2}})$ for the metric and similarity learning problems. As mentioned in Remark 2, Theorem 1 can also be applied to the metric and similarity learning (see Remark 2 for more details).

2.2 Application to Pairwise Least Squares Regression

We apply our main results to study the generalization performance of pairwise learning problems with deep ReLU networks. Specifically, we show that the optimal excess generalization error rate can be achieved for pairwise least squares regression, where constructing a hypothesis space consisting of the structured deep ReLU neural networks according to the specific form of the true predictor plays an important role.

2.2.1 A novel approximation of the true predictor

We first construct a novel structured deep ReLU network as an approximation of the true predictor f_ρ. Then, considering the hypothesis space consisting of the networks with this structure, we establish a sharp oracle inequality in the order of $O((\frac{\log(n)}{n})^{\frac{1}{2-\beta}})$ for general anti-symmetric losses. We will show in the next subsection that, the excess error rate that matches the minimax lower bound for least squares regression [Györfi et al. 2006; Schmidt-Hieber 2020] can be obtained when the least squares loss is considered.

In pairwise learning, we aim to learn a predictor $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that determines whether the sample x has higher priority than the sample x' or not. Note the value predicted by f depends on the order of the input samples x, x'. We often expect f to be anti-symmetric, i.e., $f(x, x') = -f(x', x)$. The following assumption is introduced to guarantee that the true predictor f_ρ satisfies the anti-symmetric property.

ASSUMPTION 5. *The loss ℓ is anti-symmetric with respect to any predicted values and labels, i.e., for any possible predicted value $t \in \mathbb{R}$ and labels $y, y' \in \mathcal{Y}$, there holds*

$$\ell(t, y, y') = \ell(-t, y', y).$$

It's obvious that the hinge loss $\ell(t, y, y') = (1 - sgn(y - y')t)_+$ for ranking and the least squares loss $\ell(t, y, y') = (t - y + y')^2$ for regression satisfy this assumption.

A good approximation of a model should be one that has the same structure as itself. Hence, the basic idea of designing an alternative of f_ρ is to construct a model that has the same structure as f_ρ. The following proposition shows that the true predictor for a pairwise problem has an anti-symmetric structure, which suggests the structure of the approximation of f_ρ. The proof of Proposition 1 can be found in Section 2.3.2.

PROPOSITION 1. *Under Assumption 5, the true predictor f_ρ is anti-symmetric, i.e., for almost $x, x' \in \mathcal{X}$, there holds $f_\rho(x, x') = -f_\rho(x', x)$. Furthermore, there holds*

$$f_\rho(x, x') = \frac{1}{2} f_\rho(x, x') - \frac{1}{2} f_\rho(x', x). \tag{2.5}$$

The nice decomposition of f_ρ given in the above proposition indeed tells us how to find a targeted approximation, i.e., we consider making use of a series of ReLU networks to approximate f_ρ with the structure in (2.5). Before introducing our approximation, we give the definition of deep ReLU neural networks, which are the basis of our structured networks. We denote $\sigma(t) = (t)_+$ the ReLU activation function acting componentwise on the vectors. Let $w_0 \in \mathbb{N}^+$ be the dimension of the input space, $w_L = 1$ and $w_l \in \mathbb{N}^+$ be the width of the l-th layer for $l = 1, \ldots, L-1$, where $L \in \mathbb{N}$ is the depth of the network. A deep ReLU network h from $\widetilde{\mathcal{X}}$ to $\mathbb{R}$ with depth L has the form

$$h(\widetilde{x}) = a^\top \sigma(T_{L-1}\sigma(T_{L-2}\cdots\sigma(T_1(\widetilde{x})^\top + b_1)\cdots + b_{L-2}) + b_{L-1}) + b_L \text{ for } \widetilde{x} \in \widetilde{\mathcal{X}}, \tag{2.6}$$

where $T_l \in \mathbb{R}^{w_l \times w_{l-1}}$ indicates the connection matrix between the l-th layer and the $(l-1)$-th layer, $b_l \in \mathbb{R}^{w_l}$ is the bias, and $a \in \mathbb{R}^{w_{L-1}}$. Let $\|\cdot\|_0$ denote the number of nonzero elements of the corresponding matrices and vectors. We define the number of nonzero weights and computation units of h by $(\|a\|_0 + 1) + \sum_{l=1}^{L-1} \|T_l\|_0 + \|b_l\|_0$ and $\sum_{l=1}^{L-1} w_l$, respectively. We say a network h has the complexity (L, W, U) if its depth, the number of nonzero weights and computation units are L, W and U.

Now, we can give our structured approximation network $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as follows

$$f(x, x') = g\big(\pi_\eta(h(x, x')), \pi_\eta(h(x', x))\big) \text{ for } x, x' \in \mathcal{X}, \tag{2.7}$$

where $\pi_\eta : \mathbb{R} \mapsto \mathbb{R}$ and $g : \mathbb{R} \mapsto \mathbb{R}$ are shallow ReLU networks with fixed complexity, and $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a deep ReLU network defined in (2.6) with complexity (L, W, U). The value of (L, W, U) will be properly chosen later for specific problems (see Theorem 8 for least squares regression as an example). Let us give some explanations of (2.7). First, we use a deep ReLU network h to approximate $\frac{1}{2}f_\rho(x, x')$ and $\frac{1}{2}f_\rho(x', x)$ by $h(x, x')$ and $h(x', x)$, respectively. Note that $\|f_\rho\|_{L^\infty(\mathcal{X}\times\mathcal{X})}$ is bounded by η, we hope that the values of h also lie on $[-\eta/2, \eta/2]$. Hence, we consider improving our estimate h by projecting the values of h onto

$[-\eta/2, \eta/2]$, i.e., projecting h to $\pi_\eta(h)$ with the projection operator π_η defined as

$$\pi_\eta(t) := \begin{cases} \eta/2, & t > \eta/2 \\ t, & t \in [-\eta/2, \eta/2] \\ -\eta/2, & t < -\eta/2. \end{cases}$$

The operator $\pi_\eta(h)$ can be written as a shallow ReLU network: $\pi_\eta(t) = \sigma(t) - \sigma(t - \eta/2) - \sigma(-t) + \sigma(-t - \eta/2)$. Such an expression can be found in [Zhou et al. 2024c]. With this, two main items $\frac{1}{2} f_\rho(x, x')$ and $\frac{1}{2} f_\rho(x', x)$ are addressed. It remains to find a ReLU network to handle the difference of the values of $\pi_\eta(h)$ between the sample pair (x, x') and its reverse order pair (x', x). By noting that the difference operator can also be represented by a shallow ReLU network g, i.e., $x - y = \sigma([1, -1] \cdot [x, y]) - \sigma([-1, 1] \cdot [x, y])$, we can give our final approximation $f(x, x') = g\big(\pi_\eta(h(x, x')), \pi_\eta(h(x', x))\big)$. Figure 2.1 gives the specific structure of $f(x, x')$. Note that the complexity of a network f of the form (2.7) can

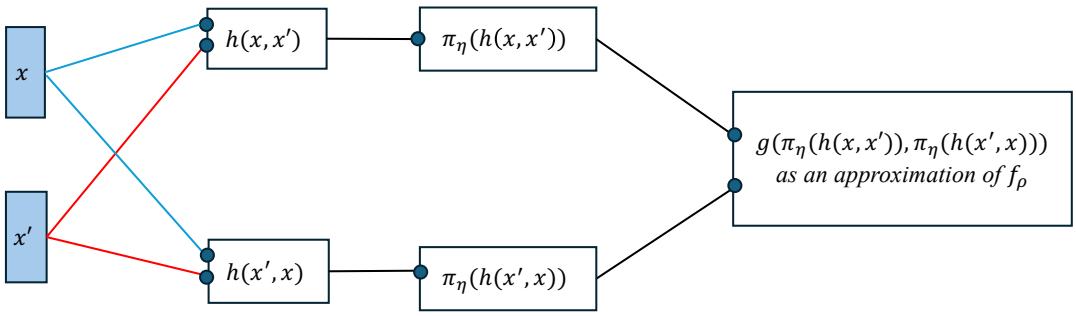

FIGURE 2.1. Structure of the designed anti-symmetric deep ReLU network (2.7) with input $x, x' \in \mathcal{X}$.

be computed by summing up the corresponding complexities of each sub-network h, π_η and g directly. Specifically, the depth of f is the summation of the depth of h, π_η and g. The number of nonzero weights and computation units of f are $2(W_h + W_{\pi_\eta}) + W_g + c$ and $2(U_h + U_{\pi_\eta}) + U_g + c'$ respectively, where W_γ and U_γ are the corresponding parameters of each sub-network $\gamma \in \{h, \pi_\eta, g\}$, and c and c' are absolute constants. The hypothesis space

$\mathcal{H}$ consists of all possible predictors of the form (2.7) is defined as

$$\mathcal{H} = \mathcal{H}(L, W, U) := \{f \text{ of form (2.7): the complexity of } f \text{ does not exceed } (L, W, U).\} \tag{2.8}$$

Here, we use the parameters $(L, W, U) \in \mathbb{N}^3$ to measure the capacity (size) of $\mathcal{H}$. As these parameters increase, the capacity of the hypothesis is getting larger.

Unlike the covering number used in Theorem 1, we employ the pseudo-dimension to characterize the hypothesis space here. An advantage of using pseudo-dimension is that any boundedness assumptions on the parameters of neural networks are not required.

DEFINITION 3. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\mathcal{F}^+ := \{(x, x', t) : f(x, x') > t, f \in \mathcal{F}\}$ be its subgraph set, then the pseudo-dimension $Pdim(\mathcal{F})$ of $\mathcal{F}$ is defined as

$$Pdim(\mathcal{F}) := VC(\mathcal{F}^+),$$

where $VC(\mathcal{F}^+)$ is the VC-dimension of $\mathcal{F}^+$. Further, if $Pdim(\mathcal{F}) < \infty$, then we call $\mathcal{F}$ a VC-class.

The following lemma reveals a relation between the covering number and pseudo-dimension [Vaart and Wellner 1996, Theorem 2.6.7] and will be used frequently later.

LEMMA 1. *For a VC-class $\mathcal{F}$ of functions with uniform bound F, one has for any probability measure ρ,*

$$\mathcal{N}(\mathcal{F}, L^2_\rho, \epsilon F) \leq C\, Pdim(\mathcal{F})(16e)^{Pdim(\mathcal{F})} \left(\frac{1}{\epsilon}\right)^{2(Pdim(\mathcal{F})-1)}$$

for an absolute constant $C > 0$ and $0 < \epsilon < 1$, where L^2_ρ denotes the L^2 norm with respect to the measure ρ.

Lemma 1 implies that any bounded VC-class satisfies Assumption 3. Specifically, for any hypothesis space $\mathcal{H}$ which is a VC-class and uniformly bounded by $\eta > 0$, $\mathcal{H}$ satisfies Assumption 3 with $s_1 = s_2 = C_{\eta, Pdim(\mathcal{H})}$ and $V_1 = V_2 = 2(Pdim(\mathcal{H}) - 1)$, respectively.

Conversely, if we further assume that the inequalities therein hold with any probability measures, then Assumption 3 can be employed as an alternative definition of a VC-class.

Now, we can establish the following theorem on an oracle inequality of the empirical minimizer $\hat{f}_z$ trained by the structured deep ReLU networks of the form (2.7). The proof of Theorem 2 can be directly derived by Theorem 1, detailed proof can be found in Section 2.3.2.

THEOREM 2. *Suppose Assumptions 1,2 and 5 hold. Let $\mathcal{H}$ be the space of structured deep ReLU networks* (2.8) *and $V = Pdim(\mathcal{H})$ be its pseudo-dimension. Let $M > 0$ and $\beta \in [0, 1]$, suppose that the shifted hypothesis space $\mathcal{F} := \{\ell(f(x, x'), y, y') - \ell(f_\rho(x, x'), y, y') : f \in \mathcal{H}\}$ has a variance-expectation bound with parameter pair (β, M). Then for any $\delta \in (0, 1/2)$, with probability at least $1 - \delta$, there holds*

$$\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho) \leq C_{\eta,K,M,\beta}\Big(\frac{V\log(n)}{n}\Big)^{\frac{1}{2-\beta}}\log(4/\delta) + C_{\eta,K}\frac{V}{n}\log^2(\delta/2) + \big(\beta+2\big)\big(\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho)\big).$$

2.2.2 Pairwise least squares regression with deep ReLU networks

Now, we focus on the pairwise least squares regression with loss $\ell(f(x, x'), y, y') = \big(f(x, x') - y + y'\big)^2$. We first estimate the approximation error with $\mathcal{H}$ of the form (2.8). Then, by combining the approximation error bound with Theorem 2 and choosing the proper capacity of $\mathcal{H}$, we derive the nearly optimal excess generalization error bound for pairwise least squares regression.

The estimates of the approximation error $\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho)$ are closely related to the true predictor f_ρ, which is known to be $f_\rho(x, x') = \tilde{f}_\rho(x) - \tilde{f}_\rho(x')$ for $x, x' \in \mathcal{X}$ [Ying and Zhou 2016]. Here $\tilde{f}_\rho(x) = \mathbb{E}[Y|X = x]$ is the regression function. Before stating our approximation results, we first introduce some notations and standard assumptions.

To define the Sobolev spaces, we assume that $\mathcal{X} = [0, 1]^p$ in the remainder of this subsection. Let $r \in \mathbb{N}$ be the smoothness index, the Sobolev space $W^{r,\infty}([0, 1]^p)$ is defined as a class consisting of functions along with their partial derivatives up to order r lying in $L^\infty([0, 1]^p)$.

The norm in $W^{r,\infty}([0,1]^p)$ is defined as

$$\left\|f\right\|_{W^{r,\infty}([0,1]^p)} := \max_{\alpha\in\mathbb{Z}_+^p:\|\alpha\|_1\leq r} \left\|D^\alpha f\right\|_{L^\infty([0,1]^p)},$$

where $\mathbb{Z}_+$ contain all non-negative integers, $\|\alpha\|_1 = \sum_{i=1}^p |\alpha_i|$ denotes the l^1 norm of α, $D^\alpha f = \frac{\partial^{\|\alpha\|_1} f}{\partial x_1^{\alpha 1}\cdots\partial x_p^{\alpha p}}$ denotes the weak partial derivative of f with order α.

The following assumption assumes that all r-th derivatives of the $\tilde{f}_\rho$ exist and their L^∞ norms are bounded.

ASSUMPTION 6. *Assume* $\|\tilde{f}_\rho\|_{W^{r,\infty}([0,1]^p)} \leq 1$ *for some* $r \in \mathbb{N}$.

Since Sobolev norm dominates $L^\infty([0,1]^p)$ norm, then we know that Assumption 6 implies Assumption 1 with $\eta = 2$.

We also require the distribution of Y to be uniformly bounded, which ensures that Assumption 2 holds.

ASSUMPTION 7. *There exists a constant* $B > 0$ *such that* $Prob\{|Y| \leq B\} = 1$.

We are in a position to give our estimate of the approximation error, whose proof can be found in Section 2.3.2.

THEOREM 3. *Suppose Assumption 6 holds, and the structured hypothesis space* $\mathcal{H}(L, W, U)$ *is defined by* (2.8)*. Then, for any* $\epsilon \in (0,1)$*, there exists a deep ReLU network* f *of form* (2.7) *with depth at most* $C_{p,r}\log(1/\epsilon)$ *and the number of nonzero weights and computation units at most* $C_{p,r}\epsilon^{-\frac{p}{r}}\log(1/\epsilon)$ *such that*

$$\left\|f - f_\rho\right\|_{L^\infty([0,1]^{2p})} \leq \epsilon.$$

Further, if we set $W = U = \lceil \exp(L) \rceil$*, then the approximation error can be bounded as follows*

$$\mathcal{D}(\mathcal{H}) \leq C_{p,r}\Big(\frac{L}{\exp(L)}\Big)^{\frac{2r}{p}}.$$

Now, we can derive the excess generalization error bound by combining Theorems 2 and 3.

THEOREM 4. *Suppose Assumptions 6 and 7 hold, and the structured hypothesis space* $\mathcal{H}(L, W, U)$ *is defined by* (2.8). *If we set* $W = U = \lceil \exp(L) \rceil$, *then for any* $\delta \in (0, 1/2)$, *with probability at least* $1 - \delta$, *there holds*

$$\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho) \leq C_{p,r,B} \frac{L^2 \exp(L) \log(n)}{n} \log^2(4/\delta) + C_{p,r} \Big(\frac{L}{\exp(L)} \Big)^{\frac{2r}{p}}.$$

Setting $L = \big\lceil \frac{p}{2r+p} \log(n) \big\rceil$, *with probability at least* $1 - \delta$, *there holds*

$$\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho) \leq C_{p,r,B} \log^\tau(n) n^{-\frac{2r}{2r+p}} \log^2(4/\delta),$$

where $\tau = \max\{3, \frac{2r}{p}\}$.

REMARK 3. Theorem 4 establishes the excess risk rate $O(n^{-\frac{2r}{2r+p}})$ for pairwise least squares regression with deep ReLU networks by choosing the parameters $L \asymp log(n)$, $W = U \asymp n^{\frac{p}{2r+p}}$, which matches the minimax lower rate for pointwise least squares regression [Györfi et al. 2006; Schmidt-Hieber 2020]. We note that the minimax lower rate $O\big(n^{-\frac{2r}{2r+s}}\big)$ of the excess risk for pairwise learning is developed in [Guo et al. 2022] by using the kernel method, where $r \in (1/2, 3/2]$ is the smoothness parameter of f_ρ, and $s \in (0, 1)$ is the effective dimension measuring the capacity of the corresponding RKHS. [Huang et al. 2023] studied ranking with deep ReLU networks when the hinge loss is considered, they derived the excess risk rate $O\big(n^{-\frac{r(\theta+1)}{2p+r(\theta+2)}}\big)$, where $r > 0$ denotes the smoothness of the target function, and $\theta > 0$ is the parameter of noise condition. As we mentioned before, the results of previous works [Clémençon et al. 2008; Rejchel 2012] cannot be used for studying the kernel methods or neural networks with the least squares loss, which demonstrates that our general results indeed help us to explore the generalization performance on a variety of problems that cannot be handled by existing approaches.

2.3 Proofs for Main Results

2.3.1 Proofs for Section 2.1

In this subsection, we present the proof of Theorem 1. The oracle inequality established in Theorem 1 is obtained by deriving upper bounds of the estimation error $S(\mathcal{H}) = \mathcal{E}(\hat{f}_z) - \mathcal{E}(f_{\mathcal{H}})$ defined in (1.1). In classical statistical learning theory with pointwise loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ [Cucker and Zhou 2007; Bartlett and Mendelson 2006], the estimation error is often bounded by $2\sup_{f\in\mathcal{H}} |\mathbb{E}[\ell(f(X), Y)] - \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)|$, where the analysis of this empirical term heavily depends on the independence of $\{\mathbb{E}[\ell(f(X), Y)] - \ell(f(X_i), Y_i)\}_{i=1}^{n}$. However, for pairwise learning with loss ℓ and predictor $f \in \mathcal{H}$, the terms $\{l(f(X_i, X_j), Y_i, Y_j)\}_{i\neq j=1}^{n}$ in the double-index summation $\mathcal{E}(f) - \mathcal{E}_z(f)$ are dependent, which cannot be handled by the standard techniques in the empirical process theory directly. We consider employing the Hoeffding decomposition [Hoeffding 1963] to overcome this dependency difficulty.

Given i.i.d. sample $S = \{Z_i\}_{i=1}^{n}$ and a symmetric kernel $q : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, a U-statistic U_n associated with kernel q is defined as $U_n = \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n} q(Z_i, Z_j)$. A degenerate U-statistic W_n is a U-statistic such that $\mathbb{E}[W_n|Z_i] = 0$ for any $i \in \{1, \cdots, n\}$. Hoeffding decomposition breaks $\mathbb{E}[U_n] - U_n$ into the summation of an i.i.d. term and a degenerate U-statistic term. i.e.,

$$\mathbb{E}[U_n] - U_n = 2T_n + W_n,$$

where

$$\begin{aligned} h(Z_i) &= \mathbb{E}[U_n] - \mathbb{E}[q(Z_i, Z)|Z_i], \\ T_n &= \frac{1}{n}\sum_{i=1}^{n} h(Z_i), \\ \hat{h}(Z_i, Z_j) &= \mathbb{E}[U_n] - h(Z_i) - h(Z_j) - q(Z_i, Z_j), \\ W_n &= \frac{1}{n(n-1)}\sum_{i\neq j=1} \hat{h}(Z_i, Z_j). \end{aligned}$$

Note for any $i \in \{1, \cdots, n\}$, there holds $\mathbb{E}[W_n|Z_i] = 0$. Then W_n is a degenerate U-statistic.

We now apply Hoeffding decomposition to the estimation error $\mathcal{E}(\hat{f}_z)-\mathcal{E}(f_{\mathcal{H}})$. For any $f \in \mathcal{H}$, from (2.4) we know the kernel defined as $q_f(z,z') := \ell(f(x,x'),y,y') - \ell(f_\rho(x,x'),y,y')$ is symmetric. Further, denote by $U_n^f, h_f, T_n^f, \hat{h}_f, W_n^f$ the corresponding Hoeffding decomposition terms associated with q_f. That is, $U_n^f = \frac{1}{n(n-1)}\sum_{i\neq j=1}^n q_f(Z_i,Z_j)$, $h_f(Z_i) = \mathbb{E}[q_f(Z_i,Z)|Z_i] - \mathbb{E}[U_n^f]$, $T_n^f = \frac{1}{n}\sum_{i=1}^n h_f(Z_i)$, $\hat{h}_f(Z_i,Z_j) = \mathbb{E}[U_n^f] - h_f(Z_i) - h_f(Z_j) - q_f(Z_i,Z_j)$, and $W_n^f = \frac{1}{n(n-1)}\sum_{i\neq j=1}^n \hat{h}_f(Z_i,Z_j)$. Then the estimation error can be decomposed as

$$
\begin{aligned}
\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_{\mathcal{H}}) &= \mathcal{E}(\hat{f}_z) - \mathcal{E}_z(\hat{f}_z) + \mathcal{E}_z(\hat{f}_z) - \mathcal{E}_z(f_{\mathcal{H}}) + \mathcal{E}_z(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \\
&\leq \mathcal{E}(\hat{f}_z) - \mathcal{E}_z(\hat{f}_z) + \mathcal{E}_z(f_{\mathcal{H}}) - \mathcal{E}(f_{\mathcal{H}}) \\
&= \{\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho) + \mathcal{E}_z(f_\rho) - \mathcal{E}_z(\hat{f}_z)\} + \{\mathcal{E}_z(f_{\mathcal{H}}) - \mathcal{E}_z(f_\rho) + \mathcal{E}(f_\rho) - \mathcal{E}(f_{\mathcal{H}})\} \\
&= \{\mathbb{E}[U_n^{\hat{f}_z}] - U_n^{\hat{f}_z}\} + \{U_n^{f_{\mathcal{H}}} - \mathbb{E}[U_n^{f_{\mathcal{H}}}]\} \\
&= 2T_n^{\hat{f}_z} + W_n^{\hat{f}_z} - 2T_n^{f_{\mathcal{H}}} - W_n^{f_{\mathcal{H}}} \\
&= \{2T_n^{\hat{f}_z} - 2T_n^{f_{\mathcal{H}}}\} + \{W_n^{\hat{f}_z} - W_n^{f_{\mathcal{H}}}\} \\
&=: S_1(\mathcal{H}) + S_2(\mathcal{H}),
\end{aligned} \tag{2.9}
$$

where in the first inequality we have used the fact that $\mathcal{E}_z(\hat{f}_z) \leq \mathcal{E}_z(f_{\mathcal{H}})$. Here, $S_1(\mathcal{H})$ consists of i.i.d. terms which can be bounded by using standard techniques in empirical process theory. For bounding the degenerate U-statistics term $S_2(\mathcal{H})$, we exploit the decoupling methods in the theory of U-processes [De la Peña and Giné 1999; Clémençon et al. 2008]. We estimates $S_1(\mathcal{H})$ and $S_2(\mathcal{H})$ in the following two steps, respectively.

Step 1. Upper bounds for $S_1(\mathcal{H})$

Let $\mathcal{G} = \{g_f(z) := \mathbb{E}\big[\ell(f(x,X),y,Y) - \ell(f_\rho(x,X),y,Y)|X=x,Y=y\big] : f \in \mathcal{H}\}$ be the function class consisting of the conditional expectation of the symmetric kernels q_f. From the definition of g_f, we know $T_n^f = \mathbb{E}[g_f(Z)] - \frac{1}{n}\sum_{i=1}^n g_f(Z_i)$ and

$$
S_1(\mathcal{H}) = 2\bigg(\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] - \frac{1}{n}\sum_{i=1}^n g_{\hat{f}_z}(Z_i) + \frac{1}{n}\sum_{i=1}^n g_{f_{\mathcal{H}}}(Z_i) - \mathbb{E}\big[g_{f_{\mathcal{H}}}(Z)\big]\bigg). \tag{2.10}
$$

We estimate $\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] - \frac{1}{n}\sum_{i=1}^n g_{\hat{f}_z}(Z_i)$ and $\frac{1}{n}\sum_{i=1}^n g_{f_{\mathcal{H}}}(Z_i) - \mathbb{E}\big[g_{f_{\mathcal{H}}}(Z)\big]$ separately. We will use the Bernstein concentration inequality to control the second term directly. The first term cannot be estimated by the Bernstein concentration inequality due to the appearance of the empirical risk minimizer $\hat{f}_z$ depending on the sample. We will bound it by using the tools (e.g., local complexities and sharp concentration inequalities) in empirical process theory [Bousquet 2002; Bartlett et al. 2005].

To guarantee that the local complexity has good properties, we enlarge the function class by introducing the star-shaped class and the star-hull of a function class as follows.

DEFINITION 4. A function class $\mathcal{F}$ is called a star-shaped class around 0 if for any $f \in \mathcal{F}$ and $\alpha \in [0,1]$, $\alpha f \in \mathcal{F}$.

DEFINITION 5. Given a function class $\mathcal{F}$. Denote by

$$\mathcal{F}^* = \{\alpha f : \alpha \in [0,1], f \in \mathcal{F}\} \tag{2.11}$$

the star hull of $\mathcal{F}$ around 0.

Intuitively speaking, a star-shaped class around 0 contains all the line segments between 0 and any point in $\mathcal{F}$. The star-hull of $\mathcal{F}$ is the smallest star-shaped class that contains $\mathcal{F}$.

Since the star hull $\mathcal{G}^*$ contains $\mathcal{G}$, we know the term $\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] - \frac{1}{n}\sum_{i=1}^n g_{\hat{f}_z}(Z_i)$ in $S_1(\mathcal{H})$ can be bounded by $\sup_{g_f \in \mathcal{G}^*}\big|\mathbb{E}[g_f] - \frac{1}{n}\sum_{i=1}^n g_f(Z_i)\big|$. However, this bound is quite loose. Indeed, this upper bound controls the deviation from the generalization errors and empirical errors simultaneously over the whole class $\mathcal{G}^*$, which might be much more larger than that of the empirical risk minimizer $\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] - \frac{1}{n}\sum_{i=1}^n g_{\hat{f}_z}(Z_i)$. To derive sharp error bounds, we introduce $\mathcal{G}_r^* := \{g_f \in \mathcal{G}^* : \mathbb{E}[g_f^2] \le r\}$ with $r > 0$, a small subset of $\mathcal{G}^*$ consisting of predictors satisfying a mild variance condition. The corresponding local complexity is defined as

$$\phi(r) = \mathbb{E}\left[\sup_{g_f \in \mathcal{G}_r^*}\left|\mathbb{E}[g_f] - \frac{1}{n}\sum_{i=1}^n g_f(Z_i)\right|\right]. \tag{2.12}$$

We will use the above notion of local complexity and the corresponding fixed point to introduce a sharp concentration inequality. To this end, we need to show that (2.12) is a sub-root function [Bartlett et al. 2005].

DEFINITION 6. A function $\psi : [0, \infty) \to [0, \infty)$ is sub-root if it is non-negative, nondecreasing and if $r \mapsto \psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$.

Intuitively, sub-root functions increase slowly. From [Bartlett et al. 2005] we know they are continuous and have a unique positive fixed point r^*. i.e., there exists a unique $r^* \in [0, \infty)$ such that $\psi(r^*) = r^*$. It is clear that (2.12) is non-negative and nondecreasing on $[0, \infty)$. Since $\mathcal{G}^*$ is star-shaped, we can show $\phi(r)/\sqrt{r}$ is nonincreasing on $[0, \infty)$ by following the arguments in the proof of Lemma 3.4 in [Bartlett et al. 2005]. Then, from the above definition we know (2.12) is sub-root and has a unique fixed point r^*.

The following lemma from [Bousquet 2002, Theorem 5.4] establishes a sharp concentration inequality in terms of the fixed point of a sub-root function.

LEMMA 2 ([Bousquet 2002]). *Let $M > 0$ and $\beta \in [0, 1]$, and $\mathcal{F}$ be a star-shaped class around 0. Suppose $\mathcal{F}$ is uniformly bounded by a constant $b > 0$ and has a variance-expectation bound with parameter pair (β, M). Let r^* be the unique fixed point of the following sub-root function $\psi(r) = \mathbb{E}\left[\sup_{f \in \mathcal{F}: \mathbb{E}[f^2] \le r} \left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}[f]\right|\right]$. Then, for any $\delta \in (0, 1)$ and $\kappa > 1$, with probability $1 - \delta$, we have*

$$\forall f \in \mathcal{F},\ \mathbb{E}[f] \le \frac{\kappa}{\kappa - 1}\frac{1}{n}\sum_{i=1}^n f(Z_i) + C_{\kappa,M,\beta}\left((r^*)^{\frac{1}{2-\beta}} + \left(\frac{b\log(1/\delta)}{n}\right)^{\frac{1}{2-\beta}}\right).$$

Lemma 2 with $\psi(r) = \phi(r)$ implies that $\mathbb{E}[g_{\hat{f}_z}(Z)] - \frac{1}{n}\sum_{i=1}^n g_{\hat{f}_z}(Z_i) \le \frac{1}{\kappa}\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] + C_{K,M,\beta}\big((r^*)^{\frac{1}{2-\beta}} + \big(\frac{b\log(1/\delta)}{n}\big)^{\frac{1}{2-\beta}}\big)$ with high probability. However, this upper bound that depends on the fixed point r^* is too rough to explore the behavior of $g_{\hat{f}_z}$. The following proposition introduces the estimates of r^* in terms of the sample size n and the capacity of the hypothesis space, which helps us get a more explicit convergence rate of $\mathbb{E}[g_{\hat{f}_z}(Z)] - \frac{1}{n}\sum_{i=1}^n g_{\hat{f}_z}(Z_i)$.

PROPOSITION 2. *Let r^* be the fixed point of the local complexity* (2.12). *Suppose we are under the same conditions as that in Theorem 1.*

- *If the capacity of $\mathcal{H}$ satisfies Assumption 3, then*

$$r^* \leq C_{\eta,L} \max\{V_1, \log(s_1)\} \frac{\log(n)}{n}.$$

- *If the capacity of $\mathcal{H}$ satisfies Assumption 4, then*

$$r^* \leq C_{\eta,L} \max\left\{ \sqrt{s_1'} \Big(\frac{1}{n}\Big)^{\frac{2}{2+V_1'}}, \frac{\log(n)}{n} \right\}.$$

PROOF. Define $\rho_n := \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ the empirical measure on $\mathcal{Z}$. For any $g_1, g_2 \in \mathcal{G}_r^*$, we have $\|g_1 - g_2\|_{L^2_{\rho_n}}^2 = \frac{1}{n}\sum_{i=1}^n |g_1(Z_i) - g_2(Z_i)|^2$. Let $\{\epsilon_i\}_{i=1}^n$ be the i.i.d. Rademacher variables, i.e., $Prob\{\epsilon_i = 1\} = Prob\{\epsilon_i = -1\} = 1/2$. Then, for $r \geq r^*$, according to the standard symmetrization method and the chaining lemma [Wainwright 2019], there holds

$$\begin{aligned} \phi(r) &\leq 2\mathbb{E}\bigg[\mathbb{E}\bigg[\sup_{g\in\mathcal{G}_r^*} \Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(Z_i)\Big| \bigg| Z_1, \cdots, Z_n\bigg]\bigg] \\ &\leq \frac{C}{\sqrt{n}}\mathbb{E}\Big[\int_0^{\sqrt{S}} \sqrt{\log \mathcal{N}(\mathcal{G}_r^*, L^2_{\rho_n}, t)}dt\Big], \end{aligned} \tag{2.13}$$

where $S := \sup_{g\in\mathcal{G}_r^*} \frac{1}{n}\sum_{i=1}^n g(Z_i)^2$.

Before proving the two cases, we need to control $\mathcal{N}(\mathcal{G}_r^*, L^2_{\rho_n}, t)$ and $\mathbb{E}[S]$. We first upper bound $\mathcal{N}(\mathcal{G}_r^*, L^2_{\rho_n}, t)$ by the covering number of $\mathcal{H}$.

Let $\mathcal{M}$ be a t-net of $\mathcal{G}$ and define $N := \lceil \frac{A}{t} \rceil$, where $A := \sup_{g_f\in\mathcal{G}} \|g_f\|_{L^2_{\rho_n}}$. We claim that $\{\frac{i}{N} g_t : i = 1, ..., N, g_t \in \mathcal{M}\}$ is a $2t$-net of $\mathcal{G}^*$. Indeed, for any $g_f^* \in \mathcal{G}^*$, we know there exists a $g_f \in \mathcal{G}$ and an $\alpha \in [i/N, (i+1)/N]$ for some $i \in \{0, ..., N-1\}$, such that $g_f^* = \alpha g_f$. Notice that there exists a $g_t \in \mathcal{M}$ such that $\|g_t - g_f\|_{L^2_{\rho_n}} \leq t$. Then we conclude that $\|(i/N)g_t - g_f^*\|_{L^2_{\rho_n}} = \|(i/N)g_t - \alpha g_f\|_{L^2_{\rho_n}} \leq i/N\|g_t - g_f\|_{L^2_{\rho_n}} + |i/N - \alpha|\|g_f\|_{L^2_{\rho_n}} \leq (i/N)t + t \leq 2t$. Further, according to the Exercise 4.2.10 in [Vershynin 2018] with the fact $\mathcal{G}_r^* \subset \mathcal{G}^*$, there holds

$$\mathcal{N}\Big(\mathcal{G}_r^*, L^2_{\rho_n}, t\Big) \leq \mathcal{N}\Big(\mathcal{G}^*, L^2_{\rho_n}, t/2\Big) \leq \mathcal{N}\Big(\mathcal{G}, L^2_{\rho_n}, t/4\Big) \left\lceil \frac{A}{t} \right\rceil. \tag{2.14}$$

Now, we estimate the covering number of $\mathcal{G}$ with repect to metric $L^2_{\rho_n}$. For any $g_{f_1}, g_{f_2} \in \mathcal{G}$, there holds

$$\begin{aligned}\left\|g_{f_1} - g_{f_2}\right\|^2_{L^2_{\rho n}} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\ell(f_1(X_i, X), Y_i, Y) - \ell(f_2(X_i, X), Y_i, Y)\big|X_i, Y_i\big]^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\big(\ell(f_1(X_i, X), Y_i, Y) - \ell(f_2(X_i, X), Y_i, Y)\big)^2\big|X_i, Y_i\big] \\ &\leq \frac{K^2}{n}\sum_{i=1}^{n} \mathbb{E}\big[\big(f_1(X_i, X) - f_2(X_i, X)\big)^2\big|X_i\big] = K^2\left\|f_1 - f_2\right\|^2_{L^2_{\mu_n \times \rho_{\boldsymbol{x}}}},\end{aligned}$$

where in the first inequality we have used the Jensen's inequality for conditional expectation, and in the second inequality we have used the Lipschitz property of the loss (Assumption 2). Hence, the above inequality implies

$$\mathcal{N}\Big(\mathcal{G}, L^2_{\rho_n}, t\Big) \leq \mathcal{N}\Big(\mathcal{H}, L^2_{\mu_n \times \rho_{\boldsymbol{x}}}, t/K\Big). \tag{2.15}$$

According to Assumptions 1 and 2, we know

$$A = \sup_{g_f \in \mathcal{G}} \|g_f\|_{L^2_{\rho_n}} \leq \sup_{g_f \in \mathcal{G}} \|g_f\|_\infty \leq K \sup_{f \in \mathcal{H}} \|f - f_\rho\|_\infty \leq 2K\eta.$$

Combining (2.14) and (2.15) together, we finally have

$$\mathcal{N}\Big(\mathcal{G}^*_r, L^2_{\rho_n}, t\Big) \leq \mathcal{N}\Big(\mathcal{H}, L^2_{\mu_n \times \rho_{\boldsymbol{x}}}, t/(4K)\Big)\left\lceil \frac{2K\eta}{t} \right\rceil. \tag{2.16}$$

Now, we are in a position to derive upper and lower bounds for $\mathbb{E}[S]$, which will be used in the estimates of the entropy integral later. Recall that $S = \sup_{g \in \mathcal{G}^*_r} \frac{1}{n}\sum_{i=1}^{n} g(Z_i)^2$. Let

$\{Z_i'\}_{i=1}^n$ be an independent copy of $\{Z_i\}_{i=1}^n$, the upper bound can be estimated as follows

$$\begin{aligned}
\mathbb{E}[S] &= \mathbb{E}\bigg[\sup_{g_f\in\mathcal{G}_r^*}\frac{1}{n}\sum_{i=1}^n g_f(Z_i)^2 - \mathbb{E}[g_f^2] + \mathbb{E}[g_f^2]\bigg] \\
&\le 2\mathbb{E}\bigg[\sup_{g_f\in\mathcal{G}_r^*}\frac{1}{n}\sum_{i=1}^n \epsilon_i g_f(Z_i)^2\bigg] + \sup_{g_f\in\mathcal{G}_r^*}\mathbb{E}[g_f^2] \\
&\le 8K\eta\mathbb{E}\bigg[\sup_{g_f\in\mathcal{G}_r^*}\frac{1}{n}\sum_{i=1}^n \epsilon_i g_f(Z_i)\bigg] + r \\
&= 8K\eta\mathbb{E}\bigg[\sup_{g_f\in\mathcal{G}_r^*}\frac{1}{n}\sum_{i=1}^n \epsilon_i g_f(Z_i) - \epsilon_i\mathbb{E}[g_f] + \epsilon_i\mathbb{E}[g_f]\bigg] + r \\
&\le 8K\eta\mathbb{E}\bigg[\sup_{g_f\in\mathcal{G}_r^*}\frac{1}{n}\sum_{i=1}^n \epsilon_i\big(g_f(Z_i) - g_f(Z_i')\big)\bigg] + 8K\eta\sup_{g_f\in\mathcal{G}_r^*}\mathbb{E}[g_f]\,\mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n\epsilon_i\bigg] + r \\
&= 8K\eta\mathbb{E}\bigg[\sup_{g_f\in\mathcal{G}_r^*}\frac{1}{n}\sum_{i=1}^n g_f(Z_i) - g_f(Z_i')\bigg] + r \le 16K\eta\phi(r) + r,
\end{aligned}$$

where in the first and third inequalities we have used symmetrization method, and in the second inequality we have used the well-known Ledoux-Talagrand contraction principle [Ledoux and Talagrand 1991] with $x \mapsto x^2$ and the definition of $\mathcal{G}_r^*$, and in the last equality we have used the fact that $\{\epsilon_i\big(g(Z_i) - g(Z_i')\big)\}_{i=1}^n$ and $\{g(Z_i) - g(Z_i')\}_{i=1}^n$ are identically distributed.

For the lower bound, we can show that there exists a $g_0 \in \mathcal{G}_r^*$ such that $\mathbb{E}[g_0^2] = r$ if $r < \sup_{g_f\in\mathcal{G}}\mathbb{E}[g_f^2]$. Indeed, if $r < \sup_{g_f\in\mathcal{G}}\mathbb{E}[g_f^2]$, by definition we know there exists a $g \in \mathcal{G}$ such that $\mathbb{E}[g^2] > r$. Setting $\alpha = \sqrt{\frac{r}{\mathbb{E}[g^2]}} \in (0,1)$ and $g_0 := \alpha g$, we know $\mathbb{E}[g_0^2] = r$ and $g_0 \in \mathcal{G}_r^*$ since $\mathcal{G}_r^*$ is star shaped. Therefore, we have the following lower bound

$$\mathbb{E}[S] \ge \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n g_0(Z_i)^2\right] = \mathbb{E}[g_0^2] = r.$$

Now, we consider the following two cases.

First case: suppose Assumption 3 holds.

According to (2.16), there holds

$$\log \mathcal{N}\Big(\mathcal{G}_r^*, L^2_{\rho_n}, t\Big) \leq \log\left\{s_1\Big(\frac{4K}{t}\Big)^{V_1}\left\lceil\frac{2K\eta}{t}\right\rceil\right\} \leq C_{\eta,K}\max\{V_1, \log(s_1)\}\log\Big(\frac{2K\eta}{t}\Big).$$

Then, (2.13) can be bounded as follows

$$\begin{aligned}
\phi(r) &\leq C_{\eta,K}\sqrt{\frac{\max\{V_1,\log(s_1)\}}{n}}\,\mathbb{E}\bigg[\int_0^{\sqrt{S}}\sqrt{\log\Big(\frac{2K\eta}{t}\Big)}dt\bigg] \\
&\leq C_{\eta,K}\sqrt{\frac{\max\{V_1,\log(s_1)\}}{n}}\,\mathbb{E}\bigg[\sqrt{2S\log\Big(\frac{C_{\eta,K}}{S}\Big)}\bigg] \\
&\leq C_{\eta,K}\sqrt{\frac{\max\{V_1,\log(s_1)\}}{n}}\,\sqrt{\mathbb{E}[S]\log\Big(\frac{C_{\eta,K}}{\mathbb{E}[S]}\Big)} \\
&\leq C_{\eta,K}\sqrt{\frac{\max\{V_1,\log(s_1)\}}{n}}\,\sqrt{\big(C_{\eta,K}\phi(r)+r\big)\log\Big(\frac{C_{\eta,K}}{r}\Big)}. \qquad (2.17)
\end{aligned}$$

where in the second inequality we have used Lemma 3.8 in [Mendelson 2003], and in the third inequality we have used Jensen's inequality since the function $x \mapsto \sqrt{x\log(C_{\eta,K}/x)}$ is concave, and in the last inequality we have used the upper and lower bounds of $\mathbb{E}[S]$ obtained as above.

Recall that r^* is a fixed point of $\phi(r)$, i.e., $\phi(r^*) = r^*$. Taking the limit $r \to r^*$ and by the continuity of $\phi(r)$ [Bartlett et al. 2005], there holds

$$\begin{aligned}
& r^* \leq C_{\eta,K}\sqrt{\frac{\max\{V_1,\log(s_1)\}}{n}r^*\log\left(\frac{C_{\eta,K}}{r^*}\right)} \\
\Longrightarrow\ & r^* \leq C_{\eta,K}\frac{\max\{V_1,\log(s_1)\}}{n}\log(\frac{C_{\eta,K}}{r^*}) \\
\Longrightarrow\ & r^* \leq C_{\eta,K}\max\{V_1,\log(s_1)\}\frac{\log(n)}{n}.
\end{aligned}$$

The proof of the first case is complete.

Second case: suppose Assumption 4 holds.

According to (2.16), there holds

$$\log \mathcal{N}\Big(\mathcal{G}_r^*, L^2_{\rho_n}, t\Big) \leq s_1'\Big(\frac{1}{t}\Big)^{V_1'} + \log\left(\left\lceil\frac{2\eta K}{t}\right\rceil\right)$$

Then, (2.13) can be bounded as follows

$$\begin{aligned}
\phi(r) &\leq \frac{C}{\sqrt{n}}\mathbb{E}\bigg[\int_0^{\sqrt{S}}\sqrt{s_1'\Big(\frac{1}{t}\Big)^{V_1'}}+\sqrt{\log\bigg(\Big\lceil\frac{2\eta K}{r}\Big\rceil\bigg)}dt\bigg]\\
&\leq \frac{C_{\eta,K}}{\sqrt{n}}\left(\mathbb{E}\Big[\int_0^{\sqrt{S}}\sqrt{s_1'}\Big(\frac{1}{t}\Big)^{\frac{V_1'}{2}}dt\Big]+\sqrt{\big(C_{\eta,K}\phi(r)+r\big)\log\Big(\frac{C_{\eta,K}}{r}\Big)}\right)\\
&\leq \frac{C_{\eta,K}}{\sqrt{n}}\left(\frac{2\sqrt{s_1'}}{2-V_1'}\mathbb{E}\Big[S^{\frac{2-V_1'}{4}}\Big]+\sqrt{\big(C_{\eta,K}\phi(r)+r\big)\log\Big(\frac{C_{\eta,K}}{r}\Big)}\right)\\
&\leq \frac{C_{\eta,K}}{\sqrt{n}}\left(\frac{2\sqrt{s_1'}}{2-V_1'}\Big(C_{\eta,K}\phi(r)+r\Big)^{\frac{2-V_1'}{4}}+\sqrt{\big(C_{\eta,K}\phi(r)+r\big)\log\Big(\frac{C_{\eta,K}}{r}\Big)}\right),
\end{aligned}$$

where in the first inequality we have used the inequality $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for $a,b\geq 0$, and the second inequality follows from (2.17) directly, and in the last inequality we have used Jensen's inequality with the concave function $x\mapsto(x)^{\frac{2-V_1'}{4}}$.

Taking the limit $r\to r^*$, there holds

$$\begin{aligned}
& r^*\leq\sqrt{\frac{C_{\eta,K}}{n}}\left(\frac{\sqrt{s_1'}(r^*)^{\frac{2-V_1'}{4}}}{2-V_1'}+\sqrt{r^*\log\Big(\frac{C_{\eta,K}}{r^*}\Big)}\right)\\
\implies & r^*\leq\sqrt{\frac{C_{\eta,K}}{n}}\max\left\{\frac{\sqrt{s_1'}(r^*)^{\frac{2-V_1'}{4}}}{2-V_1'},\sqrt{r^*\log\Big(\frac{C_{\eta,K}}{r^*}\Big)}\right\}\\
\implies & r^*\leq C_{\eta,K}\max\left\{\sqrt{s_1'}(2-V_1')^{-\frac{4}{2+V_1'}}\Big(\frac{1}{n}\Big)^{\frac{2}{2+V_1'}},\frac{\log(n)}{n}\right\}\\
\implies & r^*\leq C_{\eta,K}\max\left\{\sqrt{s_1'}\Big(\frac{1}{n}\Big)^{\frac{2}{2+V_1'}},\frac{\log(n)}{n}\right\},
\end{aligned}$$

where in the last inequality we have used the fact $(2-V_1')^{-\frac{4}{2+V_1'}}\leq 1$ since $V_1'\in(0,1)$. The proof of the proposition is complete. □

Combining Lemma 2 and Proposition 2 together, and applying Bernstein concentration inequality to $\mathbb{E}[g_{f_\mathcal{H}}(Z)]-\frac{1}{n}\sum_{i=1}^n g_{f_\mathcal{H}}(Z_i)$, we obtain the following lemma which derives an upper bound for $S_1(\mathcal{H})$.

LEMMA 3. *Assume the same conditions as that in Theorem 1.*

- *If the capacity of $\mathcal{H}$ satisfies Assumption 3, then for any $\delta \in (0, 1/2)$, with probability at least $1 - \delta/2$, there holds*

$$S_1(\mathcal{H}) \leq C_{\eta,L,M,\beta}\Big(\frac{\max\{\log(s_1), V_1\}\log(n)}{n}\Big)^{\frac{1}{2-\beta}}\log(4/\delta) + \frac{1}{2}\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] + \frac{\beta}{2}\mathcal{D}(\mathcal{H}).$$

- *If the capacity of $\mathcal{H}$ satisfies Assumption 4, then for any $\delta \in (0, 1/2)$, with probability at least $1 - \delta/2$, there holds*

$$S_1(\mathcal{H}) \leq C_{\eta,L,M,\beta}\max\left\{\sqrt{s_1'}\Big(\frac{1}{n}\Big)^{\frac{2}{(2+V_1')(2-\beta)}}, \Big(\frac{\log(n)}{n}\Big)^{\frac{1}{2-\beta}}\log(4/\delta)\right\} + \frac{1}{2}\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] + \frac{\beta}{2}\mathcal{D}(\mathcal{H}).$$

PROOF. Recall that in (2.10), we have

$$S_1(\mathcal{H}) = 2\bigg(\mathbb{E}\big[g_{\hat{f}_z}\big] - \frac{1}{n}\sum_{i=1}^{n} g_{\hat{f}_z}(Z_i) + \frac{1}{n}\sum_{i=1}^{n} g_{f_\mathcal{H}}(Z_i) - \mathbb{E}\big[g_{f_\mathcal{H}}\big]\bigg).$$

We first estimate $\frac{1}{n}\sum_{i=1}^{n} g_{f_\mathcal{H}}(Z_i) - \mathbb{E}[g_{f_\mathcal{H}}]$. Notice $\|g_{f_\mathcal{H}} - \mathbb{E}[g_{f_\mathcal{H}}]\|_{L^\infty(\mathcal{Z})} \leq 2\|g_{f_\mathcal{H}}\|_{L^\infty(\mathcal{Z})} \leq 2K\|f_\mathcal{H} - f_\rho\|_{L^\infty(\mathcal{X})} \leq 4K\eta$. Since $g_\mathcal{H}$ is data-independent, we know $g_\mathcal{H}(Z_1), \ldots, g_\mathcal{H}(Z_n)$ are i.i.d. variables. Then, by applying the one-sided Bernstein concentration inequality, for any $\epsilon > 0$, there holds

$$Prob\bigg\{\frac{1}{n}\sum_{i=1}^{n} g_{f_\mathcal{H}}(Z_i) - \mathbb{E}[g_{f_\mathcal{H}}] > \epsilon\bigg\} \leq \exp\left\{-\frac{n\epsilon^2}{2\big(\mathbb{E}[g_{f_\mathcal{H}}^2] + \frac{4}{3}K\eta\epsilon\big)}\right\}.$$

For any $\delta \in (0, 1/2)$, suppose $\epsilon = \epsilon^*$ is the solution of the equation $-\frac{n\epsilon^2}{2(\mathbb{E}[g_{f_\mathcal{H}}^2] + \frac{4}{3}K\eta\epsilon)} = \log\left(\frac{\delta}{4}\right)$. Solving ϵ^* and plug it into the above probability inequality. Then, we know with

probability at least $1-\delta/4$, there holds

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} g_{f_{\mathcal{H}}}(Z_i) - \mathbb{E}[g_{\mathcal{H}}] &\leq \frac{\frac{4}{3}K\eta\log(4/\delta) + \sqrt{\left(\frac{4}{3}K\eta\log(4/\delta)\right)^2 + 2n\mathbb{E}[g_{\mathcal{H}}^2]\log(4/\delta)}}{n} \\ &\leq \frac{8K\eta\log(4/\delta)}{3n} + \sqrt{\frac{2\mathbb{E}[g_{\mathcal{H}}^2]\log(4/\delta)}{n}} \\ &\leq \frac{8K\eta\log(4/\delta)}{3n} + \sqrt{\frac{2M\log(4/\delta)}{n}\mathcal{D}(\mathcal{H})^{\beta}}. \\ &\leq \frac{8K\eta\log(4/\delta)}{3n} + \frac{2-\beta}{2}\Big(\frac{2M\log(4/\delta)}{n}\Big)^{\frac{1}{2-\beta}} + \frac{\beta}{2}\mathcal{D}(\mathcal{H}) \\ &\leq C_{\eta,K,M,\beta}\Big(\frac{1}{n}\Big)^{\frac{1}{2-\beta}}\log(4/\delta) + \frac{\beta}{2}\mathcal{D}(\mathcal{H}), \end{aligned} \tag{2.18}$$

where $\mathcal{D}(\mathcal{H}) = \mathbb{E}[g_{f_{\mathcal{H}}}] = \inf_{f\in\mathcal{H}}\mathcal{E}(f) - \mathcal{E}(f_\rho)$ is the approximation error, in the second inequality we have used the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, in the third inequality we have used Jensen's inequality for conditional expectation and the variance-expectation bound for shifted hypothesis space. Indeed, there holds $\mathbb{E}[g_{\mathcal{H}}^2] \leq M\left(\mathbb{E}[g_{\mathcal{H}}]\right)^{\beta} = M\mathcal{D}(\mathcal{H})^{\beta}$. In the last second inequality we have used Young's inequality $ab \leq \dfrac{1}{p}a^p + \dfrac{1}{q}b^q$ with $a = \sqrt{\frac{2M\log(4/\delta)}{n}}$, $b = \sqrt{\mathcal{D}(\mathcal{H})^{\beta}}$, $p = \frac{2}{2-\beta}$, and $q = \frac{2}{\beta}$, in the last inequality we have used the fact $\left(\log(4/\delta)\right)^{1/(2-\beta)} \leq \log(4/\delta)$ since $\frac{1}{2-\beta} \leq 1$ and $\log(4/\delta) > 1$.

Now we estimate $\mathbb{E}\big[g_{\hat{f}_z}\big] - \frac{1}{n}\sum_{i=1}^{n} g_{\hat{f}_z}(Z_i)$. Notice $g_{\hat{f}_z} \in \mathcal{G} \subset \mathcal{G}^*$ and $\|g_f - \mathbb{E}[g_f]\|_{L^\infty(\mathcal{Z})} \leq 4K\eta$ for any $g_f \in \mathcal{G}^*$. Applying Lemma 2 with $\mathcal{F} = \mathcal{G}^*$, $b = 4K\eta$ and $\kappa = 4$, we know with probability at least $1-\delta/4$, there holds

$$\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] - \frac{1}{n}\sum_{i=1}^{n} g_{\hat{f}_z}(Z_i) \leq \frac{1}{4}\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] + C_{\eta,K,M,\beta}\Big(\big(r^*\big)^{\frac{1}{2-\beta}} + \Big(\frac{\log(4/\delta)}{n}\Big)^{\frac{1}{2-\beta}}\Big)$$

Combining above inequality with (2.18) together, we know with probability at least $1-\delta/2$, there holds

$$S_1(\mathcal{H}) \leq C_{\eta,K,M,\beta}\bigg(\big(r^*\big)^{\frac{1}{2-\beta}} + \Big(\frac{1}{n}\Big)^{\frac{1}{2-\beta}}\log(4/\delta)\bigg) + \frac{1}{2}\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] + \frac{\beta}{2}\mathcal{D}(\mathcal{H}).$$

First case: Suppose Assumption 3 holds.

According to Proposition 2, with probability at least $1-\delta/2$, there holds

$$S_1(\mathcal{H}) \leq C_{\eta,K,M,\beta}\Big(\frac{\max\{V_1, \log(s_1)\}\log(n)}{n}\Big)^{\frac{1}{2-\beta}} \log(4/\delta) + \frac{1}{2}\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] + \frac{\beta}{2}\mathcal{D}(\mathcal{H}). \tag{2.19}$$

Second case: Suppose Assumption 4 holds.

According to Proposition 2, with probability at least $1-\delta/2$, there holds

$$\begin{aligned} S_1(\mathcal{H}) \leq C_{\eta,K,M,\beta} \max\left\{\sqrt{s_1'}\Big(\frac{1}{n}\Big)^{\frac{2}{(2+V_1')(2-\beta)}}, \Big(\frac{\log(n)}{n}\Big)^{\frac{1}{2-\beta}} \log(4/\delta)\right\} \\ + \frac{1}{2}\mathbb{E}\big[g_{\hat{f}_z}(Z)\big] + \frac{\beta}{2}\mathcal{D}(\mathcal{H}). \end{aligned}$$

The proof of the lemma is complete. □

Step 2. Upper bounds for $S_2(\mathcal{H})$

We begin by introducing some notations that will be used frequently later. Let $\mathcal{Q} := \{h(z,z') : h(z,z') = h(z',z) \text{ for almost } z, z' \in \mathcal{Z}\}$ be some class of symmetric functions from $\mathcal{Z}\times\mathcal{Z}$ to $\mathbb{R}$. Denote a empirical probability measure on $\mathcal{Z}\times\mathcal{Z}$ by $\xi := \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}\delta_{(Z_i,Z_j)}$. For any $h \in \mathcal{Q}$, we define the L_ξ^2 norm of h by

$$\|h\|_{L_\xi^2} = \left(\frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}|h(Z_i,Z_j)|\right)^{\frac{1}{2}}.$$

Further, we define $\mathcal{W} := \{\hat{h}_f : f \in \mathcal{H}\}$ as the class consisting of the kernel of the degenerate U-statistics defined before.

We introduce a probability bound of the degenerate U-statistic, which controls the degenerate U-statistic in terms of the expectations of the corresponding Rademacher chaos and Rademacher complexities.

LEMMA 4 ([Clémençon et al. 2008]). *Define the supremum of a degenerate U-statistic over the hypothesis space $\mathcal{H}$ as*

$$Z = \sup_{f\in\mathcal{H}}\left|\sum_{i\neq j=1}^{n}\hat{h}_f(Z_i,Z_j)\right|,$$

where $\hat{h}_f(Z_i, Z_j) = \mathbb{E}[U_n^f] - h_f(Z_i) - h_f(Z_j) - q_f(Z_i, Z_j)$. *Then there exists an absolute constant* $C > 0$ *such that for all* n *and* $t > 0$,

$$Prob\big\{Z > C\mathbb{E}[Z_\epsilon]+t\big\} \leq \exp\left(-\frac{1}{C}\min\left(\left(\frac{t}{\mathbb{E}[U_\epsilon]}\right)^2, \frac{t}{\mathbb{E}[M]+Fn}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right),$$

where $\epsilon_1, ..., \epsilon_n$ *are i.i.d. Rademacher variables and* $M = \sup_{f\in\mathcal{H}, k=1,...,n} \big|\sum_{i=1}^n \epsilon_i \hat{h}_f(Z_i, Z_k)\big|$, $F = \sup_{f\in\mathcal{H}} \|\hat{h}_f\|_{L^\infty(\mathcal{Z}\times\mathcal{Z})}$, $Z_\epsilon = \sup_{f\in\mathcal{H}} \big|\sum_{i\neq j=1}^n \epsilon_i\epsilon_j \hat{h}_f(Z_i, Z_j)\big|$ *and* $U_\epsilon = \sup_{f\in\mathcal{H}} \sup_{\alpha:\|\alpha\|_2\leq 1} \sum_{i\neq j=1}^n \epsilon_i\alpha_j \hat{h}_f(Z_i, Z_j)$.

Notice that

$$S_2(\mathcal{H}) = -W_n^{\hat{f}_z} + W_n^{f_\mathcal{H}} \leq \frac{2}{n(n-1)} Z.$$

For any $\delta \in (0, 1/2)$, we set the equation

$$\exp\left(-\frac{1}{C}\min\left(\left(\frac{t}{\mathbb{E}[U_\epsilon]}\right)^2, \frac{t}{\mathbb{E}[M]+Fn}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right) = \delta/2,$$

and solve it for variable t, then by Lemma 4 we know with probability at least $1 - \delta/2$, there holds

$$S_2(\mathcal{H}) \leq \frac{C\log^2(2/\delta)}{n^2}\Big(\mathbb{E}[Z_\epsilon] + \mathbb{E}[U_\epsilon] + \mathbb{E}[M] + Fn\Big).$$

Therefore, to derive upper bounds for $S_2(\mathcal{H})$, we suffice to estimate the Rademchaer chaos $\mathbb{E}[Z_\epsilon]$ and the Rademacher complexities $\mathbb{E}[U_\epsilon]$ and $\mathbb{E}[M]$.

We first estimate $\mathbb{E}[Z_\epsilon]$. Since the Rademacher chaos $Z_\epsilon = \sup_{f\in\mathcal{H}} \big|\sum_{i\neq j=1}^n \epsilon_i\epsilon_j \hat{h}_f(Z_i, Z_j)\big|$ conditioned on sample S is not sub-Gaussian with respect to the metric on $\mathcal{H}$ induced by any empirical norm of $\hat{h}_f$, then we cannot apply the classical chaining methods directly [Vershynin 2018; Wainwright 2019]. Instead, we introduce two lemmas that estimate $\mathbb{E}[Z_\epsilon]$ directly. The first lemma from [Ying and Campbell 2010] established a maximal inequality of Z_ϵ. The second lemma is very similar to Theorem 2 in [Ying and Campbell 2010], which controls $\mathbb{E}[Z_\epsilon]$ in terms of an entropy integral.

LEMMA 5 ([Ying and Campbell 2010]). *Let $\{h_1, ..., h_N\} \subset \mathcal{Q}$ be a finite class of functions contained in $\mathcal{Q}$, and $\{\epsilon_1, \ldots, \epsilon_N\}$ be i.i.d. Rademacher variables, then there holds*

$$\mathbb{E}\left[\max_{k\in\{1,\ldots,N\}} \left|\frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \epsilon_i\epsilon_j h_k(Z_i, Z_j)\right| \middle| Z_1, \ldots, Z_n\right] \leq 2\sqrt{2}en\log(N+1) \max_{k\in\{1,\ldots,N\}} \|h_k\|_{L^2_\xi}.$$

With the above maximal inequality, we can bound the Rademacher chaos in terms of an entropy integral by using the same arguments in classical chaining methods [Vershynin 2018; Wainwright 2019].

LEMMA 6. *Let $\{\epsilon_1, \ldots, \epsilon_N\}$ be i.i.d. Rademacher variables, then there holds*

$$\mathbb{E}\left[\sup_{h\in\mathcal{Q}} \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \epsilon_i\epsilon_j h(Z_i, Z_j) \middle| Z_1, \ldots, Z_n\right] \leq n\Big(72\sqrt{2}e \int_0^{D/2} \log\Big(\mathcal{N}\big(\mathcal{Q}, L^2_\xi, t\big)\Big)\, dt + D\Big),$$

where $D = \sup_{h\in\mathcal{Q}} \|h\|_{L^2_\xi}$ (Note that $D = D(Z_1, \ldots, Z_n)$ depends on the sample).

PROOF. For each $k \in \mathbb{N}$, let $\mathcal{N}_k$ be a $\frac{D}{2^k}$-net of $\mathcal{Q}$ with respect to the metric L^2_ξ. From the definition of the net, we can define a map ω_k from $\mathcal{Q}$ to $\mathcal{N}_k$ such that $\big\|h - \omega_k(h)\big\|_{L^2_\xi} \leq \frac{D}{2^k}$ for all $h \in \mathcal{Q}$. Then $\sup_{h\in\mathcal{Q}} \|h - \omega_k(h)\|_{L^2_\xi} \leq \frac{D}{2^k} \to 0$ as $k \to \infty$. Further, notice that $\sup_{h\in\mathcal{Q}} \big|\sum_{i\neq j=1}^{n} \epsilon_i\epsilon_j\big(h - \omega_k(h)\big)(Z_i, Z_j)\big|$ is uniformly bounded for any k. Then, by Dominated Convergence Theorem, we know $\mathbb{E}\big[\sup_{h\in\mathcal{Q}} \big|\sum_{i\neq j=1} \epsilon_i\epsilon_j\big(h - \omega_k(h)\big)(Z_i, Z_j)\big| Z_1, \ldots, Z_n\big] \to 0$ as $k \to \infty$.

Take an arbitrary $h_0 \in \mathcal{Q}$, we can set $\mathcal{N}_0 = \{h_0\}$ since $\omega_0(h) = h_0$ for any $h \in \mathcal{Q}$. Then, conditioned on the sample $\{Z_1, \ldots, Z_n\}$, there holds

$$
\begin{aligned}
&\mathbb{E}\Big[\sup_{h\in\mathcal{Q}}\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j h(Z_i,Z_j)\Big|\Big] \\
&\leq \mathbb{E}\Big[\sup_{h\in\mathcal{Q}}\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j\big(h-h_0\big)(Z_i,Z_j)\Big|\Big] + \mathbb{E}\Big[\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j h_0(Z_i,Z_j)\Big|\Big] \\
&\leq \mathbb{E}\Big[\sup_{h\in\mathcal{Q}}\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j\big(h-h_0\big)(Z_i,Z_j)\Big|\Big] + \Big(\mathbb{E}\Big[\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j h_0(Z_i,Z_j)\Big|^2\Big]\Big)^{1/2} \\
&= \mathbb{E}\Big[\sup_{h\in\mathcal{Q}}\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j\big(h-h_0\big)(Z_i,Z_j)\Big|\Big] + \sqrt{n(n-1)}\big\|h_0\big\|_{L^2_\xi} \\
&\leq \mathbb{E}\Big[\sup_{h\in\mathcal{Q}}\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j\big(h-\omega_k(h)\big)(Z_i,Z_j)\Big|\Big] + \sum_{m=1}^{k}\mathbb{E}\Big[\sup_{h\in\mathcal{Q}}\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j\big(\omega_m(h)-\omega_{m-1}(h)\big)(Z_i,Z_j)\Big|\Big] \\
&\quad + nD \\
&\to \sum_{k=1}^{\infty}\mathbb{E}\Bigg[\sup_{\substack{h\in\mathcal{Q},\\ (\omega_k(h),\omega_{k-1}(h))\in\\ \mathcal{N}_k\times\mathcal{N}_{k-1}}}\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j\big(\omega_k(h)-\omega_{k-1}(h)\big)(Z_i,Z_j)\Big|\Bigg] + nD \qquad (2.20)
\end{aligned}
$$

as $k \to \infty$, where in the second inequality we have used Jensen's inequality, and in the first equality we just expand the term $\big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j h_0\big|^2$ and compute its expectation, and in the last step we have used Dominated Convergence Theorem.

Denote by $|A|$ the cardinality of a set A, we know $\big|\mathcal{N}_k\times\mathcal{N}_{k-1}\big| = \big|\mathcal{N}_k\big|\big|\mathcal{N}_{k-1}\big| \leq \mathcal{N}\big(\mathcal{Q}, L^2_\xi, \frac{D}{2^k}\big)^2$. Notice that for any $h \in \mathcal{Q}$ and $k \in \mathbb{N}^+$, $\big\|\omega_k(h)-\omega_{k-1}(h)\big\|_{L^2_\xi} \leq \big\|\omega_k(h)-h\big\|_{L^2_\xi} + \big\|h-$

$\omega_{k-1}(h)\big\|_{L^2_\xi} \leq \frac{3}{2^{k-1}}D$. Then by Lemma 5, (2.20) can be bounded as follows

$$\begin{aligned}
\mathbb{E}\bigg[\sup_{h\in\mathcal{Q}}\Big|\sum_{i\neq j=1}^{n}\epsilon_i\epsilon_j h\Big|\bigg|Z_1,\ldots,Z_n\bigg] &\leq n\bigg(2\sqrt{2}e\sum_{k=1}^{\infty}\log\Big(\mathcal{N}\big(\mathcal{Q},L^2_\xi,D/2^k\big)^2+1\Big)\frac{3D}{2^{k-1}}+D\bigg)\\
&\leq n\bigg(24\sqrt{2}e\sum_{k=1}^{\infty}\int_{\frac{D}{2^{k+1}}}^{\frac{D}{2^k}}\log\Big(\mathcal{N}\big(\mathcal{Q},L^2_\xi,t\big)^2+1\Big)\,dt+D\bigg)\\
&= n\bigg(24\sqrt{2}e\int_0^{\frac{D}{2}}\log\Big(\mathcal{N}\big(\mathcal{Q},L^2_\xi,t\big)^2+1\Big)\,dt+D\bigg)\\
&\leq n\bigg(72\sqrt{2}e\int_0^{\frac{D}{2}}\log\big(\mathcal{N}\big(\mathcal{Q},L^2_\xi,t\big)\big)\,dt+D\bigg),
\end{aligned}$$

where in the last inequality we have used the inequality $\mathcal{N}\big(\mathcal{Q},L^2_\xi,t\big)^2+1\leq\mathcal{N}\big(\mathcal{Q},L^2_\xi,t\big)^3$ since $\mathcal{N}\big(\mathcal{Q},L^2_\xi,t\big)\geq 2$ for $t\leq\frac{D}{2}$. This completes the proof. □

Applying Lemma 6 with $\mathcal{Q}=\mathcal{W}=\{\hat{h}_f : f\in\mathcal{H}\}$, we know

$$\mathbb{E}\big[Z_\epsilon\big|Z_1,\ldots,Z_n\big]\leq n\bigg(72\sqrt{2}e\int_0^{F}\log\Big(\mathcal{N}\big(\mathcal{W},L^2_\xi,t\big)\Big)\,dt+F\bigg), \tag{2.21}$$

where $F:=\sup_{f\in\mathcal{H}}\big\|\hat{h}_f\big\|_{L^\infty(\mathcal{Z}\times\mathcal{Z})}\geq D$.

Now we estimate $\mathbb{E}[U_\epsilon]$. Denote by $H_f\in\mathbb{R}^{n\times n}$ a square matrix with zero diagonal such that its entries $(H_f)_{i,j}=\hat{h}_f(Z_i,Z_j)$ for $i\neq j$, and let $\epsilon=(\epsilon_1,\ldots,\epsilon_n)\in\mathbb{R}^n$ be i.i.d. Rademacher variables. Then, we know

$$\begin{aligned}
\mathbb{E}[U_\epsilon^2]=\mathbb{E}\Big[\Big(\sup_{f\in\mathcal{H}}\sup_{\|\alpha\|_2\leq 1}\alpha^\top H_f\epsilon\Big)^2\Big]&=\mathbb{E}\Big[\Big(\sup_{f\in\mathcal{H}}\|H_f\epsilon\|_2\Big)^2\Big]\\
&\leq\mathbb{E}\Big[\sup_{f\in\mathcal{H}}\|H_f\epsilon\|_2^2\Big]=\mathbb{E}\Big[\sup_{f\in\mathcal{H}}\epsilon^\top H_f^2\epsilon\Big].
\end{aligned}\tag{2.22}$$

From above, we know the estimates of $\mathbb{E}[U_\epsilon]$ reduce to the estimates of the Rademacher chaos. Indeed, we first suppose that $\mathcal{Q}=\{J_f : f\in\mathcal{H}\}$ is any class such that

$$J_f(Z_i,Z_j)=\big(H_f^2\big)_{i,j}=\sum_{k\neq i,k\neq j=1}^{n}\hat{h}_f(Z_i,Z_k)\hat{h}_f(Z_j,Z_j) \tag{2.23}$$

for all $f \in \mathcal{H}$. Then, there holds

$$\mathbb{E}\Big[\sup_{f\in\mathcal{H}} \epsilon^\top H_f^2 \epsilon\Big] = \mathbb{E}\Big[\sup_{f\in\mathcal{H}} \sum_{i,j=1}^n \epsilon_i\epsilon_j J_f(Z_i,Z_j)\Big]$$
$$= \mathbb{E}\Big[\sup_{f\in\mathcal{H}} \sum_{i\neq j=1}^n \epsilon_i\epsilon_j J_f(Z_i,Z_j)\Big] + \mathbb{E}\Big[\sup_{f\in\mathcal{H}} \sum_{i\neq j=1} \hat{h}_f^2(Z_i,Z_j)\Big]. \tag{2.24}$$

The first term of the above equality is a Rademacher chaos, which will be estimated using Lemma 6. The second term equals $\mathbb{E}\big[\sup_{f\in\mathcal{H}} \|\hat{h}_f\|_{L_\xi^2}\big]$, which will be controlled by the uniform boundedness of $\hat{h}_f$ directly. Combining (2.22) and (2.24) together, and bounding the covering number of $\mathcal{Q}$ by that of $\mathcal{W}$, we can control $\mathbb{E}[U_\epsilon]$ as follows

LEMMA 7. *Let $D = \sup_{f\in\mathcal{H}} \big\|\hat{h}_f\big\|_{L_\xi^2}$, there holds*

$$\mathbb{E}\big[U_\epsilon \big| Z_1,\ldots,Z_n\big] \leq CnD\sqrt{\int_0^{1/2} \log\big(\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big)\big)\, dt + 1}.$$

PROOF. According to (2.22) and (2.24), there holds

$$\mathbb{E}[U_\epsilon^2|Z_1,\ldots,Z_n] = \mathbb{E}\Big[\sup_{f\in\mathcal{H}} \sum_{i\neq j=1}^n \epsilon_i\epsilon_j J_f(Z_i,Z_j)\Big|Z_1,\ldots,Z_n\Big] + \mathbb{E}\Big[\sup_{f\in\mathcal{H}} \sum_{i\neq j=1}^n \hat{h}_f^2(Z_i,Z_j)\Big|Z_1,\ldots,Z_n\Big]$$
$$\leq n\bigg(72\sqrt{2}e\int_0^{D/2} \log\big(\mathcal{N}\big(\mathcal{Q}, L_\xi^2, t\big)\big)\, dt + D\bigg) + n^2 D$$
$$\leq Cn^2 D\Big(\int_0^{D/2} \log\big(\mathcal{N}\big(\mathcal{Q}, L_\xi^2, t\big)\big)\, dt + 1\Big),$$

where in the first inequality we have used Lemma 6 and the fact $D = \sup_{f\in\mathcal{H}} \|\hat{h}_f\|_{L_\xi^2}$.

Now, we are in a position to estiamte the covering number $\mathcal{N}\big(\mathcal{Q}, L_\xi^2, t\big)$. Denote by $\|H\|_F = \sqrt{\sum_{i,j=1}^2 (H_{i,j})^2}$ the Frobenius norm of a matrix H. For any $f \in \mathcal{H}$, the L_ξ^2 norm of J_f (defined in (2.23)) can be bounded as follows

$$\big\|J_f\big\|_{L_\xi^2} = \bigg(\frac{1}{n(n-1)}\sum_{i\neq j=1}^n J_f^2(Z_i,Z_j)\bigg)^{\frac{1}{2}} \leq \bigg(\frac{1}{n(n-1)}\sum_{i,j=1}^n J_f^2(Z_i,Z_j)\bigg)^{\frac{1}{2}}$$
$$= \frac{1}{\sqrt{n(n-1)}}\big\|H_f^2\big\|_F \leq \frac{1}{\sqrt{n(n-1)}}\big\|H_f\big\|_F^2 = \big\|\hat{h}_f\big\|_{L_\xi^2}^2 \leq D\big\|\hat{h}_f\big\|_{L_\xi^2}.$$

Therefore, we know $\mathcal{N}\big(\mathcal{Q}, L_\xi^2, t\big) \leq \mathcal{N}\big(\mathcal{W}, L_\xi^2, t/D\big)$ for any $t \in (0, D/2)$. Then, there holds

$$\begin{aligned}\mathbb{E}[U_\epsilon^2|Z_1,\ldots,Z_n] &\leq Cn^2D\Big(\int_0^{D/2}\log\big(\mathcal{N}\big(\mathcal{W}, L_\xi^2, t/D\big)\big)\,dt+1\Big)\\ &= Cn^2D^2\Big(\int_0^{1/2}\log\big(\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big)\big)\,dt+1\Big).\end{aligned}$$

By Jensen's inequality for conditional expectation, we know

$$\mathbb{E}[U_\epsilon|Z_1,\ldots,Z_n] \leq \big(\mathbb{E}[U_\epsilon^2|Z_1,\ldots,Z_n]\big)^{1/2}.$$

Then, we finally conclude that

$$\mathbb{E}\big[U_\epsilon|Z_1,\ldots,Z_n\big] \leq CnD\sqrt{\int_0^{1/2}\log\big(\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big)\big)\,dt+1}.$$

The proof of this lemma is complete. □

For the last term $\mathbb{E}[M]$, we can control it just by the uniform boundedness of the class $\mathcal{W}$. Indeed, recall that $F = \sup_{f\in\mathcal{H}}\big\|\hat{h}_f\big\|_{L^\infty(\mathcal{Z}\times\mathcal{Z})}$, then we immediately get the upper bounds of $\mathbb{E}[M]$ as follows

$$\mathbb{E}[M] \leq \sup_{f\in\mathcal{H},k=1,\ldots,n}\sum_{i=1}^n\big|\hat{h}_f(Z_i,Z_k)\big| \leq nF. \tag{2.25}$$

From (2.21) and Lemma 7, we know the only thing left now is the estimates of the covering number $\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big)$. Combining all the above results, and carefully controlling $\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big)$ in terms of the capacity of the hypothesis space $\mathcal{H}$, we can establish an upper bound for $S_2(\mathcal{H})$ as the following proposition.

LEMMA 8. *Suppose Assumptions 1, 2, and* (2.4) *hold, and the hypothesis space $\mathcal{H}$ is uniformly bounded by $\eta > 0$.*

- *If the capacity of $\mathcal{H}$ satisfies Assumption 3, then for any $\delta \in (0, 1/2)$, with probability at least $1 - \delta/2$, there holds*

$$S_2(\mathcal{H}) \leq C_{\eta,L}\max\{V_1, V_2, \log(s_1), \log(s_2)\}\frac{\log^2(\delta/2)}{n}.$$

- *If the capacity of $\mathcal{H}$ satisfies Assumption 4, then for any $\delta \in (0, 1/2)$, with probability at least $1 - \delta/2$, there holds*

$$S_2(\mathcal{H}) \leq C_{\eta,L} \max\{s_1', s_2'\} \frac{\log^2(\delta/2)}{n} \frac{1}{1 - \max\{V_1', V_2'\}}.$$

The following lemma studies the covering numbers of the sum of the function classes, which will be used in the proof of Lemma 8.

LEMMA 9. *Let $\big(\Omega, \Gamma, p\big)$ be a measure space. Suppose $\mathcal{W}_1, \ldots, \mathcal{W}_m \subset L^2\big(\Omega, \Gamma, p\big)$ and define $\bigoplus_{i=1}^m \mathcal{W}_i := \big\{ \sum_{i=1}^m w_i : w_i \in \mathcal{W}_i \big\}$ as the direct sum of the classes $\mathcal{W}_i$ for $i = 1 \ldots, m$. Then, for any $t > 0$, the covering numbers of above function classes satisfying*

$$\mathcal{N}\Big(\bigoplus_{i=1}^m \mathcal{W}_i, L_p^2, t \Big) \leq \prod_{i=1}^m \mathcal{N}\big(\mathcal{W}_i, L_p^2, t/m\big).$$

PROOF. Let $N_i \subset \mathcal{W}_i$ be a $\frac{t}{m}$-net of $\mathcal{W}_i$ for $i = 1, \cdots, m$. Then, we claim that $N = \bigoplus_{i=1}^m N_i := \{\sum_{i=1}^m w_i^* : w_i^* \in \mathbb{N}_i\}$ is a t-net of $\bigoplus_{i=1}^m \mathcal{W}_i$. Indeed, for any $\sum_{i=1}^m w_i \in \bigoplus_{i=1}^m \mathcal{W}_i$, we know there exists a $w_i^* \in \mathcal{W}_i$ such that $\|w_i - w_i^*\|_{L_p^2} \leq t/m$ for $i = 1, \ldots, m$. Then, we have $\| \sum_{i=1}^m w_i - \sum_{i=1}^m w_i^* \|_{L_p^2} \leq \sum_{i=1}^m \|w_i - w_i^*\|_{L_p^2} \leq t$, which means N is t-net of $\bigoplus_{i=1}^m \mathcal{W}_i$. Further, notice that $|N| \leq \prod_{i=1}^m |N_i|$ and N_i is an arbitrary $\frac{t}{m}$-net of $\mathcal{W}_i$. Then, the proof of this lemma is complete. □

We are in a position to prove Lemma 8.

PROOF OF LEMMA 8. Recall that by Lemma 4, we know with probability at least $1-\delta/2$, there holds

$$S_2(\mathcal{H}) \leq \frac{C \log^2(2/\delta)}{n^2} \Big(\mathbb{E}[Z_\epsilon] + \mathbb{E}[U_\epsilon] + \mathbb{E}[M] + Fn \Big).$$

Notice that the constant $F = \sup_{f \in \mathcal{H}} \big\| \hat{h}_f \big\|_{L^\infty(\mathcal{Z} \times \mathcal{Z})} \leq 4 \sup_{f \in \mathcal{H}} \big\| q_f \big\|_{L^\infty(\mathcal{Z} \times \mathcal{Z})} \leq 4K \sup_{f \in \mathcal{H}} \big\| f - f_\rho \big\|_{L^\infty(\mathcal{X} \times \mathcal{X})} \leq 8\eta K$. According to (2.21), Lemma 7, and (2.25), with probability at least $1 - \delta/2$, there holds

$$S_2(\mathcal{H}) \leq C_{\eta,K} \frac{1}{n} \mathbb{E}\bigg[\Big(\int_0^{8\eta K} \log\Big(\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big) \Big)\, dt + \sqrt{\int_0^{1/2} \log\big(\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big) \big)\, dt} + 1 \Big) \bigg].$$

Now, we are in a position to estimate the covering number $\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big)$. We first define three classes of functions from $\mathcal{Z} \times \mathcal{Z}$ to $\mathbb{R}$ as follows

$$\begin{aligned}
\mathcal{W}_1 &= \big\{w_f^1(z, z') = \mathbb{E}[q_f(Z, Z')] : f \in \mathcal{H}\big\} \\
\mathcal{W}_2 &= \big\{w_f^2(z, z') = -h_f(z) - h_f(z') : f \in \mathcal{H}\big\} \\
\mathcal{W}_3 &= \big\{w_f^3(z, z') = q_f(z, z') : f \in \mathcal{H}\big\}.
\end{aligned}$$

It can be seen that for any $f \in \mathcal{H}$, $\hat{h}_f = w_f^1 + w_f^2 + w_f^3$. Then, it follows that $\mathcal{W} \subset \mathcal{W}_1 + \mathcal{W}_2 + \mathcal{W}_3$. According to Lemma 9, for any $t > 0$ and Exercise 4.2.10 in [Vershynin 2018], there holds

$$\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big) \leq \mathcal{N}\big(\mathcal{W}_1, L_\xi^2, t/6\big)\mathcal{N}\big(\mathcal{W}_2, L_\xi^2, t/6\big)\mathcal{N}\big(\mathcal{W}_3, L_\xi^2, t/6\big).$$

We will estimate $\mathcal{N}\big(\mathcal{W}_i, L_\xi^2, t\big)$ for $i = 1, 2, 3$ separately. For any $f_1, f_2 \in \mathcal{H}$, there hold

$$\begin{aligned}
\big\|w_{f_1}^1 - w_{f_2}^1\big\|_{L_\xi^2} &= \big|\mathbb{E}\big[\ell(f_1(X, X'), Y, Y') - \ell(f_2(X, X'), Y, Y')\big]\big| \leq K\big\|f_1 - f_2\big\|_{L^\infty(\mathcal{X}\times\mathcal{X})}, \\
\big\|w_{f_1}^2 - w_{f_2}^2\big\|_{L_\xi^2} &\leq \big\|h_{f_1}(z) - h_{f_2}(z)\big\|_{L_\xi^2} + \big\|h_{f_1}(z') - h_{f_2}(z')\big\|_{L_\xi^2} \\
&= 2\Big(\frac{1}{n}\sum_{i=1}^n \big|h_{f_1}(Z_i) - h_{f_2}(Z_i)\big|^2\Big)^{1/2} \\
&\leq 2K\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big|f_1(X, X_i) - f_2(X, X_i)\big|^2\Big|X_i\Big]\Big)^{1/2} = 2K\big\|f_1 - f_2\big\|_{L^2_{\rho_{\mathbf{x}}\times\mu_n}}, \\
\big\|w_{f_1}^3 - w_{f_2}^3\big\|_{L_\xi^2} &= \Big(\frac{1}{n(n-1)}\sum_{i\neq j=1}^n \big|q_{f_1}(Z_i, Z_j) - q_{f_2}(Z_i, Z_j)\big|^2\Big)^{1/2} \\
&\leq K\Big(\frac{1}{n(n-1)}\sum_{i\neq j=1}^n \big|f_1(X_i, X_j) - f_2(X_i, X_j)\big|^2\Big)^{1/2} = K\big\|f_1 - f_2\big\|_{L^2_{\nu_n}}.
\end{aligned}$$

Therefore, for any $t > 0$, we have

$$\begin{aligned}
\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big) &\leq \mathcal{N}\big(\mathcal{H}, L^\infty(\mathcal{X}\times\mathcal{X}), t/6K\big)\mathcal{N}\big(\mathcal{H}, L^2_{\rho_{\mathbf{x}}\times\mu_n}, t/12K\big)\mathcal{N}\big(\mathcal{H}, L^2_{\nu_n}, t/6K\big) \\
&\leq \mathcal{N}\big([-8\eta K, 8\eta K], |\cdot|, t/6K\big)\mathcal{N}\big(\mathcal{H}, L^2_{\rho_{\mathbf{x}}\times\mu_n}, t/12K\big)\mathcal{N}\big(\mathcal{H}, L^2_{\nu_n}, t/6K\big) \\
&\leq C_{\eta,K}\frac{1}{t}\mathcal{N}\big(\mathcal{H}, L^2_{\rho_{\mathbf{x}}\times\mu_n}, t/12K\big)\mathcal{N}\big(\mathcal{H}, L^2_{\nu_n}, t/6K\big). \qquad (2.26)
\end{aligned}$$

First case: Suppose Assumption 3 holds.

From (2.26), we know

$$\begin{aligned}
&\log\Big(\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big)\Big) \\
&\le \log(C_{\eta,K}) + \log\Big(\frac{1}{t}\Big) + \log(s_1) + V_1 \log\Big(\frac{12K}{t}\Big) + \log(s_2) + V_2 \log\Big(\frac{6K}{t}\Big) \\
&\le C_{\eta,K} \max\{V_1, V_2, \log(s_1), \log(s_2)\} \log\Big(\frac{12K}{t}\Big).
\end{aligned}$$

Then, with probability at least $1-\delta/2$, $S_2(\mathcal{H})$ can be bounded as

$$\begin{aligned}
S_2(\mathcal{H}) &\le C_{\eta,K} \frac{\log^2(\delta/2)}{n} \bigg(\max\{V_1, V_2, \log(s_1), \log(s_2)\} \int_0^{8\eta K} \log\Big(\frac{12K}{t}\Big) dt \\
&\qquad + \sqrt{\max\{V_1, V_2, \log(s_1), \log(s_2)\} \int_0^{1/2} \log\Big(\frac{12K}{t}\Big) dt} + 1 \bigg) \\
&\le C_{\eta,K} \frac{\log^2(\delta/2) \max\{V_1, V_2, \log(s_1), \log(s_2)\}}{n}.
\end{aligned}$$

Second case: Suppose Assumption 4 holds.

From (2.26), we know

$$\begin{aligned}
\log\Big(\mathcal{N}\big(\mathcal{W}, L_\xi^2, t\big)\Big) &\le \log(C_{\eta,K}) + \log\Big(\frac{1}{t}\Big) + s_1' \Big(\frac{12K}{t}\Big)^{V_1'} + s_2' \Big(\frac{6K}{t}\Big)^{V_2'} \\
&\le C_{\eta,K} \max\{s_1', s_2'\} \Big(\frac{12K}{t}\Big)^{\max\{V_1', V_2'\}}.
\end{aligned}$$

Then, with probability at least $1-\delta/2$, $S_2(\mathcal{H})$ can be bounded as

$$\begin{aligned}
S_2(\mathcal{H}) &\le C_{\eta,K} \frac{\log^2(\delta/2)}{n} \bigg(\int_0^{8\eta K} \max\{s_1', s_2'\} \Big(\frac{12K}{t}\Big)^{\max\{V_1', V_2'\}} dt \\
&\qquad + \sqrt{\int_0^{1/2} \max\{s_1', s_2'\} \Big(\frac{12K}{t}\Big)^{\max\{V_1', V_2'\}} dt} + 1 \bigg) \\
&= C_{\eta,K} \max\{s_1', s_2'\} \frac{\log^2(\frac{\delta}{2})}{n} \bigg(\frac{(12K)^{\max\{V_1', V_2'\}} (8\eta K)^{1-\max\{V_1', V_2'\}}}{1-\max\{V_1', V_2'\}} \\
&\qquad + \sqrt{\frac{(12K)^{\max\{V_1', V_2'\}} (\frac{1}{2})^{1-\max\{V_1', V_2'\}}}{1-\max\{V_1', V_2'\}}} + 1 \bigg) \\
&\le C_{\eta,K} \max\{s_1', s_2'\} \frac{\log^2(\frac{\delta}{2})}{n} \frac{1}{1-\max\{V_1', V_2'\}}.
\end{aligned}$$

This completes the proof of Lemma 8. □

Now, we are in the position to present the proof of our main result.

PROOF OF THEOREM 1. Notice that $\mathbb{E}\big[g_{\hat{f}_z}\big] = \mathbb{E}\big[\ell(\hat{f}_z(X,X'),Y,Y') - \ell(\hat{f}_\rho(X,X'),Y,Y')\big] = \mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho)$ is the excess generalization error. Combining Lemma 3 and Lemma 8 together, there holds

(a) If the capacity of $\mathcal{H}$ satisfies Assumption 3, then with probability at least $1-\delta$, there holds

$$\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho) \leq C_{\eta,L,M,\beta}\Big(\frac{\max\{\log(s_1), V_1\}\log(n)}{n}\Big)^{\frac{1}{2-\beta}}\log(4/\delta)$$
$$+ C_{\eta,L}\frac{\log^2(\delta/2)\max\{V_1, V_2, \log(s_1), \log(s_2)\}}{n} + \frac{1}{2}\big(\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho)\big) + \Big(\frac{\beta}{2}+1\Big)\mathcal{D}(\mathcal{H}).$$

(b) If the capacity of $\mathcal{H}$ satisfies Assumption 4, then with probability at least $1-\delta$, there holds

$$\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho) \leq C_{\eta,L,M,\beta}\max\left\{\Big(\sqrt{s_1'}\frac{1}{n}\Big)^{\frac{2}{(2+V_1')(2-\beta)}}, \Big(\frac{\log(n)}{n}\Big)^{\frac{1}{2-\beta}}\log(4/\delta)\right\}$$
$$+ C_{\eta,L}\max\{s_1', s_2'\}\frac{\log^2(\delta/2)}{n\big(1-\max\{V_1', V_2'\}\big)} + \frac{1}{2}\big(\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho)\big) + \Big(\frac{\beta}{2}+1\Big)\mathcal{D}(\mathcal{H}).$$

By rearranging the terms of the above inequalities, we can obtain the desired results. □

2.3.2 Proofs for Section 2.2

We first present the proof of Proposition 1 as follows.

PROOF OF PROPOSITION 1. By tower property of the conditional expectation, the generalization error with a predictor f can be written as

$$\mathcal{E}(f) = \mathbb{E}\big[\mathbb{E}\big[\ell\big(f(X,X'),Y,Y'\big)\big|X,X'\big]\big],$$

which implies that the true predictor $f_\rho(x, x')$ is obtained by minimizing the inner conditional expectation for almost $x, x' \in \mathcal{X}$. Then, there holds

$$\begin{aligned} f_\rho(x, x') &= \arg\min_{t\in\mathbb{R}} \mathbb{E}\big[\ell(t, Y, Y')\big|X = x, X' = x'\big] \\ &= \arg\min_{t\in\mathbb{R}} \int_{\mathcal{Y}\times\mathcal{Y}} \ell(t, y, y')\, d\rho(y|x)\, d\rho(y'|x'). \end{aligned}$$

According to Assumption 5, we further know that

$$\begin{aligned} f_\rho(x, x') &= \arg\min_{t\in\mathbb{R}} \int_{\mathcal{Y}\times\mathcal{Y}} \ell(-t, y', y)\, d\rho(y'|x')\, d\rho(y|x) \\ &= -\arg\min_{t\in\mathbb{R}} \int_{\mathcal{Y}\times\mathcal{Y}} \ell(t, y', y)\, d\rho(y'|x')\, d\rho(y|x) \\ &= -\arg\min_{t\in\mathbb{R}} \mathbb{E}\big[\ell(t, Y', Y)\big|X' = x', X = x\big] \\ &= -f_\rho(x', x). \end{aligned}$$

The second part of the proposition can be directly derived from $f_\rho(x, x') = -f_\rho(x', x)$. The proof is completed. □

The proof of Theorem 2 can be directly derived from Theorem 1 by verifying the corresponding assumptions.

PROOF OF THEOREM 2. From [Bartlett et al. 2019] we know the space $\mathcal{H}$ of deep ReLU networks is a VC-class. Then, Lemma 1 implies that $\mathcal{H}$ satisfies Assumption 3 with $V_1 = V_2 = 2(V-1)$ and $s_1 = s_2 = C(V/2+1)(16e)^{V/2+1}\eta^V$. Further, Assumption 5 and the anti-symmetric structure of $\mathcal{H}$ and f_ρ imply (2.4). Then, we can get the desired results by applying Theorem 1 directly. □

To prove Theorem 3, we first introduce the following lemma which shows the expressive ability of deep ReLU networks for approximating functions in the Sobolev spaces.

LEMMA 10 ([Yarotsky 2017]). *For any $p, r \in \mathbb{N}$, $\epsilon \in (0, 1/2)$ and any function $f \in W^{r,\infty}([0,1]^p)$ with Sobolev norm not larger than 1, there exists a deep ReLU network h with depth at most $C_{p,r}\log(1/\epsilon)$ and the number of nonzero weights and computational units at*

most $C_{p,r}\epsilon^{-\frac{p}{r}}\log(1/\epsilon)$ *such that*

$$\|h - f\|_{L^\infty([0,1]^p)} \leq \epsilon.$$

Now, we give the proof of Theorem 3 on the estimate of the approximation error.

PROOF OF THEOREM 3. According to Lemma 10, we know there exists a deep ReLU network h with depth at most $C_{p,r}\log(1/\epsilon)$ and the number of nonzero weights and computational units at most $C_{p,r}\epsilon^{-\frac{p}{r}}\log(1/\epsilon)$ such that $\big\|h - \tilde{f}_\rho\big\|_{L^\infty([0,1]^p)} \leq \frac{\epsilon}{2}$. We construct f as

$$f(x, x') = \pi_\eta\big(h(x)\big) - \pi_\eta\big(h(x')\big) \text{ for } x, x' \in \mathcal{X}.$$

Here, we suppose $\eta = 2$ since Assumption 6 implies Assumption 1 with $\eta = 2$. To maintain consistency with (2.7), we can regard $h(x)$ and $h(x')$ as functions defined on $[0,1]^p \times [0,1]^p$, which are denoted by $(x, x') \mapsto h(x)$ and $(x, x') \mapsto h(x')$, respectively. Therefore, we know f is a deep ReLU network of form (2.7) with depth at most $C_{p,r}\log(1/\epsilon)$ and the number of nonzero weights and computational units at most $C_{p,r}\epsilon^{-\frac{p}{r}}\log(1/\epsilon)$. The approximation accuracy can be bounded as follows

$$\begin{aligned}
\big\|f - f_\rho\big\|_{L^\infty([0,1]^{2p})} &= \big\|\pi_\eta\big(h(x)\big) - \pi_\eta\big(h(x')\big) - \tilde{f}_\rho(x) + \tilde{f}_\rho(x')\big\|_{L^\infty([0,1]^{2p})} \\
&\leq 2\big\|\pi_\eta\big(h(\cdot)\big) - \tilde{f}_\rho(\cdot)\big\|_{L^\infty([0,1]^p)} \\
&\leq 2\big\|h - \tilde{f}_\rho\big\|_{L^\infty([0,1]^p)} \leq \epsilon,
\end{aligned}$$

where in the second inequality we have used the fact $\pi_\eta\big(h(\cdot)\big)$ is uniformly bounded by 1. This completes the proof of the first inequality.

From [Ying and Zhou 2016] we know for any $T \in L^2_{\rho^2_{\mathbf{x}}}(\mathcal{X} \times \mathcal{X})$, the excess risk $\mathcal{E}(T) - \mathcal{E}(f_\rho) = \big\|T - f_\rho\big\|^2_{L^2_{\rho^2_{\mathbf{x}}}}$. Then, the excess risk of f can be controlled as follows

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \big\|f - f_\rho\big\|^2_{L^\infty([0,1]^{2p})} \leq \epsilon^2.$$

According to the complexity of f, by setting parameters $W = U = \lceil \exp L \rceil$, we have $\mathcal{E}(f) - \mathcal{E}(f_\rho) \leq C_{p,r}\big(\frac{L}{\exp(L)}\big)^{\frac{2r}{p}}$. Then, the approximation error

$$\mathcal{D}(\mathcal{H}) = \inf_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}(f_\rho) \leq \mathcal{E}(f) - \mathcal{E}(f_\rho) \leq C_{p,r}\Big(\frac{L}{\exp(L)}\Big)^{\frac{2r}{p}},$$

which completes the proof. □

LEMMA 11. *Suppose Assumption 7 holds. Then, Assumption 1 holds with $\eta = 2B$, Assumption 2 holds with $K = 8B$, and the shifted hypothesis space $\{\ell\big(f(x,x'),y,y'\big) - \ell\big(f_\rho(x,x'),y,y'\big) : f \in \mathcal{H}\}$ has a variance-expectation bound with parameter pair $(1, 64B^2)$.*

PROOF. Notice $\tilde{f}_\rho(x) = \mathbb{E}[Y|X = x]$ for almost $x \in \mathcal{X}$. Then, $\|\tilde{f}_\rho\|_{L^\infty(\mathcal{X})} \leq \|Y\|_{L^\infty(\mathcal{Y})} \leq B$. Since $f_\rho(x,x') = \tilde{f}_\rho(x) - \tilde{f}_\rho(x')$, we know $\|f_\rho\|_{L^\infty(\mathcal{X}\times\mathcal{X})} \leq 2\|\hat{f}_\rho\|_{L^\infty(\mathcal{X})} \leq 2B$. It follows that Assumption 1 holds with $\eta = 2B$.

For any $t_1, t_2 \in [-2B, 2B]$, and almost $y, y' \in \mathcal{Y}$,

$$\begin{aligned}
\big|\ell\big(t_1,y,y'\big) - \ell\big(t_2,y,y'\big)\big| &\leq \Big|\big(t_1 - y + y'\big)^2 - \big(t_2 - y + y'\big)^2\Big| \\
&\leq \big|t_1 - t_2\big|\big|t_1 + t_2 - 2y + 2y'\big| \\
&\leq 8B\big|t_1 - t_2\big|.
\end{aligned}$$

Then, Assumption 2 holds with $K = 8B$.

For any $f \in \mathcal{H}$, we define $q_f = \big(f(x,x') - y + y'\big)^2 - \big(f_\rho(x,x') - y + y'\big)^2$. Then

$$\begin{aligned}
\mathbb{E}\big[q_f^2\big] &= \mathbb{E}\Big[\Big(\big(f(X,X') - Y + Y'\big)^2 - \big(f_\rho(X,X') - Y + Y'\big)^2\Big)^2\Big] \\
&= \mathbb{E}\big[\big(f(X,X') - f_\rho(X,X')\big)^2\big(f(X,X') + f_\rho(X,X') - 2Y + 2Y'\big)^2\big] \\
&\leq 64B^2\mathbb{E}\big[\big(f(X,X') - f_\rho(X,X')\big)^2\big] \\
&= 64B^2\big(\mathcal{E}(f) - \mathcal{E}(f_\rho)\big) = 64B^2\mathbb{E}[q_f],
\end{aligned}$$

where the third equality is due to the specialty of the least squares loss [Ying and Zhou 2016]. Hence, we know the shifted hypothesis space has a variance-expectation bound with parameter pair $(1, 64B^2)$. This completes the proof. □

The excess generalization error bound is obtained by combining Theorem 2 and Theorem 3 together.

PROOF OF THEOREM 4. According to Theorem 7 in [Bartlett et al. 2019], we know the $\mathcal{H}$ is a VC-class and its pseudo-dimension $V = Pdim(\mathcal{H}) \leq CLW\log(U)$. Lemma 11 implies all the conditions in Theorem 2 are satisfied. Then, combining Theorem 2 and Theorem 3 together, and setting $W = U = \lceil \exp(L) \rceil$, we know with probability at least $1-\delta$, there holds

$$\mathcal{E}(\hat{f}_z) - \mathcal{E}(f_\rho) \leq C_B \frac{L^2 \exp(L)\log(n)}{n} \log^2(4/\delta) + C_{p,r}\Big(\frac{L}{\exp(L)}\Big)^{\frac{2r}{p}}.$$

This proves the first part of the theorem.

By setting $L = \lceil \frac{p}{2r+p}\log(n) \rceil$, we get the desired rate immediately. The proof of the theorem is complete. □

CHAPTER 3

Generalization Analysis of Metric and Similarity Learning

As a typical task of pairwise learning, metric and similarity learning has attracted a large amount of interest. However, there are few works on the generalization performance of such learning task. In this chapter, we focus on the generalization analysis of the metric and similarity learning with the hinge loss. The main results in this chapter are based on [Zhou et al. 2024a]. The rest of the chapter is organized as follows. We present the generalization bounds with deep ReLU networks for metric and similarity learning in Section 3.1. Section 3.2 investigates some regular properties of the true metric. We give detailed proofs of this chapter in Section 3.3.

3.1 Generalization Analysis with Deep ReLU Networks

We begin by reformulating the metric and similarity learning problems. Recall that metric and similarity learning aims to learn a metric d from an observed sample S that estimates the distance or the similarity between a pair of observers (x, x'). The performance of d on a pair (z, z') is usually measured by $\ell\big(\tau(Y, Y')d(X, X')\big)$, where $\tau(y, y')$ is the reducing function defined by $\tau(y, y') = 1$ if $y = y'$ and $\tau(y, y') = -1$ else. In this chapter, we take $\mathcal{Y} = \{y_1, ..., y_m\} \subset \mathbb{R}$ be the set of labels.

Given by $\ell\left(\tau(y, y')d(x, x')\right)$, the generalization error (true risk) associated with a metric d is defined as

$$\mathcal{E}(d) = \mathbb{E}_{Z,Z'}\left[\ell\big(\tau(Y, Y')d(X, X')\big)\right] = \int_{\mathcal{Z}\times\mathcal{Z}} \ell\big(\tau(y, y')d(x, x')\big)d\rho(z)d\rho(z').$$

The corresponding empirical error based on the sample S is defined as

$$\mathcal{E}_z(d) = \frac{1}{n(n-1)} \sum_{i \neq j=1}^{n} \ell\big(\tau(Y_i, Y_j) d(X_i, X_j)\big).$$

Let $\hat{d}_z = \arg\min_{d \in \mathcal{H}} \mathcal{E}_z(d)$ be the minimizer of the empirical error over a hypothesis space $\mathcal{H}$, and $d_\rho = \arg\min \mathcal{E}(d)$ be the true metric (target function) that minimizes the generalization error over the space of all measurable functions on $\mathcal{X} \times \mathcal{X}$. According to (1.1), we have the following error decomposition

$$\mathcal{E}(\hat{d}_z) - \mathcal{E}(d_\rho) = \{\mathcal{E}(\hat{d}_z) - \mathcal{E}(d_\mathcal{H})\} + \{\mathcal{E}(d_\mathcal{H}) - \mathcal{E}(d_\rho)\}, \tag{3.1}$$

where $d_\mathcal{H} = \arg\min_{d \in \mathcal{H}} \mathcal{E}(d)$ is the minimizer of the generalization error over the hypothesis space $\mathcal{H}$. The terms $\mathcal{E}(\hat{d}_z) - \mathcal{E}(d_\mathcal{H})$ and $\mathcal{E}(d_\mathcal{H}) - \mathcal{E}(d_\rho)$ in (3.1) are called the estimation error and the approximation error, respectively.

Let $Z = (X, Y)$ and $Z' = (X', Y')$ be random variables independently following ρ, and $\rho(\cdot|x)$ be the conditional distribution of Y given $X = x$. We denote the conditional probability of $Y = Y'$ given $X = x, X' = x'$ as

$$\eta(x, x') = Prob\{Y = Y' | X = x, X' = x'\} \tag{3.2}$$

$$= \int_{\mathcal{Y} \times \mathcal{Y}} \mathbb{I}\{y = y'\} d\rho(y|x) d\rho(y'|x'), \tag{3.3}$$

which is the probability that two observers x and x' are affiliated to the same class. One can see that η only depends on the conditional distributions of Y conditioned on the observers. For any $x, x', x'' \in \mathcal{X}$, if $\rho(\cdot|x') = \rho(\cdot|x'')$, then there holds $\eta(x, x') = \eta(x, x'')$. In addition, it can be readily observed that $\eta(x, x')$ is symmetric, i.e., $\eta(x, x') = \eta(x', x)$ for any $x, x' \in \mathcal{X}$.

Denote by $P_x := [Prob\{Y = y_1 | X = x\}, \ldots, Prob\{Y = y_m | X = x\}] \in \mathbb{R}^m$ the probability distribution vector of Y conditioned on $X = x$. Then the conditional probability (3.2) can be

represented as the standard inner product in the Euclidean space:

$$\begin{aligned}\eta(x,x') &= \sum_{i=1}^{m} Prob\{Y = Y' = y_i | X = x, X' = x'\} \\ &= \sum_{i=1}^{m} Prob\{Y = y_i | X = x\} Prob\{Y' = y_i | X' = x'\} = \langle P_x, P_{x'} \rangle.\end{aligned}$$

In this chapter, we mainly focus on the hinge loss

$$\ell(\tau(y,y')d(x,x')) = \big(1 + \tau(y,y')d(x,x')\big)_+,$$

where $(\cdot)_+ = \max\{0, \cdot\}$.

Understanding the true metric d_ρ is crucial to estimating the approximation error. The following theorem presents an explicit form of the true metric d_ρ for the hinge loss.

THEOREM 5. *The true metric with the hinge loss can be represented as*

$$d_\rho(x,x') = sgn(1 - 2\eta(x,x')) = sgn(1 - 2\langle P_x, P_{x'} \rangle)$$

for almost every pair $x, x' \in \mathcal{X}$.

PROOF. By the tower property of the conditional expectation, the generalization error with a metric d can be written as

$$\mathcal{E}(d) = \mathbb{E}_{X,X'} \left[\mathbb{E}_{Y|X,Y'|X'} \left[\ell(\tau(Y,Y')d(X,X')) \right] \right].$$

The above observation implies that for almost every pair $x, x' \in \mathcal{X}$, the true metric $d_\rho(x,x')$ is obtained by minimizing the inner conditional expectation which can be rewritten as

$$\begin{aligned}&d_\rho(x,x') \\ &= \arg\min_{t \in \mathbb{R}} \mathbb{E}_{Y|X=x,Y'|X'=x'} \left[\ell(\tau(Y,Y')t) \right] \\ &= \arg\min_{t \in \mathbb{R}} Prob\{Y = Y' | X = x, X' = x'\} \ell(t) + Prob\{Y \neq Y' | X = x, X' = x'\} \ell(-t) \\ &= \arg\min_{t \in \mathbb{R}} \eta(x,x')\ell(t) + (1 - \eta(x,x'))\ell(-t), \end{aligned} \tag{3.4}$$

where the second equality is due to the definition of $\tau(y,y')$.

According to (3.4), the true metric with the hinge loss can be expressed as

$$
\begin{aligned}
d_\rho(x, x') &= \arg\min_{t\in\mathbb{R}} \left\{\eta(x, x')\,(1+t)_+ + (1-\eta(x, x'))\,(1-t)_+\right\} \\
&= \arg\min_{t\in\mathbb{R}} \begin{cases} (1-\eta(x, x'))(1-t), & \text{if } t < -1, \\ (2\eta(x, x') - 1)t + 1, & \text{if } t \in [-1, 1], \\ \eta(x, x')(1+t), & \text{if } t > 1. \end{cases}
\end{aligned}
$$

For the case $\eta(x, x') > 1/2$, note that the objective function is piecewise linear and tends to $+\infty$ as $t \to \pm\infty$, we can derive that $d_\rho(x, x') = -1$ and the corresponding minimum of the objective function is $2(1 - \eta(x, x'))$. For the case $\eta(x, x') < 1/2$, we can get that $d_\rho(x, x') = 1$ and the corresponding minimum is $2\eta(x, x')$. For the case $\eta(x, x') = 1/2$, we have $d_\rho(x, x') \in [-1, 1]$ and the corresponding minimum be 0. Then we can conclude that $d_\rho(x, x') = sgn(1 - 2\eta(x, x'))$, which completes the proof. □

REMARK 4. Unlike using the Mahalanobis distance to measure the similarity between sample pairs, deep metric learning [Huai et al. 2019; Kaya and Bilge 2019; Roth et al. 2020] aims to learn a nonlinear embedding function $\phi : \mathcal{X} \to \Phi \subset \mathbb{R}^s$ with $s \in \mathbb{N}^+$, such that similar data points x, x' are close in the embedding space Φ under a predefined distance function $d(\phi(x), \phi(x'))$ and far from each other if they are dissimilar. Note Theorem 5 shows that the true predictor $d_\rho(x, x') = sgn(1 - 2\langle P_x, P_{x'}\rangle)$ for the hinge loss. Then the true embedding function is $\phi(x) = P_x$ under this setting, and the distance function can be further defined as $d(\phi(x), \phi(x')) = sgn(1 - 2\langle P(x), P(x')\rangle)$. This indicates that learning a nonlinear embedding function in deep metric learning with the hinge loss is in fact learning the conditional probability P_x.

According to the specific structure of the true metric d_ρ given in Theorem 5, we can construct a structured deep network with ReLU activation as an approximation of d_ρ and further design the corresponding hypothesis space. Before that, we first introduce some assumptions and useful lemmas.

To define the Sobolev smoothness of the conditional probabilities p_i for $i = 1, \ldots, m$, we take $\mathcal{X} = [0, 1]^p$ in the remainder of this section. We define the Sobolev space $W^{r,\infty}([0, 1]^p)$

as the space of functions on $[0,1]^p$ along with their partial derivatives up to order r lying in $L^\infty([0,1]^p)$. The norm in $W^{r,\infty}([0,1]^p)$ is defined as

$$\|f\|_{W^{r,\infty}([0,1]^p)} := \max_{\alpha \in \mathbb{Z}_+^p : \|\alpha\|_1 \leq r} \|D^\alpha f\|_{L^\infty([0,1]^p)},$$

where $\|\alpha\|_1 = \sum_{i=1}^p |\alpha_i|$ denotes the l^1 norm of $\alpha = (\alpha_1, \ldots, \alpha_p) \in \mathbb{Z}_+^p$, and $D^\alpha f = \frac{\partial^{\|\alpha\|_1} f}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d}$ denotes the partial derivatives of f with order α.

For notational simplicity, denote by $p_i(x) = Prob\{Y = y_i | X = x\}$ the i-th component of P_x. The following assumption means that all r-th partial derivatives of p_i exist and their L^∞ norms are not greater than 1.

ASSUMPTION 8. *The conditional probability $p_i \in W^{r,\infty}([0,1]^p)$ has the Sobolev norm not greater than 1 for each $i = 1, \ldots, m$ (This upper bound for the Sobolev norm can be extended to any finite constant, for simplicity we set it to 1).*

In binary classification problems, we often introduce a noise condition that says the ambiguous points $Prob\{Y = 1 | X = x\} \approx 1/2$ occur with a small probability. When this condition is satisfied, it suggests that the classification problem is well-posed and learning algorithms have the potential to achieve faster convergence rates. Similarly, we can extend this notion in metric and similarity learning to suggest that the probability of $\eta(X, X') \approx 1/2$ is relatively small.

ASSUMPTION 9 (Tsybakov's noise condition). *There exist constants $\theta > 0$ and $C_\theta > 0$ such that for any $t > 0$,*

$$Prob\{|\eta(X, X') - 1/2| \leq t\} \leq C_\theta t^\theta$$

For $L \in \mathbb{N}$, denote the deep network $h : \mathbb{R}^p \to \mathbb{R}$ with the ReLU activation $\sigma(x) = \max\{x, 0\}$ and depth L as

$$h(x) = \sigma(T_L \sigma(T_{L-1} \cdots \sigma(T_1 x + b_1) \cdots + b_{L-1}) + b_L), \tag{3.5}$$

where σ acts entry-wise. For $l = 1, \ldots, L$, $T_l \in \mathbb{R}^{w_l \times w_{l-1}}$ and $b_l \in \mathbb{R}^{w_l}$ are the connection matrix and the bias of the l-th layer respectively, where $w_l \in \mathbb{N}$ is the width of the l-th layer and $w_0 = p$ is the dimension of the input space. Let the number of nonzero weights and computation units of h be $\sum_{l=1}^{L} \|T_l\|_0 + \|b_l\|_0$ and $\sum_{l=1}^{L} w_l$ respectively, where $\|\cdot\|_0$ denotes the number of nonzero elements of the corresponding matrices or vectors.

The following two lemmas established in [Yarotsky 2017] show the expressive power of deep ReLU networks in the setting of approximations in Sobolev spaces and the product function.

LEMMA 12. *For any $p, r \in \mathbb{N}$, $\epsilon \in (0, 1/2)$ and any function $f \in W^{r,\infty}([0,1]^p)$ with Sobolev norm not larger than 1, there exists a deep ReLU network h with depth at most $C_{p,r} \log(1/\epsilon)$ and the number of nonzero weights and computational units at most $C_{p,r}\epsilon^{-\frac{p}{r}} \log(1/\epsilon)$ such that*

$$\|h - f\|_{L^\infty([0,1]^p)} \leq \epsilon.$$

LEMMA 13. *For any $\epsilon \in (0, 1/2)$, there exists a deep ReLU network ϕ with the depth and the number of weights and computation units at most $C \log(1/\epsilon)$ such that*

$$\|\phi(x, y) - xy\|_{L^\infty([-1,2]^2)} \leq \epsilon \text{ and } \phi(x, y) = 0 \text{ if } xy = 0.$$

We will use a deep ReLU network h_i to approximate the conditional probability $p_i \in [0, 1]$ later. According to Lemma 12, if Assumption 8 holds, we have $h_i \in [-1, 2]$.

Now, we are ready to design a structured deep network with ReLU activation to approximate the true metric d_ρ as follows

$$d(x, x') := F_a\Big(1 - 2\sum_{i=1}^{m} \phi(h_i(x), h_i(x'))\Big), \tag{3.6}$$

where $F_a(t) := \frac{1}{a}\big(\sigma(t + a) - \sigma(t - a) - a\big)$ is a fixed two-layer ReLU network with $a > 0$ (the value of a will be selected later, see Theorems 6 and 8 below for more details), ϕ is a deep ReLU network for approximating the product function xy introduced in Lemma 13, and h_i is a sub-network has the form (3.5). Here, we assume that each h_i has the same depth.

The main idea of constructing (3.6) is to design a series of sub-networks to approximate the true metric d_ρ part by part according to its form $d_\rho(x, x') = sgn(1 - 2\sum_{i=1}^{m} p_i(x)p_i(x'))$ given by Theorem 5. Specifically, we first construct m sub-networks h_i to approximate the conditional probabilities p_i for $i = 1, \ldots, m$, respectively. Next, a deep ReLU network $\phi(h_i(x), h_i(x'))$ is introduced to approximate product function $h_i(x)h_i(x')$. Lemma 13 implies that $\phi(h_i(x), h_i(x'))$ is a reliable approximation of $h_i(x)h_i(x')$. Finally, the output layer $F_a(\cdot)$ is employed to approximate the sign function as used in [Zhou and Huo 2024]. Indeed, $F_a(\cdot)$ coincides with $sgn(\cdot)$ on the interval $(-\infty, -a) \cup (a, +\infty)$, and exhibits linearity on the interval $[-a, a]$ to serve as an approximation for the discontinuity of $sgn(\cdot)$ near the origin. The specific structure of (3.6) is described in Figure 3.1.

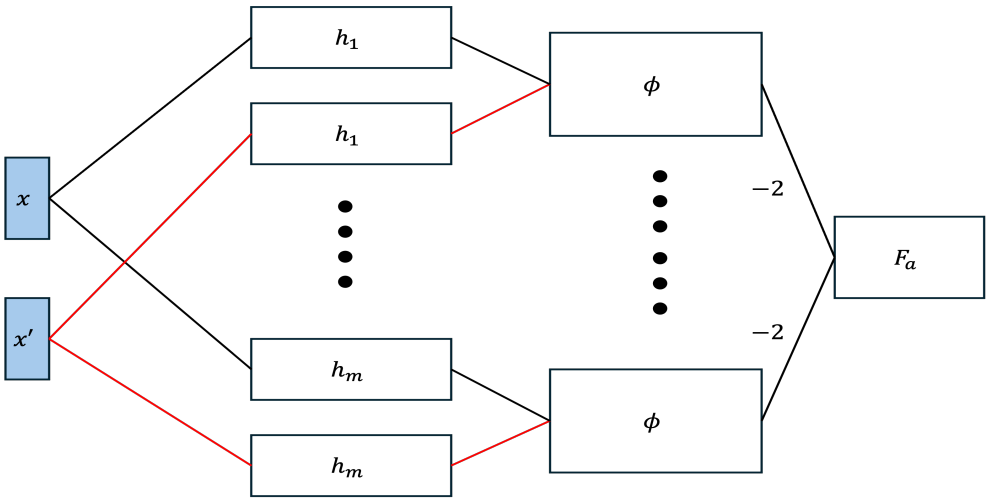

FIGURE 3.1. Structure of the designed deep network with ReLU activation (3.6) with input $x, x' \in \mathcal{X}$.

We say a network d has complexity (L, W, U) if its depth, the number of possibly nonzero weights and computation units are L, W and U. The hypothesis space consisting of the structured deep network of the form (3.6) is defined as

$$\mathcal{H} = \mathcal{H}(L, W, U) = \big\{ d(x, x') \text{ of the form (3.6) : the complexity of } d \text{ does not exceed } (L, W, U) \big\}. \tag{3.7}$$

Here, the complexity of d can be computed by summing up the corresponding parameters of sub-networks. Specifically, the depth of d is the summation of the depth of h_1, ϕ and

F_a (recall that each h_i has the same depth). The number of possibly nonzero weights and computation units of d are $2\sum_{i=1}^{m} W_{h_i} + mW_\phi + W_{F_a} + c$ and $2\sum_{i=1}^{m} U_{h_i} + mU_\phi + U_{F_a} + c$, respectively, where W_γ and U_γ denote the corresponding parameters of the sub-networks $\gamma \in \{h_1, \ldots, h_m, \phi, F_a\}$, and c is an absolute constant. The capacity (size) of the hypothesis space $\mathcal{H}$ can be measured by (L, W, U). As these parameters increase, the capacity of the hypothesis space gets larger.

The following theorem establishes approximation error bounds of the structured deep networks of the form (3.6).

THEOREM 6 (Approximation error). *Suppose Assumptions 8 and 9 hold. For any $\epsilon \in (0, 1/2)$, if the hypothesis space $\mathcal{H}$ defined in* (3.7) *has the depth $C_{p,r,m,\theta} \log(1/\epsilon)$ and the number of possibly nonzero weights and computation units $C_{p,r,m,\theta}\epsilon^{-\frac{p}{r(\theta+1)}} \log(1/\epsilon)$, then there exists a deep ReLU network $d_{\mathcal{H}} \in \mathcal{H}$ of the form* (3.6) *with $a = C_\theta \epsilon^{\frac{1}{\theta+1}}$ such that*

$$\mathcal{E}(d_{\mathcal{H}}) - \mathcal{E}(d_\rho) \leq \epsilon.$$

Theorem 6 implies that the approximation ability of $d_{\mathcal{H}}$ becomes better if the capacity of the hypothesis space $\mathcal{H}$ is larger. However, as the hypothesis space grows in size, the model may become overly flexible, which leads to increasing the estimation error. This means that there is a trade-off between the estimation and the approximation error. We will choose a proper hypothesis space to obtain the optimal excess generalization error rate (see Theorem 8).

To derive upper bounds of the estimation error, we utilize pseudo-dimension to measure the capacity of the hypothesis space. The pseudo-dimension is defined by the VC-dimension of the sub-graph set of the hypothesis space. A comprehensive definition of VC-dimension can be found in [Györfi et al. 2006; Wainwright 2019].

DEFINITION 7. Let $\mathcal{F}$ be a class of functions from $[0, 1]^p$ to $\mathbb{R}$, and $\mathcal{F}^+ := \{(x, t) \in [0, 1]^p \times \mathbb{R} : f(x) > t, f \in \mathcal{F}\}$ be the corresponding sub-graph set. The pseudo-dimension $Pdim(\mathcal{F})$ of $\mathcal{F}$ is defined as

$$Pdim(\mathcal{F}) := VC(\mathcal{F}^+),$$

where $VC(\mathcal{F}^+)$ is the VC-dimension of $\mathcal{F}^+$. Furthermore, if $Pdim(\mathcal{F}) < \infty$, then we call $\mathcal{F}$ a VC-class.

If we solely apply the uniform boundedness (first-order condition) of the hypothesis space, the estimation error bound is of the order $O(1/\sqrt{n})$ [Cao et al. 2016; Clémençon et al. 2008], which is often quite loose. To derive a tighter upper bound of the estimation error, the variance condition (second-order condition) should be taken into consideration.

DEFINITION 8. Let $\beta \in (0, 1]$ and $M > 0$, and $\mathcal{F} \subset L^2(\mathcal{X} \times \mathcal{X}, \rho \times \rho)$ is a function class with nonnegative first order moment, i.e.for any $f \in \mathcal{F}$, $\mathbb{E}[f] \geq 0$. We say $\mathcal{F}$ has a variance-expectation bound with parameter pair (β, M), if for any $f \in \mathcal{F}$,

$$\mathbb{E}[f^2] \leq M(\mathbb{E}[f])^\beta.$$

DEFINITION 9. Let $\Omega \subset \mathbb{R}^p$. We say a function $g : \Omega \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $\alpha > 0$ if for any $x, y \in \Omega$,

$$|g(x) - g(y)| \leq \alpha \|x - y\|_2,$$

where $\|\cdot\|_2$ is the Euclidean norm.

The following lemma proved in Chapter 2 shows that if the loss satisfies some mild conditions and the shifted hypothesis space has a variance-expectation bound, then a tight upper bound for the estimation error can be established.

LEMMA 14. *Let $V = Pdim(\mathcal{H})$ be the pseudo-dimension of the hypothesis space $\mathcal{H}$, and $\ell(\tau(y, y')d(x, x'))$ be the loss function for $(z, z') \in \mathcal{Z} \times \mathcal{Z}$, where d is a predictor from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. Suppose the following conditions hold for some $\alpha > 0$, $s > 0$, $\beta \in (0, 1]$ and $M > 0$.*

- *for any $y, y' \in \mathcal{Y}$, the loss function $\ell(\tau(y, y')\cdot)$ is Lipschitz continuous with Lipschitz constant α,*
- *for any $d \in \mathcal{H} \cup \{d_\rho\}$ and for almost every sample pair $(z, z') \in \mathcal{Z} \times \mathcal{Z}$, there holds $\ell(\tau(y, y')d(x, x')) = \ell(\tau(y', y)d(x', x))$ and $\|d\|_{L^\infty(\mathcal{X}\times\mathcal{X})} \leq s \in \mathbb{R}^+$,*

- *the shifted hypothesis space* $\{\ell(\tau(y,y')d(x,x')) - \ell(\tau(y,y')d_\rho(x,x')) : d \in \mathcal{H}\}$ *has a variance-expectation bound with parameter pair* (β, M)*,*

then for any $\delta \in (0, 1/2)$*, with probability at least* $1-\delta$*, there holds*

$$\mathcal{E}(\hat{d}_z) - \mathcal{E}(d_\rho) \leq C_{s,\alpha,M,\beta}\Big(\frac{V\log(n)\log^2(4/\delta)}{n}\Big)^{\frac{1}{2-\beta}} + 2(1+\beta)\Big(\mathcal{E}(d_{\mathcal{H}}) - \mathcal{E}(d_\rho)\Big).$$

The first two conditions in this lemma are easy to verify. Indeed, the Lipschitz continuity of the loss ℓ with Lipschitz constant $\alpha = 1$ simply follows from the Lipschitz continuity of the hinge loss. It is evident that the hypothesis functions and the true metric are uniformly bounded by $s = 1$, and the symmetry of the loss ℓ is obtained by the symmetries of the reducing function τ, the hypothesis functions d and the true metric d_ρ with respect to y, y' and x, x' respectively. All that remains is to check whether the shifted hypothesis space has a variance-expectation bound. The following proposition shows that this bound can be established under Tsybakov's noise condition.

PROPOSITION 3 (Variance-expectation bound). *Suppose Assumption 9 holds. For any* $d \in \mathcal{H}$*, the shifted hypothesis space* $\{\ell(\tau(y,y')d(x,x')) - \ell(\tau(y,y')d_\rho(x,x')) : d \in \mathcal{H}\}$ *has a variance-expectation bound with parameter pair* $\Big(\frac{\theta}{\theta+1}, 2^{\frac{3}{\theta+1}}C_\theta^{\frac{1}{\theta+1}}\Big)$.

The estimation error is estimated in the following theorem by combining Lemma 14 and Proposition 3 together.

THEOREM 7 (Estimation error). *Let* $V = Pdim(\mathcal{H})$ *be the pseudo-dimension of the hypothesis space* $\mathcal{H}$ *defined in* (3.7)*. For any* $\delta \in (0, 1/2)$*, with probability at least* $1-\delta$*, there holds*

$$\mathcal{E}(\hat{d}_z) - \mathcal{E}(d_\rho) \leq C_\theta\Big(\frac{V\log(n)\log^2(4/\delta)}{n}\Big)^{\frac{\theta+1}{\theta+2}} + \frac{4\theta+2}{\theta+1}\Big(\mathcal{E}(d_{\mathcal{H}}) - \mathcal{E}(d_\rho)\Big).$$

Now, we can obtain an excess generalization error bound by combining the approximation error bound (Theorem 6) and the estimation error bound (Theorem 7). Then, the optimal excess generalization error rate is established by carefully trading off the estimation error and the approximation error.

THEOREM 8 (Excess generalization error). *Suppose Assumptions 8 and 9 hold and let* $L \in \mathbb{N}$. *Consider the hypothesis space* $\mathcal{H} = \mathcal{H}(L, W, U)$ *defined in* (3.7) *with* $W = U = \lceil C_{p,r,m,\theta} \exp\{L\} \rceil$. *For any* $\delta \in (0, 1/2)$, *with probability at least* $1 - \delta$, *there holds*

$$\mathcal{E}(\hat{d}_z) - \mathcal{E}(d_\rho) \leq C_{p,r,m,\theta} \log^2(4/\delta) \left\{ \left(\frac{L^2 \exp\{L\} \log n}{n} \right)^{\frac{\theta+1}{\theta+2}} + \left(\frac{L}{\exp\{L\}} \right)^{\frac{(\theta+1)}{p}} \right\}.$$

By setting $L = \left\lceil \frac{p}{p+(\theta+2)r} \log \left(\frac{n}{\log n} \right) \right\rceil$, *there holds*

$$\mathcal{E}(\hat{d}_z) - \mathcal{E}(d_\rho) \leq C_{p,r,m,\theta} \log^2(4/\delta) \log^4(n) n^{-\frac{(\theta+1)r}{p+(\theta+2)r}}.$$

Theorem 8 shows that the excess generalization error bound is of order (up to a logarithmic term) $O(n^{-\frac{(\theta+1)r}{p+(\theta+2)r}})$. This rate is closely related to the dimension of the input space, the parameter θ in the noise condition, and the smoothness r of the conditional probabilities. When the distribution ρ has very low noise and the corresponding conditional probabilities are rather smooth, i.e., parameters θ and r are very large, then the learning rate can be of order $O(n^{-1+\epsilon})$ with a small $\epsilon > 0$.

REMARK 5. It is worth emphasizing that previous works [Cao et al. 2016; Guo and Ying 2014; Huai et al. 2019; Jin et al. 2009; Ye et al. 2019] on the study of generalization analysis for metric and similarity learning only derived the estimation error bounds. For instance, [Cao et al. 2016] established the upper bound for the estimation error of the order $O(\frac{1}{\sqrt{n}})$ for metric learning with the hinge loss, where they focused on learning the Mahalanobis distance. [Ye et al. 2019] showed that the convergence rate of the estimation error can achieve $O(\frac{1}{n})$ for metric learning with the smooth loss function and strongly convex objective. [Jin et al. 2009] established estimation error bounds of the order $O(\frac{1}{\sqrt{n}})$ via the algorithmic stability for metric learning.

3.2 Properties of True Metric with General Losses

In this section, we study the regularities of the problem setting and the true metric for metric and similarity learning with a general loss. Specifically, we first show that it is reasonable

to remove the bias term b in the loss $\ell(\tau(y, y')(d(x, x') - b))$ and assume the output space $\mathcal{Y}$ only has finite labels as what we do in the chapter. Further, we prove that the true metric is symmetric for almost $x, x' \in \mathcal{X}$, which provides a rationale for the use of symmetric models like Mahalanobis distance in metric learning. Finally, we show that the true metric between any two identical samples must be less than or equal to that between different samples for convex, non-negative and non-decreasing losses.

Removing the bias term. Unlike many works [Cao et al. 2016; Huai et al. 2019; Jin et al. 2009; Lei and Ying 2016; Roth et al. 2020; Ye et al. 2019], we do not introduce the bias term $b > 0$ in the loss function ℓ. We will show that this term can be set to 0 if we learn the true metric directly.

Denote by $\mathcal{E}_b(d) = \mathbb{E}_{Z,Z'}\left[\ell(\tau(Y, Y')(d(X, X') - b)\right]$ the generalization error of d with a bias term $b > 0$. Define $\tilde{d}_\rho := \arg\min_{d\in\mathcal{F}} \mathcal{E}_b(d)$ the true metric under this loss with a bias term. Analogous to (3.4), there holds

$$\begin{aligned}\tilde{d}_\rho(x, x') &= \arg\min_{t\in\mathbb{R}} \eta(x, x')\ell(t - b) + (1 - \eta(x, x'))\ell(b - t) \\ &= b + \arg\min_{s\in\mathbb{R}} \eta(x, x')\ell(s) + (1 - \eta(x, x'))\ell(-s) \\ &= b + d_\rho(x, x')\end{aligned}$$

for almost every pair $x, x' \in \mathcal{X}$. It is then indicated that b is redundant and we hence set $b = 0$.

Output space is finite. In almost all the theoretical and empirical literature of metric and similarity learning [Bar-Hillel et al. 2005; Cao et al. 2016; Davis et al. 2007; Guo and Ying 2014; Jin et al. 2009; Kar and Jain 2011; Lei and Ying 2016; Ye et al. 2019], the label space $\mathcal{Y}$ is often assumed to be finite or even binary, i.e., $\mathcal{Y} = \{+1, -1\}$. The finite label space means that the distribution of Y is discrete. It naturally asks what if the distribution of Y is continuous? The following proposition gives an answer to this question.

PROPOSITION 4. *If the distribution of Y is continuous, then the true metric equals to a generalized constant $c \in [-\infty, +\infty]$ almost surely.*

Proposition 4 implies that when the distribution of Y is continuous, for almost surely, $d_\rho(x, x') = c \in [-\infty, +\infty]$. The distances or similarities between almost every pair x, x' are the same. Therefore, it's reasonable to assume that the label space $\mathcal{Y}$ is finite.

The above proposition suggests that within any distribution of Y, the continuous part plays no significant role in defining the true metric since the conditional probability η solely relies on the discrete part of the distribution. In addition, this property is also determined by the reducing function τ. Note we assume that $\tau(y, y') = 1$ only when $y = y'$, where this event is impossible for a continuous distribution. Hence, labels play almost no role in the learning process under this setting.

Regularities of the true metric. In contrast to predictors in ranking problems [Agarwal and Niyogi 2009; Clémençon et al. 2008; Huang et al. 2023], the metric d is expected to reflect the distance or similarity of given objects instead of a rank or an order. Thus, the metric d is supposed to be independent of the order of the sample pair (x, x'), i.e., we expect $d(x, x') = d(x', x)$. The following proposition shows that the true metric d_ρ satisfies this symmetric property, which provides a theoretical guarantee that functions in the hypothesis space can be constructed in symmetric forms, such like the Mahalanobis distance $(x - x')^\top M(x - x')$ or the pairwise similarity function $x^\top M x'$.

PROPOSITION 5. *The true metric d_ρ is symmetric, i.e., almost surely we have*

$$d_\rho(x, x') = d_\rho(x', x).$$

In mathematical terms, the metric or distance between any two identical points is defined as zero. The distance predefined as $d(\phi(x), \phi(x')) = \|\phi(x) - \phi(x')\|_2$ in Deep Metric Learning [Huai et al. 2019; Kaya and Bilge 2019; Roth et al. 2020] and the Mahalanobis distance $d(x, x') = (x - x')^\top M(x - x')$ in traditional Distance Metric Learning [Cao et al. 2016; Ye et al. 2019] both vanish when $x = x'$. Then, for all $x, x' \in \mathcal{X}$, $d(x, x) \leq d(x, x')$. However,

this may not hold for the true metric. For example, consider the true metric with the hinge loss (see Theorem 5). Let the size of the labels be $m = 3$ and assume that $P_x = [3/5, 1/5, 1/5]$ and $P_{x'} = [1, 0, 0]$. Then $\eta(x, x) = \langle P_x, P_x \rangle = 11/25 < 3/5 = \langle P_x, P_{x'} \rangle = \eta(x, x')$, and we know that $d_\rho(x, x) = sgn(1 - 2\eta(x, x)) = 1 > -1 = sgn(1 - 2\eta(x, x')) = d_\rho(x, x')$. This indicates that x is more similar to x' than to itself. One can observe that this phenomenon is determined by the unknown distribution ρ. It is natural to investigate the sufficient conditions for the distribution ρ such that for any $x, x' \in \mathcal{X}$, there holds $d_\rho(x, x) \leq d_\rho(x, x')$.

Note the minimizer in (3.4) may not be unique, we first introduce the following definition.

DEFINITION 10. For any $x, x' \in \mathcal{X}$ and $a \in [0, 1]$, let $t^*(a) := \inf\{s \in \mathbb{R} | s \in \arg\min_{t\in\mathbb{R}} a\ell(t) + (1 - a)\ell(-t)\}$ be the infimum of all the minimizers of problem (3.4) with conditional probability a. We define $d_\rho(x, x') = t^*(\eta(x, x'))$. Then $d_\rho(x, x) \leq d_\rho(x, x')$ if $t^*(\eta(x, x)) \leq t^*(\eta(x, x'))$.

For many losses like the modified least squares $\ell(-t) = \max\{1 - t, 0\}^2$, the hinge loss $\ell(-t) = \max\{1 - t, 0\}$, the exponential loss $\ell(-t) = \exp(-t)$, and the logistic regression loss $\ell(-t) = \ln(1 + \exp(-t))$, the minimizer $d_\rho(x, x')$ is unique for $x, x' \in \mathcal{X}$ [Zhang 2004]. Definition 10 defines $d_\rho(x, x')$ as the minimum of all the minimizers, and the comparison is for the minimum values of $d_\rho(x, x)$ and $d_\rho(x, x')$. We do not make the assumption of the uniqueness of the minimizer here.

ASSUMPTION 10. *The loss function ℓ is convex, non-decreasing, and non-negative.*

This assumption is very reasonable because we often use a convex loss to implement algorithms efficiently. For any sample pair $(x, y), (x', y')$, the distance or the similarity $d(x, x')$ is supposed to be small when $\tau(y, y') = 1$, then we expect the loss $\ell(d(x, x'))$ to increase as $d(x, x')$ increases. For the case of $\tau(y, y') = -1$, the monotonicity of ℓ can be discussed similarly.

According to (3.4), when a loss ℓ is given, the true metric $d_\rho(x, x')$ is only determined by the conditional probability $\eta(x, x')$. Hence, it becomes imperative to investigate the conditions under which the property "$d_\rho(x, x) \leq d_\rho(x, x')$" holds for $\eta(x, x')$. As shown in (3.2), when

$\eta(x, x') = Prob\{Y = Y'|X = x, X' = x'\}$ is large, the observers x, x' are more likely to be affiliated to the same class. Then we expect that the distance or the similarity between them is small, further, the true distance $d_\rho(x, x')$ is expected to be greater than or equal to $d_\rho(x, x)$ if their corresponding conditional probabilities satisfy $\eta(x, x') \leq \eta(x, x)$.

For $\eta \in [0, 1]$ and $t \in \mathbb{R}$, we define

$$Q(\eta, t) = \eta\ell(t) + (1 - \eta)\ell(-t). \tag{3.8}$$

It's obvious that $t^*(\eta) \in \arg\min_{t\in\mathbb{R}} Q(\eta, t)$. The following lemma shows that $t^*(\eta)$ is non-increasing on $[0, 1]$, which is the key step to prove $d_\rho(x, x) \leq d_\rho(x, x')$ for the general case.

LEMMA 15. *Suppose Assumption 10 holds. If* $0 \leq \eta_1 \leq \eta_2 \leq 1$*, then* $t^*(\eta_2) \leq t^*(\eta_1)$.

Lemma 15 also has some implications for binary classification. For a binary classification problem, we aim to learn a classifier $f : \mathcal{X} \to \{-1, 1\}$ from a given sample S. The performance of a classifier f is usually measured by its generalization error $\mathbb{E}_Z[\ell(-Yf(X))]$ with a loss ℓ that is convex, non-decreasing, and nonnegative (examples of losses can be found in the paragraph below Definition 10). Similar to the arguments of (3.4) in the proof of Theorem 5, one can derive that, for almost every $x \in \mathcal{X}$, the true predictor (target function) f_ρ with loss ℓ can be written as $f_\rho(x) = \arg\min_{t\in\mathbb{R}} p(x)\ell(-t) + (1 - p(x))\ell(t)$, where $p(x) = Prob\{Y = 1|X = x\}$ is the conditional probability. Therefore, by setting $Q(p, t) = p\ell(t) + (1 - p)\ell(-t)$, we get $f_\rho(x) = -\arg\min_{t\in\mathbb{R}} Q(p(x), t) = -t^*(p(x))$ for almost every $x \in \mathcal{X}$. Note Lemma 15 shows that $t^*(\cdot)$ is non-increasing on $[0, 1]$. Then we know that the value of the true predictor $f_\rho(x) = -t^*(p(x))$ increases as the conditional probability $p(x)$ increases, which implies that the tendency of the sample x being classified to class $\{1\}$ by the true predictor is also increasing.

The following property can be directly obtained by using Lemma 15.

PROPOSITION 6. *Suppose Assumption 10 holds. If* $\eta(x, x') \leq \min\{\eta(x, x), \eta(x', x')\}$*, then*

$$\max\{d_\rho(x, x), d_\rho(x', x')\} \leq d_\rho(x, x').$$

Define $\mathcal{P} = \{P_x \in \mathbb{R}^m | x \in \mathcal{X}\}$ as the conditional distribution family. One can easily observe that $\mathcal{P}$ is a subset of the probability simplex

$$\Delta_m = \left\{ (p_1, \ldots, p_m) \in \mathbb{R}^m \mid \sum_{i=1}^{m} p_i = 1, p_i \geq 0 \right\}.$$

The condition in Theorem 6 can be written as $\langle P_x, P_{x'} \rangle \leq \min\{\|P_x\|_2^2, \|P_{x'}\|_2^2\}$ for almost $x, x' \in \mathcal{X}$. In the aspects of geometry, it requires that the inner product of any two probability vectors in $\mathcal{P}$ is less than or equal to both the square of their Euclidean norms. One example satisfying this condition is that $\mathcal{P}$ is a subset of the intersection of the probability simplex Δ_m and the ball centered at the origin with radius $r \in [0, 1]$. In this case, the equality of the condition holds only when $P_x = P_{x'}$.

Furthermore, $\mathcal{P}$ has an empty interior in the sub-topology of probability simplex Δ_m. Otherwise, we can find a small enough vector ϵ such that $P_x + \epsilon \in \mathcal{P}$ and $\langle P_x, P_x + \epsilon \rangle > \min\{\|P_x\|^2, \|P_x + \epsilon\|^2\}$, where P_x is a relative interior point of $\mathcal{P}$.

3.3 Proofs on Metric and Similarity Learning

3.3.1 Proofs for Section 3.1

PROOF OF THEOREM 6. Note $|\tau(y, y')| = 1$ and $|sgn(t)| \leq 1$ for any $y, y' \in \mathcal{Y}$ and $t \in \mathbb{R}$. For any metric $d : [0, 1]^p \times [0, 1]^p \to \mathcal{Y}$ with $\|d\|_{L^\infty([0,1]^{2p})} \leq 1$, the conditional excess risk can be written as

$$\begin{aligned}
&\mathbb{E}_{Y|X=x,Y'|X'=x'} \left[\ell(\tau(Y, Y')d(x, x')) - \ell(\tau(Y, Y')d_\rho(x, x'))\right] \\
&= \mathbb{E}_{Y|X=x,Y|X'=x'} \left[\big(1 + \tau(Y, Y')d(x, x')\big)_+ - \big(1 + \tau(Y, Y')sgn(1 - 2\eta(x, x'))\big)_+\right] \\
&= \mathbb{E}_{Y|X=x,Y|X'=x'} \left[\big(1 + \tau(Y, Y')d(x, x')\big) - \big(1 + \tau(Y, Y')sgn(1 - 2\eta(x, x'))\big)\right] \\
&= \mathbb{E}_{Y|X=x,Y|X'=x'} \left[\tau(Y, Y')\big(d(x, x') - sgn(1 - 2\eta(x, x'))\big)\right] \\
&= (2\eta(x, x') - 1)(d(x, x') - sgn(1 - 2\eta(x, x')) \\
&= |2\eta(x, x') - 1||d(x, x') - sgn(1 - 2\eta(x, x')|,
\end{aligned}$$

where in the first equality we have used Theorem 5, the second equality follows from the fact that the terms inside $(\cdot)_+$ are nonnegative, and the last equality is obtained by discussing the sign of $2\eta - 1$ and the fact $\|d\|_{L^\infty([0,1]^{2p})} \leq 1$.

Therefore, for $\delta \in (0,1)$ and $d \in \mathcal{H}$, the excess error

$$\begin{aligned}
&\mathcal{E}(d) - \mathcal{E}(d_\rho) \\
&= \mathbb{E}_{X,X'}\left[\mathbb{E}_{Y|X,Y'|X'}\left[\ell(\tau(Y,Y')d(X,X')) - \ell(\tau(Y,Y')d_\rho(X,X'))\right]\right] \\
&= \mathbb{E}_{X,X'}\left[|2\eta(X,X') - 1|\,|d(X,X') - sgn(1-2\eta(X,X'))|\right] \\
&= \int_{|1-2\eta(x,x')|>2\delta} |2\eta(x,x') - 1|\,|d(x,x') - sgn(1-2\eta(x,x'))|\,d\rho_X(x)d\rho_X(x') \\
&\quad + \int_{|1-2\eta(x,x')|\leq 2\delta} |2\eta(x,x') - 1|\,|d(x,x') - sgn(1-2\eta(x,x'))|\,d\rho_X(x)d\rho_X(x') \\
&\leq \int_{|1-2\eta(x,x')|>2\delta} |2\eta(x,x') - 1|\,|d(x,x') - sgn(1-2\eta(x,x'))|\,d\rho_X(x)d\rho_X(x') \\
&\quad + 4C_\theta\delta^{\theta+1},
\end{aligned} \tag{3.9}$$

where in the last step we have used Assumption 9 and the condition $|\eta - 1/2| \leq \delta$.

Let us analyze the first term in (3.9). Note that Lemma 12 with $f = p_i$ implies that for each p_i with approximation accuracy $\frac{\delta}{8m}$, there exists a ReLU network h_i with depth at most $C_{p,r,m}\log(1/\delta)$ and the number of possibly nonzero weights and computation units at most $C_{p,r,m}\delta^{-\frac{p}{r}}\log(1/\delta)$. Since $p_i(x) \in [0,1]$ and $|h_i(x) - p_i(x)| \leq \frac{\delta}{8m} < 1$ for any $x \in [0,1]^p$ and $i = 1,\ldots,m$, we know $h_i(x) \in [-1,2]$ for any $x \in [0,1]^p$ and $i = 1,\ldots,m$.

Let ϕ be the ReLU network in Lemma 13 with approximation accuracy $\frac{\delta}{8m}$, and $F_a(x) = \frac{1}{\delta}(\sigma(x+\delta) - \sigma(x-\delta) - \delta)$ with $a = \delta$. Consider constructing $d_{\mathcal{H}}$ using the above sub-networks, i.e.

$$d_{\mathcal{H}}(x,x') = F_a\Big(1 - 2\sum_{i=1}^m \phi(h_i(x), h_i(x'))\Big).$$

We know $d_{\mathcal{H}}$ is a deep ReLU network from $[0,1]^p \times [0,1]^p$ to $[-1,1]$ with depth at most $C_{p,r,m}\log(1/\delta)$ and the number of possibly nonzero weights and computation units at most $C_{p,r,m}\delta^{-\frac{p}{r}}\log(1/\delta)$.

We claim that the sign of the term $1-2\sum_{i=1}^{m}\phi(h_i(x),h_i(x'))$ inside F_a coincides with $1-2\eta(x,x')$ if $|1-2\eta(x,x')|>2\delta$, which can be proved by showing that $|(1-2\sum_{i=1}^{m}\phi(h_i(x),h_i(x'))-(1-2\eta(x,x'))|<\delta$. Because when $1-2\eta(x,x')>2\delta$ or $1-2\eta(x,x')<-2\delta$, we must have $1-2\sum_{i=1}^{m}\phi(h_i(x),h_i(x'))>\delta$ or $1-2\sum_{i=1}^{m}\phi(h_i(x),h_i(x'))<-\delta$. Indeed, for any $x,x'\in[0,1]^p$, there holds

$$\begin{aligned}
&\Big|\Big(1-2\sum_{i=1}^{m}\phi(h_i(x),h_i(x'))\Big)-(1-2\eta(x,x'))\Big|\\
&\leq 2\sum_{i=1}^{m}\Big|\phi(h_i(x),h_i(x'))-p_i(x)p_i(x')\Big|\\
&\leq 2\sum_{i=1}^{m}\Big|\phi(h_i(x),h_i(x'))-h_i(x)h_i(x')\Big|+\Big|h_i(x)h_i(x')-p_i(x)p_i(x')\Big|\\
&\leq 2\sum_{i=1}^{m}\Big(\frac{\delta}{8m}+|h_i(x)||h_i(x')-p_i(x')|+|p_i(x)||h_i(x)-p_i(x)|\Big)\\
&\leq 2\sum_{i=1}^{m}\Big(\frac{\delta}{8m}+\frac{\delta}{4m}+\frac{\delta}{8m}\Big)=\delta.
\end{aligned}$$

According to the definition of F_a and recall that $a=\delta$, we know $F_a(t)=1$ if $t>\delta$ and $F_a(t)=-1$ if $t<-\delta$. Therefore, we conclude that $d_{\mathcal{H}}(x,x')-sgn(1-2\eta(x,x'))=0$ if $|1-2\eta(x,x')|>2\delta$. Thus, the first term in (3.9) vanishes.

Taking $d=d_{\mathcal{H}}$ in the above approximation error bound, we last have

$$\mathcal{E}(d_{\mathcal{H}})-\mathcal{E}(d_\rho)\leq 4C_\theta\delta^{\theta+1}.$$

By setting $\epsilon=4C_\theta\delta^{\theta+1}$ and recall $a=\delta$, we get the desired results. □

PROOF OF PROPOSITION 3. As in the proof of Theorem 6, for any $d\in\mathcal{H}$, we can show

$$\begin{aligned}
\mathcal{E}(d)-\mathcal{E}(d_\rho)&=\mathbb{E}_{X,X'}\left[|2\eta(X,X')-1|\,|d(X,X')-sgn(1-2\eta(X,X'))|\right]\\
&=\int_{\mathcal{X}\times\mathcal{X}}|2\eta(x,x')-1|\,|d(x,x')-sgn(1-2\eta(x,x'))|\,d\rho_X(x)d\rho_X(x').
\end{aligned}$$

Let $q_d(z,z'):=\ell(\tau(y,y')d(x,x'))-\ell(\tau(y,y')d_\rho(x,x'))$. Note that $\tau(y,y')d(x,x')\in[-1,1]$ for any $z,z'\in\mathcal{Z}$ and $d\in\mathcal{H}\cup\{d_\rho\}$. Then there holds $\ell(\tau(y,y')d(x,x'))=1+\tau(y,y')d(x,x')$.

Since $|\tau(y,y')|^2=1$, we have

$$\begin{aligned}
&\mathbb{E}_{Z,Z'}[q_d^2(Z,Z')]\\
&=\mathbb{E}_{X,X'}[|d(X,X')-d_\rho(X,X')|^2]\\
&=\int_{\mathcal{X}\times\mathcal{X}}|d(x,x')-sgn(1-2\eta(x,x'))|^2\,d\rho_X(x)d\rho_X(x')\\
&=\int_{|2\eta(x,x')-1|>t}|d(x,x')-sgn(1-2\eta(x,x'))|^2\,d\rho_X(x)d\rho_X(x')\\
&\quad+\int_{|2\eta(x,x')-1|\le t}|d(x,x')-sgn(1-2\eta(x,x'))|^2\,d\rho_X(x)d\rho_X(x')\\
&\le\int_{|2\eta(x,x')-1|>t}|d(x,x')-sgn(1-2\eta(x,x'))|\,\frac{|2\eta(x,x')-1|}{t}d\rho_X(x)d\rho_X(x')\\
&\quad+4Prob\{|\eta(X,X')-1/2|\le t/2\}\\
&\le\frac{\mathcal{E}(d)-\mathcal{E}(d_\rho)}{t}+4C_\theta(\frac{t}{2})^\theta,
\end{aligned}$$

where $t>0$, and in the last inequality we have used Tsybakov's noise condition directly.

By choosing $t=\left(\frac{\mathcal{E}(d)-\mathcal{E}(d_\rho)}{2^{2-\theta}C_\theta}\right)^{\frac{1}{\theta+1}}$, we obtain

$$\mathbb{E}_{Z,Z'}[q_d^2(Z,Z')]\le 2^{\frac{3}{\theta+1}}C_\theta^{\frac{1}{\theta+1}}\Big(\mathcal{E}(d)-\mathcal{E}(d_\rho)\Big)^{\frac{\theta}{\theta+1}},$$

which completes the proof. □

Now, we can give the proof of the excess generalization error bound by combining Theorem 6 and Theorem 7.

PROOF OF THEOREM 8. According to Theorem 7 in [Bartlett et al. 2019], we know $V\le CLW\log U$, where W and U are the number of possibly nonzero weights and computation units, respectively. By setting $W=U=\lceil C_{p,r,m,\theta}\exp\{L\}\rceil$ and applying Theorem 6 with $\epsilon=C_{p,r,m,\theta}(\frac{L}{\exp\{L\}})^{\frac{r(\theta+1)}{p}}$ and $a=C_{p,r,m,\theta}(\frac{L}{\exp\{L\}})^{\frac{r}{p}}$, the approximation error can be estimated as

$$\mathcal{E}(d_{\mathcal{H}})-\mathcal{E}(d_\rho)\le C_{p,r,m,\theta}\Big(\frac{L}{\exp\{L\}}\Big)^{\frac{(\theta+1)r}{p}}.$$

According to Theorem 7 and applying the above two error bounds of the pseudo-dimension and the approximation error, we have, with probability at least $1-\delta$,

$$\mathcal{E}(\hat{d}_z) - \mathcal{E}(d_\rho) \leq C_{p,r,m,\theta} \log^2(4/\delta) \left\{ \left(\frac{L^2 \exp\{L\} \log n}{n} \right)^{\frac{\theta+1}{\theta+2}} + \left(\frac{L}{\exp\{L\}} \right)^{\frac{(\theta+1)}{p}} \right\}.$$

The proof is completed by setting $L = \left\lceil \frac{p}{p+(\theta+2)r} \log \left(\frac{n}{\log n} \right) \right\rceil$. □

3.3.2 Proofs for Section 3.2

We first present the proof of Proposition 4 as follows.

PROOF OF PROPOSITION 4. For almost every pair $x, x' \in \mathcal{X}$, from (3.4) we know that

$$d_\rho(x, x') = \arg\min_{t \in \mathbb{R}} \eta(x, x')\ell(t) + (1 - \eta(x, x'))\ell(-t). \tag{3.10}$$

Note that the distribution of Y is continuous. Then we know $Prob\{Y = Y'\} = 0$, and hence $\eta(x, x') = 0$ for almost every pair $x, x' \in \mathcal{X}$. Combining this fact with (3.10), we have

$$d_\rho(x, x') = \arg\min_{t \in \mathbb{R}} \ell(-t).$$

This implies that $d_\rho(x, x')$ is identical to a generalized constant $c \in [-\infty, +\infty]$ almost surely. The proof is completed. □

PROOF OF PROPOSITION 5. From (3.4), we know that the proof directly follows from the symmetry of $\eta(x, x')$. □

The proof of Lemma 15 is given as follows.

PROOF OF LEMMA 15. When $\eta_1 = \eta_2$, $\eta_1 = 0$ or $\eta_2 = 1$, the proof is trivial. Then we only consider the strict inequality in the condition such that $0 < \eta_1 < \eta_2 < 1$.

We first show that for any fixed $\eta \in (0, 1)$, if there exists a $t < t_1$ such that $Q(\eta, t) < Q(\eta, t_1)$, then we must have $t^*(\eta) < t_1$. We prove this by contradiction. Suppose $t^*(\eta) = t_1$ or $t^*(\eta) > t_1$. The first case is impossible because it then follows that $Q(\eta, t) < Q(\eta, t^*(\eta))$,

which contradicts to the definition of $t^*(\eta)$; for the second case, we have

$$\begin{aligned}&\frac{t^*(\eta)-t_1}{t^*(\eta)-t}Q(\eta,t)+\frac{t_1-t}{t^*(\eta)-t}Q(\eta,t^*(\eta))\\<&\frac{t^*(\eta)-t_1}{t^*(\eta)-t}Q(\eta,t_1)+\frac{t_1-t}{t^*(\eta)-t}Q(\eta,t_1)Q(\eta,t_1),\end{aligned}$$

where the inequality uses the assumption $Q(\eta,t_2) < Q(\eta,t_1)$ and the definition of $t^*(\eta)$. Since $\frac{t^*(\eta)-t_1}{t^*(\eta)-t_2}t_2+\frac{t_1-t_2}{t^*(\eta)-t_2}t^*(\eta)=t_1$, this strict inequality contradicts to the convexity of $Q(\eta,\cdot)$.

From the above discussion, to prove $t^*(\eta_2)\le t^*(\eta_1)$, we just have to show that there exists a $t<t^*(\eta_1)$ such that $Q(\eta_2,t)<Q(\eta_2,t^*(\eta_1))$ or $t^*(\eta_1)$ is also a minimizer of $Q(\eta_2,\cdot)$. We are interested in the behavior of ℓ at point t when t is on the left side of $t^*(\eta_1)$. Since ℓ is non-decreasing, the possibilities of the behavior can be divided into the following two cases.

Case 1: There exists a $t<t^*(\eta_1)$ such that $\ell(t)=\ell(t^*(\eta_1))$. In this case we have

$$\begin{aligned}Q(\eta_2,t)-Q(\eta_2,t^*(\eta_1))&=(1-\eta_2)(\ell(-t)-\ell(-t^*(\eta_1)))\\&=\big((1-\eta_2)/(1-\eta_1)\big)\big((1-\eta_1)(\ell(-t)-\ell(-t^*(\eta_1)))\big)\\&=\big((1-\eta_2)/(1-\eta_1)\big)\big(Q(\eta_1,t)-Q(\eta_1,t^*(\eta_1))\big)\\&<0,\end{aligned}$$

where the first and thrid equalities use the assumption $\ell(t)=\ell(t^*(\eta_1))$, the last inequality follows from the definition of $t^*(\eta_1)$. Therefore, we conclude that $t^*(\eta_2)<t^*(\eta_1)$.

Case 2: There exists a $t<t^*(\eta_1)$ such that $\ell(t)<\ell(t^*(\eta_1))$. By convexity of ℓ, we know $\ell(t_1)<\ell(t^*(\eta_1))<\ell(t_2)$ for any t_1,t_2 such that $t_1<t^*(\eta_1)<t_2$. Denote by $l'_+(\cdot)$ and $l'_-(\cdot)$ the right and left derivative functions of $\ell(\cdot)$ respectively. (side derivatives must exist for monotonic functions.) The definition of $t^*(\eta_1)$ indicates that $Q(\eta_1,t_1)>Q(\eta_1,t^*(\eta_1))$ and $Q(\eta_1,t_2)\ge Q(\eta_1,t^*(\eta_1))$, where this can be rewritten in the following inequalities

$$\frac{\ell(-t_1)-\ell(-t^*(\eta_1))}{\ell(t^*(\eta_1))-\ell(t_1)}>\frac{\eta_1}{1-\eta_1}\ge\frac{\ell(-t_2)-\ell(-t^*(\eta_1))}{\ell(t^*(\eta_1))-\ell(t_2)}.$$

Divide both the numerators and denominators of the left hand side and the right hand side of the above inequality by $t^*(\eta_1)-t_1$ and $t^*(\eta_1)-t_2$ respectively. By taking limit as $t_1 \to t^*(\eta_1)_-$ and $t_2 \to t^*(\eta_1)_+$, we get

$$\frac{\ell'_+(-t^*(\eta_1))}{\ell'_-(t^*(\eta_1))} \geq \frac{\eta_1}{1-\eta_1} \geq \frac{\ell'_-(-t^*(\eta_1))}{\ell'_+(t^*(\eta_1))}.$$

According to the convexity of ℓ and the fact $\ell(t_1) < \ell(t^*(\eta_1))$, we know that $0 < l'_-(t^*(\eta_1)) \leq l'_+(t^*(\eta_1))$. Then the above inequality is well defined. Since $\eta_2 > \eta_1$, we know $\frac{\eta_2}{1-\eta_2} > \frac{\eta_1}{1-\eta_1}$. If $\frac{\eta_2}{1-\eta_2} > \frac{\ell'_+(-t^*(\eta_1))}{\ell'_-(t^*(\eta_1))}$, then by the definition of side derivative, there exists a $t < t^*(\eta_1)$ such that $\frac{\eta_2}{1-\eta_2} > \frac{\ell(-t)-\ell(-t^*(\eta_1))}{\ell(t^*(\eta_1))-\ell(t)}$, which is equivalent to $Q(\eta_2, t) < Q(\eta_2, t^*(\eta_1))$, hence we conclude $t^*(\eta_2) < t^*(\eta_1)$. If $\frac{\eta_2}{1-\eta_2} \in (\frac{\eta_1}{1-\eta_1}, \frac{\ell'_+(-t^*(\eta_1))}{\ell'_-(t^*(\eta_1))}] \subset [\frac{\ell'_+(-t^*(\eta_1))}{\ell'_-(t^*(\eta_1))}, \frac{\ell'_-(-t^*(\eta_1))}{\ell'_+(t^*(\eta_1))}]$, this is equivalent to that $t^*(\eta_1)$ is also a minimizer of $Q(\eta_2, \cdot)$, hence we conclude $t^*(\eta_2) \leq t^*(\eta_1)$. The proof is then completed. □

PROOF OF PROPOSITION 6. Let $\eta_1 = \eta(x, x'), \eta_2 = \eta(x, x), \eta_3 = \eta(x', x')$, then $d_\rho(x, x') = t^*(\eta_1), d_\rho(x, x) = t^*(\eta_2)$ and $d_\rho(x', x') = t^*(\eta_3)$. By Lemma 15, we get the desired result immediately. □

CHAPTER 4

Optimal Rates for Gradient Descent Methods with Deep ReLU Networks

---

In this chapter, we provide comprehensive generalization analysis for both GD and SGD with deep ReLU networks. The rest of the chapter is organized as follows. We first review some further related works in Section 4.1. In Section 4.2, we introduce the problem setting. Section 4.3 presents our main results for GD and SGD. We give detailed proofs of this chapter in Section 4.4.

4.1 Other Related Work on GD and SGD

In this section, we review some further works which are closely related to this chapter.

Uniform Convergence Approach: An important line to study the generalization bounds of neural networks is based on uniform convergence, resulting in algorithm-independent analysis [Bartlett et al. 2017; Frei et al. 2023; Golowich et al. 2018; Neyshabur et al. 2015; Parhi and Nowak 2022]. In such analysis, the capacity of the hypothesis space is controlled by using capacity measures such as Rademacher complexities and covering numbers. This type of bound has been used to study gradient descent methods because of its generality. Specifically, the generalization bounds of order $\mathcal{O}(1/\sqrt{n})$ have been established by trading off between the degree to which the algorithm fits the sample and the complexity of the solution through this approach [Arora et al. 2019; Chen et al. 2021a; Ji and Telgarsky 2020].

Generalization under Structured Data Distribution: Considerable works provided generalization analysis of neural networks for classification problems under certain assumptions on the structure of the data distribution [Allen-Zhu et al. 2019a; Brutzkus et al. 2018; Cao

and Gu 2020; Cao and Gu 2019; Nacson et al. 2019; Nitanda et al. 2019; Taheri et al. 2024]. For example, when data are generated by a linearly separable function, [Brutzkus et al. 2018] studied the misclassification error of SGD for training two-layer neural networks with Leaky ReLU activation. [Li and Liang 2018] focused on the problem of learning two-layer ReLU networks via SGD and showed that a small misclassification error can be achieved in the test data set when the data comes from mixtures of well-separated distributions. [Cao and Gu 2020] and [Cao and Gu 2019] established algorithm-dependent misclassification error bounds for deep ReLU networks trained by GD and SGD, respectively, under the assumption that the data can be separated by certain random feature models [Rahimi and Recht 2008; Cao and Gu 2019] with a margin. Recently, [Ji and Telgarsky 2020] and [Nitanda et al. 2019] studied the generalization performance of two-layer networks with ReLU and smooth activation, respectively, and showed that GD and SGD can achieve a small misclassification error under the separation margin of the corresponding kernel assumptions.

Algorithmic Stability Approach: Recently, some works have considered utilizing algorithmic stability to study the generalization performance of gradient descent methods for neural networks [Richards and Rabbat 2021; Richards and Kuzborskij 2021; Lei et al. 2022; Taheri and Thrampoulidis 2024; Wang et al. 2025a]. In particular, [Richards and Rabbat 2021] showed that when the two-layer network with smooth activation is studied, the empirical objective is smooth and weakly convex, the convexity parameter decaying with the order $\mathcal{O}(1/\sqrt{m})$. Based on these observations, [Lei et al. 2022; Richards and Kuzborskij 2021] provided excess population risk bounds of the order $\mathcal{O}(1/\sqrt{n})$ for GD and SGD with polynomial network width. [Wang et al. 2025a] extended their results to multilayer with a general scaling setting. Recent work [Taheri and Thrampoulidis 2024] focusing on logistics loss refined the curvature analysis by showing that empirical risks enjoy a self-bounded weak convexity in a realizable scenario. Very recently, [Taheri et al. 2024] extended their results to learn deep nets. However, the stability-based methods mentioned above all study neural networks with smooth activation functions.

NTK approach: There has been a large amount of literature studying gradient descent methods for overparameterized neural networks in the NTK regime [Allen-Zhu et al. 2019b;

Arora et al. 2019; Du et al. 2018; Du et al. 2019; Guo et al. 2024; Kuzborskij and Szepesvári 2022; Suh et al. 2021; Hu et al. 2021]. For two-layer ReLU neural networks, [Du et al. 2018] demonstrated that randomly initialized GD converges to a globally optimal solution of the empirical risk with a linear convergence rate as long as the network width $m \gtrsim n^6/\lambda_0^4$ with high probability. Here, λ_0 is the smallest eigenvalue of the corresponding NTK Gram matrix. Based on this observation, [Arora et al. 2019] provided a data-dependent generalization bound $\frac{\sqrt{\mathbf{y}^\top(\mathbf{H}^\infty)^{-1}\mathbf{y}}}{\sqrt{n}} + \widetilde{\mathcal{O}}(\frac{1}{\sqrt{n}})$ utilizing the Rademacher complexity under condition $m \gtrsim n^8/\lambda_0^3$. Here, $\mathbf{y}$ is the label vector of the training data and $\mathbf{H}^\infty$ is the NTK Gram matrix. [Kuzborskij and Szepesvári 2022] derived the generalization bound $\mathcal{O}(n^{-\frac{2}{2+d}})$ of GD to learn the target function with additive noise that is uniformly bounded and Lipschitz when $m \gtrsim (n/\lambda_0)^6$. The works mentioned above all require the positivity of the NTK Gram matrix. The works most related to ours are the recent analysis of GD/SGD on two-layer neural networks for the least-square regression [Braun et al. 2024; Cao et al. 2024; Nguyen and Mücke 2024; Nitanda and Taiji 2021]. All these works removed the positive assumption of the NTK Gram matrix to provide generalization bounds. In particular, [Nitanda and Taiji 2021; Nguyen and Mücke 2024] established the optimal excess risk rate $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$ for SGD and GD, respectively. However, all these works focus on neural networks with smooth activation functions.

Gradient descent methods in kernel setting: Finally, we investigate the generalization results of GD and SGD in the context of kernel methods [Cesa-Bianchi et al. 2004; Dieuleveut and Bach 2016; Lin and Rosasco 2017; Pillaud-Vivien et al. 2018; Ying and Pontil 2008]. Most work on SGD within the RKHS framework only provided the generalization analysis allowing a single pass over the training data [Dieuleveut and Bach 2016; Guo and Shi 2019; Ying and Pontil 2008]. The first analysis on multi-pass SGD is [Rosasco and Villa 2015], where a cyclic gradient selection strategy is considered. [Lin and Rosasco 2017] showed that multi-pass SGD with a stochastic choice of the gradient can achieve optimal excess risk rate $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$ for any $\beta > 0$ satisfying $2\beta + \gamma > 1$, matching those of ridge regression [Caponnetto and De Vito 2007; Smale and Zhou 2007]. They also extended their results to mini-batch SGD and GD. Recently, [Mücke et al. 2019] developed the generalization bounds for mini-batch SGD with tail averaging.

4.2 Problem Formulation on GD and SGD

In this chapter, we assume for any $x \in \mathcal{X} \subset \mathbb{R}^p$ and $y \in \mathcal{Y}$, $\|x\|_2 = 1$ and $|y| \leq 1$, where $\|\cdot\|_2$ is the standard Euclidean norm. Denote by $S = \{z_i = (x_i, y_i) : i = 1, \ldots, n\}$ a training dataset drawn from an unknown distribution ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Based on S, we aim to build a predictor $f : \mathcal{X} \to \mathbb{R}$, whose performance is measured by the expected risk

$$\mathcal{E}(f) := \frac{1}{2}\mathbb{E}_{(x,y)\sim\rho}[(y - f(x))^2].$$

Since the distribution ρ is unknown in practice, we instead minimize the empirical risk defined by

$$\mathcal{E}_z(f) := \frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2.$$

A minimizer of the expected risk is the regression function $f_\rho(x) = \mathbb{E}[y|x]$, where $\mathbb{E}[\cdot|x]$ denotes the conditional expectation given x.

In this chapter, we are interested in a prediction model f parameterized by $\mathbf{W}$ in some parameter space $\mathcal{W}$ with a neural network structure. In particular, we focus on L-layer deep ReLU neural networks with width m of the form

$$f_{\mathbf{W}}(x) = \mathbf{a}^\top\sqrt{\frac{2}{m}}\sigma\Big(\mathbf{W}^L \cdots \sqrt{\frac{2}{m}}\sigma\big(\mathbf{W}^1 x\big)\Big), \tag{4.1}$$

where $x \in \mathcal{X}$ is the input, $\sigma(\cdot) = \max\{\cdot, 0\}$ is the ReLU activation, $\mathbf{W} = \big(\mathbf{W}^1, \ldots, \mathbf{W}^L\big) \in \mathcal{W}$ with $\mathcal{W} := \mathbb{R}^{m\times d} \times (\mathbb{R}^{m\times m})^{L-1}$ denoting the collection of weight matrices for all layers, and $\mathbf{a} = (a_1, \ldots, a_m)^\top \in \mathbb{R}^m$ is the weight vector of the output layer. In the above formulation, $\mathbf{W}^1 \in \mathbb{R}^{m\times d}$ and $\mathbf{W}^l \in \mathbb{R}^{m\times m}$ for $l = 2, \ldots, L$ is the weight of the l-hidden layer. We denote $(\mathbf{w}_r^l)^\top$ the r-th row of $\mathbf{W}^l$ for $l \in [L] := \{1, \ldots, L\}$. For the simplicity of argument, we assume m is even. We use the notations $\mathcal{E}(\mathbf{W}) = \mathcal{E}(f_{\mathbf{W}})$, $\mathcal{E}_z(\mathbf{W}) = \mathcal{E}_z(f_{\mathbf{W}})$, and $\ell(\mathbf{W}; z) = \frac{1}{2}(y - f_{\mathbf{W}}(x))^2$.

In this chapter, we are concerned with two notable algorithms to minimize the empirical risk, i.e., GD and SGD. We will consider symmetric initialization of GD and SGD, which are widely used in the theoretical analysis of neural networks [Kuzborskij and Szepesvári 2022; Nguyen and Mücke 2024; Nitanda and Taiji 2021; Xu and Zhu 2024]. Especially, we

adopt Gaussian initialization for all weights while the weights of the last layer are initialized additionally using the symmetric weights and uniform initialization for the output layer weight defined as follows:

$$\begin{aligned}
&\textit{for the first } [L-1] \textit{ layer: } \quad \mathbf{w}_r^1(0) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_p) \textit{ and } \mathbf{w}_r^l(0) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_m) \textit{ for all } r \in [m], \\
&\textit{for the last layer: } \quad \mathbf{w}_r^L(0) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_m) \textit{ for } r \in \big[\frac{m}{2}\big], \textit{ and } \mathbf{w}_{r+\frac{m}{2}}^L(0) = \mathbf{w}_r^L(0), \\
&\textit{for the output layer: } \quad a_r \overset{\text{i.i.d.}}{\sim} \textit{Unif}\left(\{-1, 1\}\right) \textit{for } r \in \big[\frac{m}{2}\big], \textit{ and } a_{r+\frac{m}{2}} = -a_r.
\end{aligned} \tag{4.2}$$

Symmetric initialization is mainly used to ensure that the initial function $f_{\mathbf{W}(0)}(x)$ is 0 for any $x \in \mathcal{X}$, which simplifies theoretical analysis. As noted in [Nguyen and Mücke 2024; Nitanda and Taiji 2021], this requirement can be relaxed by taking into account the additional error caused by non-symmetric initialization. Moreover, this symmetric trick does not affect the concentration properties of the NTK for deep ReLU networks (see the discussion in Section 4.4.5). For a differentiable function F on $\mathcal{W}$, we denote $\partial F(\mathbf{W}_0) = \frac{\partial F(\mathbf{W})}{\partial \mathbf{W}}|_{\mathbf{W}=\mathbf{W}_0}$ and $\partial_l F(\mathbf{W}_0) = \frac{\partial F(\mathbf{W})}{\partial \mathbf{W}^l}|_{\mathbf{W}=\mathbf{W}_0}$ for all $l \in [L]$.

DEFINITION 11 (Gradient Descent). Let $\mathbf{W}(0) \in \mathcal{W}$ be the initialization generated by (4.2) and $\eta > 0$ be the step size. GD updates $\{\mathbf{W}(k) : k \in \mathbb{N}\}$ by

$$\mathbf{W}(k+1) = \mathbf{W}(k) - \eta \partial \mathcal{E}_z(\mathbf{W}(k)). \tag{4.3}$$

DEFINITION 12 (Stochastic Gradient Descent). Let $\mathbf{W}(0) \in \mathcal{W}$ be the initialization generated by (4.2) and $\eta > 0$ be the step size. SGD updates $\{\mathbf{W}(k) : k \in \mathbb{N}\}$ by

$$\mathbf{W}(k+1) = \mathbf{W}(k) - \eta \partial \ell(\mathbf{W}(k); z_{i_k}), \tag{4.4}$$

where i_k is uniformly drawn from $[n]$.

We are interested in the generalization performance of a model $f_{\mathbf{W}}$ trained by GD and SGD with T iterations, measured in terms of the *excess population risk*

$$\varepsilon_{risk}(f_{\mathbf{W}(T)}) = \mathcal{E}(f_{\mathbf{W}(T)}) - \mathcal{E}(f_\rho),$$

i.e., the discrepancy between the expected risks of $f_{\mathbf{W}(T)}$ and f_ρ. For the least squares regression, it has been shown in [Cucker and Zhou 2007] that $\varepsilon_{risk}(f_{\mathbf{W}(T)})$ can be further cast as

$$\varepsilon_{risk}(f_{\mathbf{W}(T)}) = \frac{1}{2}\|f_{\mathbf{W}(T)} - f_\rho\|_\rho^2.$$

Here, $\|\cdot\|_\rho$ is the L_2-norm defined as $\|f\|_\rho = (\int_{\mathcal{X}} |f(x)|^2 d\rho_x(x))^{1/2}$ where ρ_x denotes the marginal distribution of ρ on $\mathcal{X}$.

In the remainder of the chapter, we focus on studying $\|f_{\mathbf{W}(T)} - f_\rho\|_\rho^2$. The key idea of the analysis is to introduce kernel methods as a bridge between the neural network and the best model f_ρ. To this end, we require the concept of the neural tangent kernel (NTK) [Jacot et al. 2018]. In our setting, the NTK $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with symmetric initialization is defined, for any $x, x' \in \mathcal{X}$, by

$$K(x, x') = 2\mathbb{E}[\sigma(U^{L-1}(x))\sigma(U^{L-1}(x'))]q^L(x, x'), \tag{4.5}$$

where $\{(U^l(x), U^l(x'))\}_{l=1}^{L-1}$ are pairs of bivariate normal variables defined iteratively by $(U^l(x), U^l(x')) \sim \mathcal{N}(0, \Sigma^{l-1}(x, x'))$ with

$$\Sigma^{l-1}(x,x') = 2\begin{pmatrix} \mathbb{E}[\sigma^2(U^{l-1}(x))] & \mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))] \\ \mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))] & \mathbb{E}[\sigma^2(U^{l-1}(x'))] \end{pmatrix},$$

$$\Sigma^0(x,x') = \begin{pmatrix} 1 & \langle x, x'\rangle_2 \\ \langle x, x'\rangle_2 & 1 \end{pmatrix}$$

and

$$q^l(x, x') = \big(\pi - \arccos(p^{l-1}(x, x'))\big)/\pi$$

with

$$p^{l-1}(x, x') = \frac{\mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))]}{\sqrt{\mathbb{E}[\sigma^2(U^{l-1}(x))]\mathbb{E}[\sigma^2(U^{l-1}(x'))]}}.$$

Note that for all $x, x \in \mathcal{X}$ and $l \in [L]$, $\mathbb{E}[\sigma(U^l(x))\sigma(U^l(x'))]$ is fixed and deterministic, and does not involve any randomness.

Let $\mathcal{H}_K$ be the RKHS associated with the kernel K, with inner product and induced norm denoted by $\langle\cdot,\cdot\rangle_K$ and $\|\cdot\|_K$, respectively. Let $\mathcal{L}^2_{\rho_x} = \{f : \mathcal{X} \to \mathbb{R} : \|f\|_\rho < \infty\}$ be the space of square-integrable functions on $\mathcal{X}$ with respect to ρ_x. We introduce the integral operator $\mathbf{L} : \mathcal{L}^2_{\rho_x} \to \mathcal{L}^2_{\rho_x}$, defined by $\mathbf{L}f = \int_{\mathcal{X}} K(\cdot, x)f(x)d\rho_x(x)$. One can show $\int_{\mathcal{X}} K(x,x)d\rho_x(x) \leq 1$ (see Property 1 in Section 4.4.1), hence $\mathbf{L}$ is a compact, self-adjoint and positive operator, which has the eigen-decomposition $\mathbf{L}f = \sum_{i=1}^{\infty} \lambda_i \langle f, \Phi_i\rangle_{\mathcal{L}^2_{\rho_x}} \Phi_i$. Here, $\{(\lambda_i, \Phi_i)\}$ are pairs of eigenvalues and orthogonal eigenfunctions in $\mathcal{L}^2_{\rho_x}$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, and $\{\Phi_i\}_{i=1}^{\infty}$ forms an orthonormal basis of $\mathcal{L}^2_{\rho_x}$. For $s \in \mathbb{R}$, we define the power $\mathbf{L}^s$ as, for any $f \in \mathcal{L}^2_{\rho_x}$, $\mathbf{L}^s(f) = \sum_{i=1}^{\infty} \lambda_i^s \langle f, \Phi_i\rangle_{\mathcal{L}^2_{\rho_x}} \Phi_i$. For a bounded and positive linear operator A on a separable Hilbert space $\mathcal{H}$ with orthonormal basis $\{e_i\}_{i=1}^{\infty}$, the trace of A is defined by $tr(A) = \sum_{i=1}^{\infty} \langle Ae_i, e_i\rangle_{\mathcal{H}}$.

To analyze the performance of kernel methods, we impose the following standard assumptions on the capacity of the hypothesis space and the complexity of f_ρ.

ASSUMPTION 11 (Effective dimension). *For some $\gamma \in [0,1]$ and $c_\gamma \geq 1$, there holds $tr\big(\mathbf{L}(\mathbf{L} + \lambda\mathbf{I})^{-1}\big) = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \lambda} \leq c_\gamma \lambda^{-\gamma}$ for all $\lambda > 0$.*

In the above assumption, the quantity $tr\big(\mathbf{L}(\mathbf{L} + \lambda\mathbf{I})^{-1}\big)$ is called as the effective dimension [Caponnetto and De Vito 2007] or the degrees of freedom [Zhang 2005]. Note that $\mathbf{L}$ is a trace class operator satisfying $tr(\mathbf{L}) = \sum_i \lambda_i = \int_{\mathcal{X}} K(x,x)d\rho_x(x) \leq 1$, then Assumption 11 is always true for $\gamma = 1$ and $c_\gamma = 1$. In this sense, the case of $\gamma = 1$ is often referred to as the capacity independent setting [Cucker and Zhou 2007]. Assumption 11 holds true if $\mathbf{L}$ is of finite rank (corresponds to $\gamma = 0$) or the eigenvalues $\{\lambda_i\}$ satisfy a polynomial decaying condition $\lambda_i \lesssim i^{-1/\gamma}$ for $\gamma \in (0,1]$. The specific decay rates of the eigenvalues have been studied for some specific settings [Bach 2017; Bietti and Mairal 2019; Bietti and Bach 2021; Hu et al. 2021; Scetbon and Harchaoui 2021]. For example, under the assumption that the input x is uniformly distributed on a unit sphere, [Hu et al. 2021] showed that the eigenvalues of the NTK associated with two-layer ReLU networks decay as $\lambda_i \asymp i^{-\frac{p}{p-1}}$.

ASSUMPTION 12 (Source condition). *There exist $\beta > 0$ and $B > 0$, such that $\|\mathbf{L}^{-\beta} f_\rho\|_\rho \leq B$.*

Assumption 12 is commonly used in nonparametric regression [Cucker and Smale 2002], which quantifies the smoothness (regularity) of the regression function f_ρ. The larger the value of β, the smoother f_ρ becomes and, consequently, the more stringent the assumption. In particular, if $\beta = 1/2$, then this assumption indicates $f_\rho \in \mathcal{H}_K$, which implies that there exists at least one minimizer of population risk belonging to the RKHS $\mathcal{H}_K$.

4.3 Main Results on Optimal Rates for GD and SGD

Before presenting our main results, we first introduce some necessary definitions and notations. Given the initialization $\mathbf{W}(0)$, define the feature map $\Phi_m : \mathcal{X} \to \mathcal{W}$ by

$$\Phi_m(x) = \partial f_{\mathbf{W}(0)}(x) = (\partial_1 f_{\mathbf{W}(0)}(x), \ldots, \partial_L f_{\mathbf{W}(0)}(x)).$$

With this feature map, we define a PSD kernel $K^m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$K^m(x, x') = \left\langle \partial f_{\mathbf{W}(0)}(x), \partial f_{\mathbf{W}(0)}(x') \right\rangle_2 = \sum\nolimits_{l=1}^{L} \left\langle \partial_l f_{\mathbf{W}(0)}(x), \partial_l f_{\mathbf{W}(0)}(x') \right\rangle_2, \tag{4.6}$$

where $\langle \cdot, \cdot \rangle_2$ is the inner product of a vector or a matrix. Here, K^m can be seen as a random feature approximation of the NTK K. According to Theorem 4.21 in [Steinwart and Christmann 2008], there exists a unique RKHS $\mathcal{H}_m$ associated with the kernel K^m given by

$$\mathcal{H}_m = \{f : \mathcal{X} \to \mathbb{R} \ : \ \exists\, \mathbf{W} \in \mathcal{W} \text{ such that } f(x) = \langle \mathbf{W}, \Phi_m(x) \rangle_2\},$$

whose corresponding norm is defined, for any $f \in \mathcal{H}_m$, by

$$\|f\|_{\mathcal{H}_m} = \inf\{(\textstyle\sum_{l=1}^{L} \|\mathbf{W}^l\|_2^2)^{1/2} : \mathbf{W} \in \mathcal{W} \text{ with } f(x) = \langle \mathbf{W}, \Phi_m(x) \rangle_2\}.$$

We further define the linear approximation of $f_{\mathbf{W}}$ at the initialization $\mathbf{W}(0)$ by

$$f_{\mathbf{W}}^{\text{lin}}(x) = f_{\mathbf{W}(0)}(x) + \langle \partial f_{\mathbf{W}(0)}(x), \mathbf{W} - \mathbf{W}(0) \rangle_2.$$

Let $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$ and $\mathbf{K}^m = (K^m(x_i, x_j))_{i,j=1}^n$ be the Gram matrices with kernels K and K^m, respectively. For a function ψ on an arbitrary space Ω, we denote $\|\psi\|_\infty = \sup_{\omega \in \Omega} |\psi(\omega)|$.

4.3.1 Optimal Rates for Gradient Descent

Define the functions $K_x^m \in \mathcal{H}_m, K_x \in \mathcal{H}_K$ by $K_x^m(x') = K^m(x, x')$ and $K_x(x') = K(x, x')$ for any $x, x' \in \mathcal{X}$. If we regard the empirical risk $\mathcal{E}_z(\cdot)$ as a functional on the RKHS $\mathcal{H}_m$ and $\mathcal{H}_K$, the iteration of GD for least-square regression in $\mathcal{H}_m$ and $\mathcal{H}_K$ can be defined as

$$g_{k+1}^m = g_k^m - \frac{\eta}{n}\sum_{i=1}^n (g_k^m(x_i) - y_i)K_{x_i}^m \text{ for any } k \in \mathbb{N} \text{ with } g_0^m = 0, \tag{4.7}$$

$$g_{k+1} = g_k - \frac{\eta}{n}\sum_{i=1}^n (g_k(x_i) - y_i)K_{x_i} \text{ for any } k \in \mathbb{N} \text{ with } g_0 = 0. \tag{4.8}$$

Let $\mathbf{W}(T)$, g_T^m and g_T be produced by (4.3), (4.7) and (4.8) with T iterations, respectively. Consider $f_{\mathbf{W}(T)}^{\text{lin}}$, g_T^m and g_T as bridges connecting $f_{\mathbf{W}(T)}$ and f_ρ, we introduce the following error decomposition

$$\varepsilon_{risk}\big(f_{\mathbf{W}(T)}\big) \lesssim \big\|f_{\mathbf{W}(T)} - f_{\mathbf{W}(T)}^{\text{lin}}\big\|_\rho^2 + \big\|f_{\mathbf{W}(T)}^{\text{lin}} - \mathbf{S}_m g_T^m\big\|_\rho^2 + \big\|\mathbf{S}_m g_T^m - \mathbf{S}g_T\big\|_\rho^2 + \big\|\mathbf{S}g_T - f_\rho\big\|_\rho^2, \tag{4.9}$$

where $\mathbf{S}_m : \mathcal{H}_m \hookrightarrow \mathcal{L}_{\rho_x}^2$ and $\mathbf{S} : \mathcal{H}_K \hookrightarrow \mathcal{L}_{\rho_x}^2$ are the inclusion mappings that map $g_T^m \in \mathcal{H}_m$ to $\mathbf{S}_m g_T^m \in \mathcal{L}_{\rho_x}^2$ and $g_T \in \mathcal{H}_K$ to $\mathbf{S}g_T \in \mathcal{L}_{\rho_x}^2$, respectively. We will state the estimates for the above four terms in the subsequent context respectively and present the upper bounds of $\varepsilon_{risk}(f_{\mathbf{W}(T)})$. We assume $\eta T \geq 1$ and denote $C \geq 1$ as an absolute value which may differ from line to line.

We begin by estimating the term $\|f_{\mathbf{W}(T)} - f_{\mathbf{W}(T)}^{\text{lin}}\|_\rho^2$ on the right-hand side of (4.9). Since the population distribution ρ is unknown, in the following proposition we employ the $\|\cdot\|_\infty$-norm to control the $\|\cdot\|_\rho$-norm of $f_{\mathbf{W}(T)} - f_{\mathbf{W}(T)}^{\text{lin}}$. In this sense, the established bound is the worst-case one which holds true for any population distribution ρ. The detailed proof is deferred to Section 4.4.3.

PROPOSITION 7. *Let $\delta \in (0, 1)$. Assume $\eta \leq 1/5$ and*

$$m \gtrsim L^{22} p^3 (\eta T)^7 \log^3(m/\delta). \tag{4.10}$$

Then, with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ *over the initialization* $(\mathbf{a}, \mathbf{W}(0))$*, there holds*

$$\|f_{\mathbf{W}(T)} - f^{lin}_{\mathbf{W}(T)}\|_\rho^2 \leq \|f_{\mathbf{W}(T)} - f^{lin}_{\mathbf{W}(T)}\|_\infty^2 \lesssim \frac{L^{\frac{14}{3}}(\eta T)^{\frac{4}{3}}}{m^{\frac{1}{3}}}.$$

The estimate of $\|f_{\mathbf{W}(T)} - f^{\text{lin}}_{\mathbf{W}(T)}\|_\rho^2$ is controlled by showing the trajectory of GD/SGD is always near the initialization, which critically depends on forward and backward propagation estimates at random initialization. The work [Xu and Zhu 2024] provided such estimates with upper bounds depend exponentially on the number of layers L. Applying these estimates to our approach leads to the unexpected bound $\|f_{\mathbf{W}(T)} - f^{\text{lin}}_{\mathbf{W}(T)}\|_\rho^2 \lesssim C^L m^{-\frac{1}{3}}\text{Poly}(\eta T)$ valid when $m \gtrsim C^L\text{Poly}(\eta T, d)$. Meanwhile, [Zou et al. 2020; Chen et al. 2021b] conducted fine-grained analyses for forward and backward propagation for classification problems, significantly relaxing the required network width from an exponential to a polynomial scaling. However, their approach cannot be directly applied to our setting, as we require these results to hold uniformly over the entire input space $\mathcal{X}$ to control the $\|\cdot\|_\infty$-norm, while their results are typically restricted to the training dataset S. We extend their results from the finite training set S to the full input space $\mathcal{X}$, reducing the requirement of the width to a polynomial scaling.

Proposition 8 presents the estimation of $\|f^{\text{lin}}_{\mathbf{W}(T)} - \mathbf{S}_m g_T^m\|_\rho^2$, whose proof can be found in Section 4.4.3.

PROPOSITION 8. *Let* $\delta \in (0,1)$*. Assume* $\eta \leq 1/5$*,* $\eta T \leq n(36\log(2n/\delta))^{-1}$ *and* (4.10)*. Then, with probability at least* $1 - L\exp\left(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})\right) - \delta$ *over the initialization* $(\mathbf{a}, \mathbf{W}(0))$ *and sampling, there holds*

$$\|f^{lin}_{\mathbf{W}(T)} - \mathbf{S}_m g_T^m\|_\rho^2 \lesssim \frac{L^{\frac{14}{3}}(\eta T)^{\frac{7}{3}}}{m^{\frac{1}{3}}}.$$

Propositions 7 and 8 jointly demonstrate the almost equivalence of the GD trajectories for a deep ReLU network and for the corresponding NTK K^m under overparameterization, i.e., $\|f_{\mathbf{W}(T)} - \mathbf{S}_m g_T^m\|_\rho^2 \lesssim L^{\frac{14}{3}}(\eta T)^{\frac{7}{3}} m^{-\frac{1}{3}}$ under the condition $m \gtrsim \text{Poly}(L, d, \eta T)$. This implies the larger the width of the network m, the closer the two trajectories are and the more the behavior of $f_{\mathbf{W}(T)}$ is similar to that of g_T^m. [Nitanda and Taiji 2021] established this estimate

for the trajectory of the SGD average stream. They showed that these two trajectories behave almost the same when the network width m scales exponentially with n. Subsequently, [Cao et al. 2024] reduced the requirement of m to the polynomial degree. However, both of these two works are limited to two-layer networks with smooth activation. Our result demonstrates that even for deep networks with non-smooth ReLU activation, a polynomially large width is sufficient to ensure the alignment of the learning trajectories.

We estimate $\|\mathbf{S}_m g_T^m - \mathbf{S} g_T\|_\rho^2$ in the following proposition, whose proof is deferred to Section 4.4.3.

PROPOSITION 9. *Let $\delta \in (0,1)$. Assume $\eta \leq 1/4$ and (4.10). Then, with probability at least $1 - L\exp\left(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})\right) - \delta$ over the initialization $(\mathbf{a}, \mathbf{W}(0))$, there holds*

$$\|\mathbf{S}_m g_T^m - \mathbf{S} g_T\|_\rho^2 \leq \|g_T^m - g_T\|_\infty^2 \lesssim (\eta T)^4 \|K^m - K\|_\infty^2 \lesssim \frac{L(\eta T)^4}{m^{\frac{1}{3}}}.$$

The above proposition shows that the distance between the GD iterates in $\mathcal{H}_m$ and $\mathcal{H}_K$ can be controlled by the discrepancy between their respective kernels, K^m and K. In fact, this result can be extended to any pair of RKHS with bounded kernels. Specifically, for arbitrary RKHS $\mathcal{H}_1, \mathcal{H}_2$ with bounded kernels K^1, K^2, let g_T^1 and g_T^2 denote the corresponding GD iterates (defined analogously to (4.7) with K^m replaced by K^1 and K^2), respectively. Then, it follows that $\|g_T^1 - g_T^2\|_\infty \lesssim (\eta T)^2 \|K^1 - K^2\|_\infty$. In addition, the work [Xu and Zhu 2024] proved that $\|K^m - K\|_\infty \lesssim C^L m^{-\frac{1}{6}}\sqrt{d}$ valid when m depends exponentially on L. We improved their bound to $\|K^m - K\|_\infty \lesssim m^{-\frac{1}{6}}\sqrt{pL}$ with the reduced condition $m \gtrsim \mathrm{Poly}(L, n, d)$. More details can be found in Lemma 20.

Finally, we provide an estimate for the last term, $\|\mathbf{S} g_T - f_\rho\|_\rho^2$, which captures the performance of GD within $\mathcal{H}_K$. The detailed proof is presented in Section 4.4.3.

PROPOSITION 10. *Suppose Assumptions 11 and 12 hold. Assume $\eta \leq 1$ and $\eta T \leq n(9\log(2n/\delta))^{-1}$. Then, with probability at least $1 - \delta$ over sampling, there holds*

$$\|\mathbf{S} g_T - f_\rho\|_\rho^2 \lesssim \left(\frac{\eta T}{n^2} + \frac{(\eta T)^\gamma}{n} + \frac{(\eta T)^{1-2\beta}}{n}\right)\log^4\left(\frac{T}{\delta}\right) + (\eta T)^{-2\beta}.$$

Propositions 9 and 10 together provide an estimate for $\|\mathbf{S}_m g_T^m - f_\rho\|_\rho^2$. In this remark, we highlight the technical novelty of our approach. For two-layer neural networks, previous work [Carratino et al. 2018; Nitanda et al. 2019] estimated this term by introducing an intermediate term $h^m \in \mathcal{H}_m$, separately bounding $\|\mathbf{S}_m(g_T^m - h^m)\|_\rho^2$ and $\|\mathbf{S}_m h^m - f_\rho\|_\rho^2$. Here, h^m is either the minimizer of the regularized population risk over $\mathcal{H}_m$ [Nitanda and Taiji 2021] or GD for the population risk in $\mathcal{H}_m$ [Nguyen and Mücke 2024]. One can show that $\|\mathbf{S}_m(g_T^m - h_m)\|_\rho^2 \lesssim n^{-\frac{2\beta}{2\beta+\tilde{\gamma}}}$ with $\tilde{\gamma}$ the effective dimension of $\mathcal{H}_m$. Hence, to achieve optimal rates, it is essential to demonstrate that the effective dimension of $\mathcal{H}_m$ matches that of $\mathcal{H}_K$, i.e., $\tilde{\gamma} = \gamma$. As discussed in the introduction, this equivalence naturally holds for $\gamma = 1$. When $\gamma < 1$, it is established by treating K^m as a sum of i.i.d. random kernels with mean K. However, this structure is not valid for deep ReLU networks, as the gradient $\partial_{\mathbf{w}_r^l} f_{\mathbf{W}(0)}(x)$ is influenced not only by the weights $\mathbf{w}_r^l(0)$ of the l-th layer but also by the weights of all preceding layers. In contrast, we introduce $g_T \in \mathcal{H}_K$ as an intermediate term, and separately estimate $\|\mathbf{S}_m g_T^m - \mathbf{S} g_T\|_\rho^2$ and $\|\mathbf{S} g_T - f_\rho\|_\rho^2$ in Propositions 9 and 10.

Combining the above four propositions, we now present our main result on the excess population risk of GD for deep ReLU networks. The detailed proof is deferred to Section 4.4.3.

THEOREM 9. *Suppose Assumptions 11 and 12 hold. For any $\delta \in (0,1)$, assume $\eta \leq 1/5$, $\eta T \leq n(36\log(8n/\delta))^{-1}$ and* (4.10)*. Then, with probability at least $1 - L\exp\big(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})\big) - \delta$ over initialization $(\mathbf{a}, \mathbf{W}(0))$ and sampling, there holds*

$$\mathcal{E}(f_{\mathbf{W}(T)}) - \mathcal{E}(f_\rho) \lesssim \frac{L^{\frac{14}{3}}(\eta T)^4}{m^{\frac{1}{3}}} + \Big(\frac{\eta T}{n^2} + \frac{(\eta T)^\gamma}{n} + \frac{(\eta T)^{1-2\beta}}{n}\Big)\log^4\Big(\frac{T}{\delta}\Big) + (\eta T)^{-2\beta}.$$

We point out that our result does not need the widely adopted positivity assumption on the NTK Gram matrix $\mathbf{K}^m$ to learn ReLU networks [Arora et al. 2019; Du et al. 2018], i.e., the smallest eigenvalue of $\mathbf{K}^m$ is strictly larger than 0. Previous work [Nitanda and Taiji 2021; Su and Yang 2019] has shown that this assumption can be overly restrictive, as the smallest eigenvalue of $\mathbf{K}^m$ tends to zero as the size of the training set increases.

The following corollary, derived from Theorem 9, shows that when the network width scales polynomially with the sample size n, dimension p, and the number of layers L, GD with

a deep ReLU network can achieve the optimal excess risk rate $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$, with a gradient complexity of $\mathcal{O}(n^{1+\frac{1}{2\beta+\gamma}})$.

COROLLARY 1. *Suppose Assumptions 11 and 12 hold and* $2\beta+\gamma > 1$. *For any* $\delta \in (0,1)$, *assume that* $n \geq (36(2\beta+\gamma)\beta^{-1})^{\frac{2\beta+\gamma}{\beta}} \frac{16}{\delta}$ *and* $m \gtrsim L^{14}\max\{L^8 p^3 n^{\frac{7}{2\beta+\gamma}} \log^3(npL/\delta), n^{\frac{6\beta+12}{2\beta+\gamma}}\}$. *Choosing* $T = \lceil n^{\frac{1}{2\beta+\gamma}} \rceil$ *and* $\eta \leq 1/5$ *as a constant yields that, with a probability of at least* $1-\delta$ *over initialization* $(\mathbf{a}, \mathbf{W}(0))$ *and sampling, there holds*

$$\mathcal{E}(f_{\mathbf{W}(T)}) - \mathcal{E}(f_\rho) \lesssim n^{-\frac{2\beta}{2\beta+\gamma}} \log^4\left(\tfrac{n}{\delta}\right).$$

Under Assumptions 11 and 12, the work [Lin and Rosasco 2017] proved that the optimal excess risk rate $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$ can be achieved by GD in the kernel setting, with a gradient complexity of $\mathcal{O}(n^{1+\frac{1}{2\beta+\gamma}})$, when $2\beta+\gamma > 1$. Corollary 1 shows, provided that the network width scales polynomially with n, p, and L, GD with a deep ReLU network can replicate the classical results in the kernel setting. It implies that the learning capability of GD with a deep ReLU network is competitive with that of the classical kernel regime. Moreover, as β and γ increase, the required network width m and the gradient complexity become less restrictive. In particular, in the capacity independent case, that is, $\gamma = 1$, the optimal rate $\mathcal{O}(n^{-\frac{2\beta}{2\beta+1}})$ can be derived that matches the kernel setting [Ying and Pontil 2008].

Discussion with the existing work.

The study most relevant to our work on GD is [Nguyen and Mücke 2024; Wang et al. 2025b], which provided the excess population risk bounds for two-layer neural networks with smooth and ReLU activation, respectively. Specifically, [Nguyen and Mücke 2024] established the optimal excess risk $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$ under Assumptions 11 and 12, assuming the network width $m \gtrsim \text{Poly}(p,n)$. In their analysis, the smoothness of the activation function plays a crucial role especially for ensuring the boundedness of the second partial derivatives of $f_{\mathbf{W}}$ at $\mathbf{W}(0)$. [Wang et al. 2025b] extended their results to two-layer ReLU networks. They proved that the minimax-optimal excess risk rate can be achieved by GD when the network width depends polynomially on n and p. In contrast, our work provides the first minimax-optimal excess risk rates for deep neural networks with ReLU activation functions under the

Work	Activation	Layer	Setting	Width
GD [Nguyen and Mücke 2024]	smooth	shallow	$\beta > 0$ and $2\beta + \gamma > 1$	$\Omega(\text{Poly}(n, p))$
one-pass SGD [Nitanda and Taiji 2021]	smooth	shallow	$\beta \in [1/2, 1]$	$\Omega(\exp(n))$
GD/SGD [Wang et al. 2025b]	ReLU	shallow	$\beta > 0$ and $2\beta + \gamma > 1$	$\Omega(\text{Poly}(\log(n), p)e^{\mathcal{C}^L})$
GD/SGD Ours	ReLU	deep	$\beta > 0$ and $2\beta + \gamma > 1$	$\Omega(\text{Poly}(L, n, p))$

TABLE 4.1. Results of GD and SGD with minimax optimal rates $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$ achieved for the least-square regression.

same assumptions, provided $m \gtrsim \text{Poly}(L, p, n)$. Besides, [Kuzborskij and Szepesvári 2022] studied the generalization performance of GD for two-layer ReLU networks under the positive eigenvalue assumption of the NTK Gram matrix, focusing on learning target functions with additive noise that is uniformly bounded and Lipschitz. [Lai et al. 2023] showed that gradient flow in two-layer ReLU networks can achieve a generalization bound of $\mathcal{O}(n^{-\frac{2}{3}})$ when $p = 1$ and $\beta = 1/2$. Table 4.1 summarizes the comparison of our results with the related work.

4.3.2 Optimal Rates for Stochastic Gradient Descent

In this subsection, we present our main results for SGD. We begin by introducing the kernel SGD in the RKHS $\mathcal{H}_m$ based on the random feature approximation K^m

$$f_{k+1}^m = f_k^m - \eta(f_k^m(x_{i_k}) - y_{i_k})K_{x_{i_k}}^m \text{ with } f_0^m = 0. \tag{4.11}$$

Let $\mathbf{W}(T)$ and f_T^m be produced by (4.4) and (4.11) with T iterations, respectively. We consider

$$\mathbb{E}_{\mathcal{A}}\left[\varepsilon_{risk}\left(f_{\mathbf{W}(T)}\right)\right] \lesssim \mathbb{E}_{\mathcal{A}}\left[\left\|f_{\mathbf{W}(T)} - f_{\mathbf{W}(T)}^{\text{lin}}\right\|_\rho^2\right] + \mathbb{E}_{\mathcal{A}}\left[\left\|f_{\mathbf{W}(T)}^{\text{lin}} - \mathbf{S}_m f_T^m\right\|_\rho^2\right] + \mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{S}_m f_T^m - f_\rho\right\|_\rho^2\right], \tag{4.12}$$

where $\mathbb{E}_{\mathcal{A}}[\cdot]$ denotes the expectation with respect to $\{i_k : k \in [T]\}$. In the subsequent context, we will state the estimates for the three terms on the right-hand side of (4.12).

First, we provide the estimate of the first term $\mathbb{E}_{\mathcal{A}}[\|f_{\mathbf{W}(T)} - f^{\text{lin}}_{\mathbf{W}(T)}\|_\rho^2]$, whose proof can be found in Section 4.4.4. Similar to GD, we use $\|\cdot\|_\infty$-norm to control the $\|\cdot\|_\rho$-norm of $f_{\mathbf{W}(T)} - f^{\text{lin}}_{\mathbf{W}(T)}$.

PROPOSITION 11. *Let $\delta \in (0,1)$. Assume $\eta \leq 1/5$, $\eta T \geq 1$ and*

$$m \gtrsim L^{26} p^3 (\eta T)^7 \log^3(m/\delta). \tag{4.13}$$

Then, with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ over initialization $(\mathbf{a}, \mathbf{W}(0))$, there holds

$$\mathbb{E}_{\mathcal{A}}[\|f_{\mathbf{W}(T)} - f^{lin}_{\mathbf{W}(T)}\|_\rho^2] \leq \|f_{\mathbf{W}(T)} - f^{lin}_{\mathbf{W}(T)}\|_\infty^2 \lesssim \frac{L^{\frac{14}{3}}(\eta T)^{\frac{4}{3}}}{m^{\frac{1}{3}}}.$$

The following proposition presents the estimation of the second term $\mathbb{E}_{\mathcal{A}}[\|f^{\text{lin}}_{\mathbf{W}(T)} - \mathbf{S}_m f_T^m\|_\rho^2]$. Due to the randomness of SGD, the proof strategy of Proposition 8 cannot be directly extended to SGD. Instead of estimating the $\|\cdot\|_\rho$-norm of the error term, we control the stronger $\|\cdot\|_\infty$-norm here. The detailed proof is deferred to Section 4.4.4.

PROPOSITION 12. *Let $\delta \in (0,1)$. Assume $\eta \leq 1/5$, $\eta T \geq 1$ and condition* (4.13) *hold. Then, with probability at least $1 - L\exp\big(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})\big) - \delta$ over initialization $(\mathbf{a}, \mathbf{W}(0))$, there holds*

$$\|f^{lin}_{\mathbf{W}(T)} - \mathbf{S}_m f_T^m\|_\rho^2 \leq \|f^{lin}_{\mathbf{W}(T)} - f_T^m\|_\infty^2 \lesssim \frac{L^{\frac{20}{3}}(\eta T)^{\frac{10}{3}}}{m^{\frac{1}{3}}}.$$

Finally, we establish an upper bound for the last term on the right-hand side of (4.12) in the following proposition. To this end, we first estimate the distance between the SGD and GD iterates in the RKHS $\mathcal{H}_m$, i.e, $f_T^m - g_T^m$. This intermediate step, combined with the result of Proposition 10, will allow us to complete the proof of the proposition, which is provided in Section 4.4.4.

PROPOSITION 13. *Suppose Assumptions 11 and 12 and* (4.13) *hold. Let $\delta \in (0,1)$ and $T \in \mathbb{N}$. Assume $0 < \eta \leq (32(\log(T)+1))^{-1}$ and $(\eta T)^{-1} \geq n^{-1}\log(6n/\delta)$. Then, with probability at least $1 - L\exp\big(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})\big) - \delta$ over initialization $(\mathbf{W}(0), \mathbf{a})$*

and sampling

$$\mathbb{E}_{\mathcal{A}}\big[\|\mathbf{S}_m f_T^m - f_\rho\|_\rho^2\big] \lesssim \frac{L^{\frac{20}{3}}(\eta T)^4}{m^{\frac{1}{3}}} + \Big(\eta + \frac{\eta T}{n^2} + \frac{(\eta T)^\gamma + (\eta T)^{1-2\beta}}{n}\Big)\log^4\Big(\frac{T}{\delta}\Big) + (\eta T)^{-2\beta}.$$

Combining the above three propositions, we present our main result on the excess population risk of SGD with deep ReLU networks as follows. The detailed proof is deferred to Section 4.4.4.

THEOREM 10. *Suppose Assumptions 11 and 12 and* (4.13) *hold. For any* $\delta \in (0,1)$*, assume* $0 < \eta \leq (32(\log(T)+1))^{-1}$ *and* $1 \leq \eta T \leq n(36\log(12n/\delta))^{-1}$*. Then, with probability at least* $1 - L\exp\big(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})\big) - \delta$ *over initialization* $(\mathbf{a}, \mathbf{W}(0))$ *and sampling, there holds*

$$\mathbb{E}_{\mathcal{A}}\big[\mathcal{E}(f_{\mathbf{W}(T)}) - \mathcal{E}(f_\rho)\big] \lesssim \frac{L^{\frac{20}{3}}(\eta T)^4}{m^{\frac{1}{3}}} + \Big(\eta + \frac{\eta T}{n^2} + \frac{(\eta T)^\gamma + (\eta T)^{1-2\beta}}{n}\Big)\log^4\Big(\frac{T}{\delta}\Big) + (\eta T)^{-2\beta}.$$

The following corollary, derived from Theorem 10, shows that when the network width scales polynomially with n, p and L, SGD can achieve the optimal excess risk rate $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$ with lower computational cost (in terms of gradient complexity) than GD in Corollary 1.

COROLLARY 2. *Suppose Assumptions 11 and 12 hold and* $2\beta + \gamma > 1$*. For any* $\delta \in (0,1)$*, assume* $n \geq (72(2\beta+\gamma))^{2(2\beta+\gamma)}(\frac{24}{\delta})$ *and* $m \gtrsim L^{20}\max\{L^6 p^3 n^{\frac{7}{2\beta+\gamma}}\log^3(npL/\delta), n^{\frac{6\beta+12}{2\beta+\gamma}}\}$*. Choosing* $T = \lceil n^{\frac{2\beta+1}{2\beta+\gamma}}\rceil$ *and* $\eta = (72\log(24n/\delta))^{-1}n^{-\frac{2\beta}{2\beta+\gamma}}$ *yields that, with probability at least* $1-\delta$ *over initialization* $(\mathbf{a}, \mathbf{W}(0))$ *and sampling, there holds*

$$\mathbb{E}_{\mathcal{A}}[\mathcal{E}(f_{\mathbf{W}(T)}) - \mathcal{E}(f_\rho)] \lesssim n^{-\frac{2\beta}{2\beta+\gamma}}\log^2(n)\log^{2\beta}(n/\delta).$$

Our results suggest that, provided a sufficiently wide network width, SGD with deep ReLU networks can recover the classical results of SGD [Dieuleveut and Bach 2016; Lin and Rosasco 2017] in the kernel setting with the same gradient complexity under similar assumptions.

Comparison with the existing work.

Several works studying generalization performance of deep ReLU networks trained by SGD in the NTK regime [Cao and Gu 2019; Chen et al. 2021b; Zou et al. 2020; Xu and Zhu 2024]. However, most of them focus on classification problem [Cao and Gu 2020; Chen et al. 2021b; Zou et al. 2020]. For regression problems, [Xu and Zhu 2024] studied one-pass SGD in the streaming data setting for deep ReLU networks and demonstrated that the average prediction error $\mathbb{E}_S[(\varepsilon_{risk}(f_{\mathbf{W}(T)}))^{\frac{1}{2}}]$ can converge to zero in expectation, provided that the width of the network scales exponentially with the number of layers L. The precise convergence rate was not specified in [Xu and Zhu 2024]. Under Assumptions 11 and 12, [Nitanda and Taiji 2021] established minimax-optimal rates for one-pass SGD in two-layer neural networks with smooth activations, assuming the network width m scales exponentially with n. [Wang et al. 2025b] extended their results to two-layer ReLU networks and showed that SGD can achieved the minimax-optimal rates under the condition that $m \gtrsim \text{Poly}(n, p)$. We focus on deep neural networks, showing that SGD for deep neural networks can achieve the optimal rates under the condition $m \gtrsim \text{Poly}(L, n, p)$.

4.4 Proofs for Optimal Rates for GD and SGD

Section 4.4.1 introduces the uniform concentration of the NTK. Section 4.4.2 presents some necessary lemmas. Sections 4.4.3 and 4.4.4 give all proofs for both GD and SGD. Section 4.4.5 discusses the properties of the NTK for deep ReLU networks with non-symmetric initialization.

4.4.1 Proofs for Concentration of the NTK

In this section, we provide the uniform concentration of the NTK in our setting.

Denote by $\mathbb{I}\{\cdot\}$ the indicator function (i.e., taking the value 1 if the argument holds true, and 0 otherwise). Given an input $x \in \mathcal{X}$, the L-layer ReLU network can be expressed as the

following specific form

$$f_{\mathbf{W}}(x) = \mathbf{a}^\top \sqrt{\frac{2}{m}} \mathbf{D}^L(x) \mathbf{W}^L \cdots \sqrt{\frac{2}{m}} \mathbf{D}^1(x) \mathbf{W}^1 x, \tag{4.14}$$

where $\mathbf{D}^l(x)$ with $l \in [L]$ is the diagonal sign matrix defined by

$$\mathbf{D}^l(x) = \mathrm{diag}\big\{\mathbb{I}\{\langle \mathbf{w}_r^l, o^{l-1}(x)\rangle_2 \geq 0\}\big\} \in \mathbb{R}^{m\times m} \tag{4.15}$$

with $o^0(x) = x$ and

$$o^{l-1}(x) = \sqrt{\frac{2}{m}} \mathbf{D}^{l-1}(x) \mathbf{W}^{l-1} \cdots \sqrt{\frac{2}{m}} \mathbf{D}^1(x) \mathbf{W}^1 x \text{ for } l = 2, \ldots, L. \tag{4.16}$$

Here, $o^{l-1}(x)$ can be regarded as the output of the $(l-1)$-th layer. By further defining $(V_L^L(x))^\top = \sqrt{\frac{2}{m}} \mathbf{D}^L(x)$ and

$$\big(\mathbf{V}_L^l(x)\big)^\top = \sqrt{\frac{2}{m}} \mathbf{D}^L(x) \mathbf{W}^L \cdots \sqrt{\frac{2}{m}} \mathbf{D}^l(x) \text{ for } l \in [L-1], \tag{4.17}$$

we can rewrite $f_{\mathbf{W}}(x)$ as

$$f_{\mathbf{W}}(x) = \mathbf{a}^\top \big(\mathbf{V}_L^l(x)\big)^\top \mathbf{W}^l o^{l-1}(x) = \big\langle \mathbf{V}_L^l(x) \mathbf{a} \big(o^{l-1}(x)\big)^\top, \mathbf{W}^l \big\rangle_2.$$

The above observation implies that

$$\frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}^l} = \mathbf{V}_L^l(x) \mathbf{a} \big(o^{l-1}(x)\big)^\top.$$

Denote $\|\cdot\|_{op}$ the operator norm of a matrix or an operator. For any $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathcal{W}$, let $\|\mathbf{W} - \widetilde{\mathbf{W}}\|_{op,\infty} = \max_{l\in[L]} \|\mathbf{W}^l - \widetilde{\mathbf{W}}^l\|_{op}$, and, for any $R > 0$, $\mathcal{B}_R(\widetilde{\mathbf{W}}) = \big\{\mathbf{W} \in \mathcal{W} : \|\mathbf{W} - \widetilde{\mathbf{W}}\|_{op,\infty} \leq R\big\}$.

Let $\mathbf{D}_0^l(x)$, $o_0^l(x)$ and $\mathbf{V}_{L,0}^l$ be defined as (4.15), (4.16) and (4.17) with $\mathbf{W} = \mathbf{W}(0)$ for all $l \in [L]$. The following lemma shows that only the performance of the last layer plays a role in defining K^m under the symmetric initialization.

LEMMA 16. *For any $x \in \mathcal{X}$, there holds*

$$\mathbf{a}^\top \mathbf{D}_0^L(x) \mathbf{W}^L(0) = 0 \text{ and } \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^l(0)} = 0 \text{ for any } l \in [L-1].$$

Further,

$$K^m(x,x') = \Big\langle \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}, \frac{\partial f_{\mathbf{W}(0)}(x')}{\partial \mathbf{W}^L(0)} \Big\rangle_2 \textit{ for all } x, x' \in \mathcal{X}.$$

PROOF. Note the r-th row of $\mathbf{D}_0^L(x)\mathbf{W}^L(0)$ is $\mathbb{I}\{\langle \mathbf{w}_r^L(0), o_0^{L-1}(x)\rangle_2 \geq 0\}(\mathbf{w}_r^L(0))^\top$. Since $a_r = -a_{r+\frac{m}{2}}$ and $\mathbf{w}_r^L(0) = \mathbf{w}_{r+\frac{m}{2}}^L(0)$ for all $r \in [\frac{m}{2}]$, there holds

$$\begin{aligned}
\mathbf{a}^\top \mathbf{D}_0^L(x)\mathbf{W}^L(0) &= \sum_{r=1}^{m} a_r \mathbb{I}\{\langle \mathbf{w}_r^L(0), o_0^{L-1}(x)\rangle_2 \geq 0\}(\mathbf{w}_r^L(0))^\top \\
&= \sum_{r=1}^{\frac{m}{2}} a_r \mathbb{I}\{\langle \mathbf{w}_r^L(0), o_0^{L-1}(x)\rangle_2 \geq 0\}(\mathbf{w}_r^L(0))^\top + \sum_{r=1}^{\frac{m}{2}} a_{r+\frac{m}{2}} \mathbb{I}\{\langle \mathbf{w}_{r+\frac{m}{2}}^L(0), o_0^{L-1}(x)\rangle_2 \geq 0\}(\mathbf{w}_r^L(0))^\top \\
&= \sum_{r=1}^{\frac{m}{2}} a_r \mathbb{I}\{\langle \mathbf{w}_r^L(0), o_0^{L-1}(x)\rangle_2 \geq 0\}(\mathbf{w}_r^L(0))^\top - \sum_{r=1}^{\frac{m}{2}} a_r \mathbb{I}\{\langle \mathbf{w}_r^L(0), o_0^{L-1}(x)\rangle_2 \geq 0\}(\mathbf{w}_r^L(0))^\top = 0.
\end{aligned}$$

It further implies

$$\big(\mathbf{V}_{L,0}^l(x)\mathbf{a}\big)^\top = \mathbf{a}^\top \sqrt{\frac{2}{m}}\mathbf{D}_0^L(x)\mathbf{W}^L(0)\cdots\sqrt{\frac{2}{m}}\mathbf{D}_0^l(x) = 0.$$

Combining this observation with $\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^l(0)} = \mathbf{V}_{L,0}^l(x)\mathbf{a}\big(o_0^{l-1}(x)\big)^\top$, we know $\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^l(0)} = 0$ for any $l \in [L-1]$. The first two results of the lemma are proved.

Finally, from (4.6) we get

$$K^m(x,x') = \sum_{l=1}^{L} \Big\langle \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^l(0)}, \frac{\partial f_{\mathbf{W}(0)}(x')}{\partial \mathbf{W}^l(0)} \Big\rangle_2 = \Big\langle \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}, \frac{\partial f_{\mathbf{W}(0)}(x')}{\partial \mathbf{W}^L(0)} \Big\rangle_2, \tag{4.18}$$

which completes the proof. □

The following lemma shows that the initial weights $\mathbf{W}^l(0)$ are bounded by $\mathcal{O}(\sqrt{m})$ with high probability.

LEMMA 17 (Theorem 4.4.5 in [Vershynin 2018]). *With probability at least* $1 - L\exp(-Cm)$ *over the random choice of* $\mathbf{W}(0)$*, there exists an absolute constant* $c_0 > 1$ *such that for any* $l \in [L]$*, there holds*

$$\|\mathbf{W}^l(0)\|_{op} \leq c_0\sqrt{m}. \tag{4.19}$$

In the rest of the proofs, we will assume that the event $\{\|\mathbf{W}^l(0)\|_{op} \leq c_0\sqrt{m} \text{ for all } l \in [L]\}$ holds unless otherwise specified.

We require the following useful lemma which can be found in [Zou et al. 2018] (Corollary A.2, Lemmas A.8, B.1 and B.3 with $m_L = m/2, m_{L-1} = \cdots = m_1 = m$). We note that in the following lemma, Assumptions 3.4 and 3.5 in [Zou et al. 2018] are removed and the training dataset S is replaced by a finite subset $\mathcal{D}$ of $\mathcal{X}$. Denote $\|\cdot\|_0$ the l^0-norm which is the number of nonzero entries of a matrix or a vector.

LEMMA 18. *Let $\mathcal{D} \subset \mathcal{X}$ be a finite subset of $\mathcal{X}$ with cardinality $|\mathcal{D}| = u$. For any $\delta \in (0,1)$, the following statements hold with probability at least $1-\delta$ over the random choice of $\mathbf{W}(0)$ for all $\hat{x} \in \mathcal{D}$.*

(a) Assume $m \geq C\log(uL/\delta)$. For all $l \in [L]$, there holds

$$\big|\|o_0^l(\hat{x})\|_2 - 1\big| \leq Cl\sqrt{\frac{\log(uL/\delta)}{m}}.$$

(b) Assume $m \geq C\log(uL^2/\delta)$. For all $1 \leq l_1 < l_2 \leq L$, there holds

$$\Big\|\sqrt{\frac{2}{m}}\mathbf{W}^{l_2}(0)\prod_{h=l_1}^{l_2-1}\sqrt{\frac{2}{m}}\mathbf{D}_0^h(\hat{x})\mathbf{W}^h(0)\Big\|_{op} \leq CL.$$

(c) Let $R_{op} \geq 1$ and $s \in \mathbb{N}$ with $s \leq m$. Assume $m \geq CL^6\max\{s\log(m), R_{op}^2\}$ and $s \geq C\log(uL^2/\delta)$. Then, for any $\widehat{\mathbf{W}} \in \mathcal{W}$ satisfying $\|\widehat{\mathbf{W}}\|_{op,\infty} \leq R_{op}$, and for all $\hat{x} \in \mathcal{D}, l \in [L]$ and any diagonal matrices $\widehat{\mathbf{D}}^l \in \mathbb{R}^{m\times m}$ satisfying $\|\widehat{\mathbf{D}}^l\|_0 \leq s$ and $\widehat{\mathbf{D}}^l, \mathbf{D}_0^l(\hat{x}) + \widehat{\mathbf{D}}^l \in [-1,1]^{m\times m}$, there holds

$$\Big\|\prod_{h=l_1}^{l_2}\sqrt{\frac{2}{m}}\big(\mathbf{D}_0^h(\hat{x}) + \widehat{\mathbf{D}}^h\big)\big(\mathbf{W}^h(0) + \widehat{\mathbf{W}}^h\big)\Big\|_{op} \leq CL \text{ for all } 1 \leq l_1 < l_2 \leq L. \tag{4.20}$$

(d) Let $R_{op} \geq 1$. Assume $m \geq C\max\{L^{22}pR_{op}^2\log^3(m), L^3\log^3(uL/\delta)\}$. Then, for any $\mathbf{W} \in \mathcal{W}$ satisfying $\|\mathbf{W} - \mathbf{W}(0)\|_{op,\infty} \leq R_{op}$ and all $\hat{x} \in \mathcal{D}$, there holds

$$\|o^l(\hat{x}) - o_0^l(\hat{x})\|_2 \leq \frac{ClLR_{op}}{\sqrt{m}}. \tag{4.21}$$

Lemma 18 applies only to the finite subset $\mathcal{D}$ of $\mathcal{X}$. In the following lemma, we extend their results to the entire input space $\mathcal{X} = S^{p-1}$.

LEMMA 19. *Let $\delta \in (0,1)$. The following statements hold with probability at least $1-\delta$ over initialization $\mathbf{W}(0)$ for all $l \in [L]$.*

(a) Assume $m \gtrsim pL\log(\frac{m}{\delta})$, there holds $\sup_{x\in\mathcal{X}} \big|\|o_0^l(x)\|_2 - 1\big| \le Cl\sqrt{\frac{pL\log(m/\delta)}{m}}$.
(b) Assume $m \gtrsim pL\log(\frac{m}{\delta})$, there holds $\sup_{x\in\mathcal{X}} \|\mathbf{V}_{L,0}^l(x)\|_{op} \le \frac{CL}{\sqrt{m}}$.
(c) Assume $m \gtrsim pL^3\log(\frac{m}{\delta})$, there holds $\sup_{x\in\mathcal{X}} \|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\|_2 \le 2$.

PROOF. We first prove part (a). Let $\mathcal{D}$ be a $m^{-\frac{1}{2}}(\sqrt{2}c_0)^{-L}$-net of $\mathcal{X}$. We know for any $x \in \mathcal{X}$, there exists $\hat{x} \in \mathcal{D}$ such that $\|x - \hat{x}\|_2 \le m^{-\frac{1}{2}}(\sqrt{2}c_0)^{-L}$. Then, for $l \in [L]$, there holds

$$
\begin{aligned}
&\|o_0^l(x) - o_0^l(\hat{x})\|_2 \\
&= \sqrt{\frac{2}{m}}\big\|\sigma\big(\mathbf{W}^l(0)o_0^{l-1}(x)\big) - \sigma\big(\mathbf{W}^l(0)o_0^{l-1}(\hat{x})\big)\big\|_2 \le \sqrt{\frac{2}{m}}\big\|\mathbf{W}^l(0)(o_0^{l-1}(x) - o_0^{l-1}(\hat{x}))\big\|_2 \\
&\le \sqrt{\frac{2}{m}}\big\|\mathbf{W}^l(0)\big\|_{op}\|o_0^{l-1}(x) - o_0^{l-1}(\hat{x})\|_2 \le \sqrt{2}c_0\|o_0^{l-1}(x) - o_0^{l-1}(\hat{x})\|_2,
\end{aligned}
$$

where we have used 1-Lipschitz continuity of the ReLU and $\|\mathbf{W}^l(0)\|_{op} \le c_0\sqrt{m}$.

Applying the above inequality recursively on l, we know $\|o_0^l(x) - o_0^l(\hat{x})\|_2 \le (\sqrt{2}c_0)^l\|x - \hat{x}\|_2 \le \frac{1}{\sqrt{m}}$. Note $\mathcal{X} = S^{p-1}$ is the unit sphere and $\mathcal{D}$ is a $m^{-\frac{1}{2}}(\sqrt{2}c_0)^{-L}$-net of $\mathcal{X}$. From Corollary 4.2.13 in [Vershynin 2018], we know the covering number of $\mathcal{X}$ satisfy $|\mathcal{D}| \le (3\sqrt{m})^p(\sqrt{2}c_0)^{pL}$. Combining part (a) of Lemma 18 with $u = (3\sqrt{m})^p(\sqrt{2}c_0)^{pL}$ and the condition $m \gtrsim pL\log(\frac{m}{\delta})$, we know with probability at least $1-\delta$, there holds

$$
\big|\|o_0^l(\hat{x})\|_2 - 1\big| \le Cl\sqrt{\frac{pL\log(m/\delta)}{m}} \text{ for all } \hat{x} \in \mathcal{D}.
$$

Combining this with the above inequality $\|o_0^l(x) - o_0^l(\hat{x})\|_2 \le \frac{1}{\sqrt{m}}$, there holds $\big|\|o_0^l(x)\|_2 - 1\big| \le \|o_0^l(x) - o_0^l(\hat{x})\|_2 + \big|\|o_0^l(\hat{x})\|_2 - 1\big| \le Cl\sqrt{\frac{pL\log(m/\delta)}{m}}$ for all $x \in \mathcal{X}$. The first part is proved.

Now, we turn to prove part (b). From Lemma 32 in [Xu and Zhu 2024] we know the cardinality of the set $\{(\mathbf{D}_0^1(x),\ldots,\mathbf{D}_0^L(x)) \in \mathbb{R}^{L\times m\times m} : x\in\mathcal{X}\}$ is at most m^{pL}. Therefore, there exists a subset $\mathcal{D}\subset\mathcal{X}$ with $|\mathcal{D}|\le m^{pL}$ such that

$$\sup_{x\in\mathcal{X}}\|\mathbf{V}_{L,0}^l(x)\|_{op} = \sup_{x\in\mathcal{D}}\|\mathbf{V}_{L,0}^l(x)\|_{op} \text{ for all } l\in[L].$$

Note part (b) of Lemma 18 with $u=m^{pL}, l_2=L, l_1=l+1$ and the condition $m\gtrsim pL\log(\frac{m}{\delta})$ implies that with probability at least $1-\delta$,

$$\begin{aligned}
&\sup_{x\in\mathcal{D}}\|\mathbf{V}_{L,0}^l(x)\|_{op} = \sup_{x\in\mathcal{D}}\Big\|\sqrt{\frac{2}{m}}\mathbf{D}_0^L(x)\mathbf{W}^L(0)\cdots\sqrt{\frac{2}{m}}\mathbf{D}_0^{l+1}(x)\mathbf{W}^{l+1}(0)\sqrt{\frac{2}{m}}\mathbf{D}_0^l(x)\Big\|_{op}\\
&\le \sup_{x\in\mathcal{D}}\big\|\mathbf{D}_0^L(x)\big\|_{op}\Big\|\sqrt{\frac{2}{m}}\mathbf{W}^L(0)\prod_{h=l+1}^{L}\sqrt{\frac{2}{m}}\mathbf{D}_0^h(x)\mathbf{W}^h(0)\Big\|_{op}\Big\|\sqrt{\frac{2}{m}}\mathbf{D}_0^l(x)\Big\|_{op}\le\frac{CL}{\sqrt{m}}.
\end{aligned}$$

Hence,

$$\sup_{x\in\mathcal{X}}\|\mathbf{V}_{L,0}^l(x)\|_{op}\le\frac{CL}{\sqrt{m}},$$

which completes the proof of part (b).

It remains to prove the last part. Note

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\Big\|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}^L(0)}\Big\|_2 &= \sup_{x\in\mathcal{X}}\Big\|\mathbf{V}_{L,0}^L(x)\mathbf{a}\big(o_0^{L-1}(x)\big)^\top\Big\|_2 = \sup_{x\in\mathcal{X}}\Big\|\sqrt{\frac{2}{m}}\mathbf{D}_0^L(x)\mathbf{a}\big(o_0^{L-1}(x)\big)^\top\Big\|_2\\
&\le\sqrt{2}\sup_{x\in\mathcal{X}}\|o_0^{L-1}(x)\|_2\le\sqrt{2}\big(\sup_{x\in\mathcal{X}}\big|\|o_0^{L-1}(x)\|_2-1\big|+1\big)\\
&\le\sqrt{2}\Big(CL\sqrt{\frac{pL\log(m/\delta)}{m}}+1\Big)\le 2,
\end{aligned}$$

where the last second inequality follows from the first part of this lemma, and the last inequality used the condition $m\gtrsim pL^3\log(m/\delta)$. The proof of the lemma is completed. □

The following property shows that K is bounded.

PROPERTY 1. *For any $x,x'\in\mathcal{X}$, there holds $|K(x,x')|\le 1$.*

PROOF. By the definition of $U^l(x)$, we know

$$\mathbb{E}[\sigma^2(U^l(x))] = \frac{1}{2}\mathbb{E}[(U^l(x))^2] = \mathbb{E}[\sigma^2(U^{l-1}(x))].$$

Recursively applying this equality, we have $\mathbb{E}[\sigma^2(U^l(x))] = \mathbb{E}[\sigma^2(U^1(\mathbf{x}))] = \mathbb{E}_{w\sim\mathcal{N}(0,1)}[\sigma^2(w)] = 1/2$. Then, according to Cauchy-Schwarz inequality, for all $x, x' \in \mathcal{X}$ and $l \in [L]$, there holds

$$\left|2\mathbb{E}[\sigma(U^l(x))\sigma(U^l(x'))]\right| \leq \sqrt{2\mathbb{E}[\sigma^2(U^l(x))]}\sqrt{2\mathbb{E}[\sigma^2(U^l(x'))]} = 1. \tag{4.22}$$

Further, according to the definition of $q^l(x, x')$, there holds $|q^l(x, x')| \leq 1$. Then, for $x, x' \in \mathcal{X}$, $K(x, x')$ can be uniformly bounded by

$$\begin{aligned}|K(x, x')| &= \left|2\mathbb{E}[\sigma(U^{L-1}(x))\sigma(U^{L-1}(x'))]\right||q^L(x, x')| \\ &\leq \sqrt{2\mathbb{E}[\sigma^2(U^{L-1}(x))]}\sqrt{2\mathbb{E}[\sigma^2(U^{L-1}(x'))]} = 1.\end{aligned}$$

This completes the proof. □

The work [Du et al. 2019] provided the concentration of the NTK for deep ReLU networks over the training data. i.e., $\sup_{i,j\in[n]} |K^m(x_i, x_j) - K(x_i, x_j)| \to 0$ as $m \to \infty$. [Xu and Zhu 2024] extended their result and showed the concentration uniformly over $\mathcal{X}$, i.e, $\|K^m - K\|_\infty \lesssim C^L m^{-\frac{1}{6}}\sqrt{p}$ assuming exponential scaling of m with L. In the following lemma, we improve their results with relaxed condition on m, which is pivotal for reducing the requirement on m from $C^L\text{Poly}(n, p)$ to $\text{Poly}(n, p, L)$ in Corollaries 1 and 2.

LEMMA 20. *Let $\delta \in (0, 1)$. Assume $m \gtrsim pL^3\log(\frac{m}{\delta})$. With probability at least $1 - L\exp\left(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})\right) - \delta$ over the random choice of $(\mathbf{a}, \mathbf{W}(0))$, there holds*

$$\|K^m - K\|_\infty \lesssim \sqrt{L}m^{-\frac{1}{6}} + \sqrt{pL\log(m)m^{-1}}.$$

PROOF. Note (4.22) and $\left|(\mathbf{V}_{L,0}^L(x)\mathbf{a})^\top\mathbf{V}_{L,0}^L(x')\mathbf{a}\right| = \frac{1}{m}|\mathbf{a}^\top\mathbf{D}_0^L(x)\mathbf{D}_0^L(x')\mathbf{a}| \leq 1$. From the definitions of K^m and K, there holds

$$\begin{aligned}&\|K^m - K\|_\infty \\ &= \sup_{x,x'\in\mathcal{X}}\left|\langle o_0^{L-1}(x), o_0^{L-1}(x')\rangle_2(\mathbf{V}_{L,0}^L(x)\mathbf{a})^\top\mathbf{V}_{L,0}^L(x')\mathbf{a} - 2\mathbb{E}[\sigma(U^{L-1}(x))\sigma(U^{L-1}(x'))]q^L(x, x')\right| \\ &\leq \sup_{x,x'\in\mathcal{X}}\left|\langle o_0^{L-1}(x), o_0^{L-1}(x')\rangle_2 - 2\mathbb{E}[\sigma(U^{L-1}(x))\sigma(U^{L-1}(x'))]\right| \cdot \left|(\mathbf{V}_{L,0}^L(x)\mathbf{a})^\top\mathbf{V}_{L,0}^L(x')\mathbf{a}\right|\end{aligned}$$

$$
\begin{aligned}
&+ \sup_{x,x'\in\mathcal{X}} \big|2\mathbb{E}[\sigma(U^{L-1}(x))\sigma(U^{L-1}(x'))]\big| \cdot \Big|(\mathbf{V}_{L,0}^L(x)\mathbf{a})^\top \mathbf{V}_{L,0}^L(x')\mathbf{a} - tr\big(\mathbf{V}_{L,0}^L(x)^\top \mathbf{V}_{L,0}^L(x')\big)\Big| \\
&+ \sup_{x,x'\in\mathcal{X}} \big|2\mathbb{E}[\sigma(U^{L-1}(x))\sigma(U^{L-1}(x'))]\big| \cdot \Big|tr\big(\mathbf{V}_{L,0}^L(x)^\top \mathbf{V}_{L,0}^L(x')\big) - q^L(x,x')\Big| \\
&\le \sup_{x,x'\in\mathcal{X}} \big|\langle o_0^{L-1}(x), o_0^{L-1}(x')\rangle_2 - 2\mathbb{E}[\sigma(U^{L-1}(x))\sigma(U^{L-1}(x'))]\big| \\
&+ \sup_{x,x'\in\mathcal{X}} \Big|(\mathbf{V}_{L,0}^L(x)\mathbf{a})^\top \mathbf{V}_{L,0}^L(x')\mathbf{a} - tr\big(\mathbf{V}_{L,0}^L(x)^\top \mathbf{V}_{L,0}^L(x')\big)\Big| \\
&+ \sup_{x,x'\in\mathcal{X}} \Big|tr\big(\mathbf{V}_{L,0}^L(x)^\top \mathbf{V}_{L,0}^L(x')\big) - q^L(x,x')\Big| \\
&=: \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3,
\end{aligned}
$$

The estimates of the above three terms $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ are given as follows.

Estimate of $\mathcal{E}_1$: The estimate of $\mathcal{E}_1$ follows the same proof steps as in Lemma 6 in [Xu and Zhu 2024]. According to Lemma 6 in [Xu and Zhu 2024], one can get that $\mathcal{E}_1 \lesssim LC^L m^{-\frac{1}{3}}$. We improve this estimate from $LC^L m^{-\frac{1}{3}}$ to $Lm^{-\frac{1}{3}}$ by using more finer estimates of initialization terms. Specifically, instead of using their estimate $\sup_x \|o_0^l(x)\|_2 \le c_0^l$ in Lemma 30, we apply the tight estimate $\sup_x \|o_0^l(x)\|_2 \le \sup_x \big|\|o_0^l(x)\|_2 - 1\big| + 1 \le C$ according to part (a) of Lemma 19 and the condition $m \gtrsim pL^3 \log(m/\delta)$. In addition, we set V_0 to be a $c_0^{-L} m^{-2}$-net of the S^{p-1} rather than a m^{-2}-net. Then, following the same steps of the proof of Lemma 6 in [Xu and Zhu 2024], with probability at least $1 - L\exp(O(pL\log(m)) - \Omega(m^{\frac{1}{3}}))$ over initialization $\mathbf{W}(0)$, there holds

$$\mathcal{E}_1 \lesssim Lm^{-\frac{1}{3}}.$$

Estimate of $\mathcal{E}_2$: Similar to the proof of the estimate of $\mathcal{E}_1$, by using more finer estimates $\sup_x \|o_0^l(x)\|_2 \le C, \sup_{x\in\mathcal{X}} \|\mathbf{V}_{L,0}^L(x)\|_{op} \le m^{-\frac{1}{2}}$, following the same proof steps of Lemma 7 in [Xu and Zhu 2024], we can show that

$$\mathcal{E}_2 \lesssim m^{-\frac{1}{3}}.$$

Estimate of $\mathcal{E}_3$: Similar to the above arguments, we use the estimates $\sup_x \|o_0^l(x)\|_2 \leq C$ and $\sup_{x\in\mathcal{X}} \|\mathbf{V}_{L,0}^L(x)\|_{op} \leq m^{-\frac{1}{2}}$ to improve the proof of Lemma 8 in [Xu and Zhu 2024] and get

$$\mathcal{E}_3 \lesssim \frac{\sqrt{L}}{m^{\frac{1}{6}}} + \sqrt{\frac{pL\log(m)}{m}}.$$

Combining the above estimates of $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ completes the proof of this lemma. □

4.4.2 Useful Lemmas

In this section, we present some useful lemmas for proving main results of both GD and SGD with deep ReLU networks.

Recall that $\mathbf{D}_0^l(x)$, $o_0^l(x)$ and $\mathbf{V}_{L,0}^l(x)$ are defined as (4.15), (4.16) and (4.17) with $\mathbf{W} = \mathbf{W}(0)$ for all $l \in [L]$, and $\mathcal{B}_R(\widetilde{\mathbf{W}}) = \big\{\mathbf{W} \in \mathcal{W} : \|\mathbf{W} - \widetilde{\mathbf{W}}\|_{op,\infty} = \max_{l\in[L]} \|\mathbf{W}^l - \widetilde{\mathbf{W}}^l\|_{op} \leq R\big\}$.

LEMMA 21. *Let $\delta \in (0,1)$. Assume $m \gtrsim L^{22}p^3R_{op}^2\log^3(m/\delta)$ and $R_{op} \geq 1$. Then, with probability at least $1-\delta$ over initialization $\mathbf{W}(0)$, the following statement holds for any $\mathbf{W} \in \mathcal{B}_{R_{op}}\big(\mathbf{W}(0)\big)$,*

$$\sup_{x\in\mathcal{X}} \|o^l(x) - o_0^l(x)\|_2 \lesssim \frac{lLR_{op}}{\sqrt{m}}.$$

PROOF. Let $\mathcal{D}$ be a $\frac{1}{C^L\sqrt{m}}$-net of $\mathcal{X}$. We know for any $x \in \mathcal{X}$, there exists $\hat{x} \in \mathcal{D}$ such that $\|x - \hat{x}\|_2 \leq \frac{1}{C^L\sqrt{m}}$. Note the condition for m implies $R_{op} \leq \sqrt{m}$. Then, similar to the proof of part (a) of Lemma 19, we know $\|o^l(x) - o^l(\hat{x})\|_2 \leq (\sqrt{2}+\sqrt{2}c_0)^l\|x-\hat{x}\|_2 \leq \frac{1}{\sqrt{m}}$ and $\|o_0^l(x) - o_0^l(\hat{x})\|_2 \leq \frac{1}{\sqrt{m}}$. Note $\mathcal{X} = S^{p-1}$ is the unit sphere. From Corollary 4.2.13 in [Vershynin 2018], it holds that $|\mathcal{D}| \leq (3\sqrt{m})^p C^{pL}$. Then, applying part (d) of Lemma 18 with $u = (3\sqrt{m})^p C^{pL}$, there holds

$$\begin{aligned}\|o^l(x) - o_0^l(x)\|_2 &\leq \|o^l(x) - o^l(\hat{x})\|_2 + \|o^l(\hat{x}) - o_0^l(\hat{x})\|_2 + \|o_0^l(\hat{x}) - o_0^l(x)\|_2 \\ &\lesssim \frac{1}{\sqrt{m}} + \frac{lLR_{op}}{\sqrt{m}} + \frac{1}{\sqrt{m}} \lesssim \frac{lLR_{op}}{\sqrt{m}}.\end{aligned}$$

This completes the proof. □

LEMMA 22. *Let* $\delta \in (0,1)$. *Assume* $m \gtrsim L^{22}p^3R_{op}^2\log^3(m/\delta)$ *and* $R_{op} \geq 1$. *For any* $\mathbf{W} \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$, *with probability at least* $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ *over initialization* $(\mathbf{a}, \mathbf{W}(0))$, *for any* $l \in [L]$, *there holds*

$$\sup_{x\in\mathcal{X}} \|\mathbf{D}^l(x) - \mathbf{D}_0^l(x)\|_0 \leq \big(LmR_{op}\big)^{\frac{2}{3}}.$$

PROOF. Let $R' > 0$ which will be chosen later. For $x \in \mathcal{X}$ and $l \in [L]$, define the diagonal matrix

$$\mathbf{E}^l(x) = diag\big\{\mathbb{I}\big\{|\langle \mathbf{w}_r^l(0), o_0^{l-1}(x)\rangle_2| \leq R'\big\}\big\}_{r=1}^m \in \{0,1\}^{m\times m}.$$

Note $\sup_x \|\mathbf{D}^l(x) - \mathbf{D}_0^l(x)\|_0 \leq \sup_x \|(\mathbf{D}^l(x) - \mathbf{D}_0^l(x))(\mathbf{I} - \mathbf{E}^l(x))\|_0 + \sup_x \|\mathbf{E}^l(x)\|_0$. We will estimate $\sup_x \|(\mathbf{D}^l(x) - \mathbf{D}_0^l(x))(\mathbf{I} - \mathbf{E}^l(x))\|_0$ and $\sup_x \|\mathbf{E}^l(x)\|_0$ separately in the following proof.

Estimate of $\sup_x \big\|(\mathbf{D}^l(x) - \mathbf{D}_0^l(x))(\mathbf{I} - \mathbf{E}^l(x))\big\|_0$**:** From the definition, if the absolute value of (r,r)-th entry of the diagonal matrix $(\mathbf{D}^l(x) - \mathbf{D}_0^l(x))(\mathbf{I} - \mathbf{E}^l(x))$ is 1, then $|\langle \mathbf{w}_r^l(0), o_0^{l-1}(x)\rangle_2| > R'$ and $\mathbb{I}\{\langle \mathbf{w}_r^l, o^{l-1}(x)\rangle_2 \geq 0\} \neq \mathbb{I}\{\langle \mathbf{w}_r^l(0), o_0^{l-1}(x)\rangle_2 \geq 0\}$. Then, there holds

$$|\langle \mathbf{w}_r^l, o^{l-1}(x)\rangle_2 - \langle \mathbf{w}_r^l(0), o_0^{l-1}(x)\rangle_2|^2 \geq |\langle \mathbf{w}_r^l(0), o_0^{l-1}(x)\rangle_2|^2 > (R')^2.$$

Therefore, we have

$$\begin{aligned}
&\sup_x \big\|(\mathbf{D}^l(x) - \mathbf{D}_0^l(x))(\mathbf{I} - \mathbf{E}^l(x))\big\|_0 \\
&\leq \frac{1}{(R')^2}\sup_x \sum_{r=1}^m \big(\langle \mathbf{w}_r^l, o^{l-1}(x)\rangle_2 - \langle \mathbf{w}_r^l(0), o_0^{l-1}(x)\rangle_2\big)^2 \\
&= \frac{1}{(R')^2}\sup_x \big\|\mathbf{W}^l o^{l-1}(x) - \mathbf{W}^l(0) o_0^{l-1}(x)\big\|_2^2 \\
&\leq \frac{1}{(R')^2}\sup_x \Big(\big\|\mathbf{W}^l - \mathbf{W}^l(0)\big\|_{op}\big\|o^{l-1}(x) - o_0^{l-1}(x) + o_0^{l-1}(x)\big\|_2 + \big\|\mathbf{W}^l(0)\big\|_{op}\big\|o^{l-1}(x) - o_0^{l-1}(x)\big\|_2\Big)^2 \\
&\leq \frac{1}{(R')^2}\sup_x \Big(R_{op}\big(\big\|o^{l-1}(x) - o_0^{l-1}(x)\big\|_2 + C\big) + c_0\sqrt{m}\big\|o^{l-1}(x) - o_0^{l-1}(x)\big\|_2\Big)^2,
\end{aligned}$$

where in the last inequality we have used $\sup_x \|o_0^{l-1}(x)\|_2 \leq C$ implied by part (a) of Lemma 19 and the condition for m, and $\|\mathbf{W}^l(0)\|_{op} \leq c_0\sqrt{m}$.

Combining the above inequality with Lemma 21 and noting $m \gtrsim L^{22}p^3R_{op}^2\log^3(m/\delta)$, we get

$$\sup_x \big\|(\mathbf{D}^l(x) - \mathbf{D}_0^l(x))(\mathbf{I} - \mathbf{E}^l(x))\big\|_0 \lesssim \frac{1}{(R')^2}\Big(R_{op}\Big(\frac{L^2R_{op}}{\sqrt{m}} + C\Big) + L^2R_{op}\Big) \lesssim \frac{L^2R_{op}^2}{(R')^2}.$$

Estimates of $\sup_x \big\|\mathbf{E}^l(x)\big\|_0$: The proof is similar to that of Lemma 11 in [Xu and Zhu 2024], we give the proof here for the sake of completeness.

Denoting the function class $\mathcal{F} = \{\mathbb{I}\{|\langle\,\cdot\,, o^{l-1}(x)\rangle_2| \leq R'\} : x \in \mathcal{X}\}$, there holds

$$\sup_x \frac{1}{m}\|\mathbf{E}^l(x)\|_0 = \sup_x \frac{1}{m}\sum_{r=1}^m \mathbb{I}\big\{\big|\langle \mathbf{w}_r^l(0), o_0^{l-1}(x)\rangle_2\big| \leq R'\big\} = \sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{r=1}^m f(\mathbf{w}_r^l(0)).$$

To control the right hand side of the above equality, we need to estimate the VC-dimension of $\mathcal{F}$. We first fixed $(\mathbf{W}^1(0), \ldots, \mathbf{W}^{l-1}(0))$. Denote $\mathcal{D}^{l-1} = \{(\mathbf{D}_0^1(x), \ldots, \mathbf{D}_0^{l-1}(x)) : x \in \mathcal{X}\} \subset \mathbb{R}^{(l-1)\times m\times m}$. From Lemma 32 in [Xu and Zhu 2024] we know the cardinality of $\mathcal{D}^{l-1}$ is less than $m^{p(l-1)}$, i.e., $|\mathcal{D}^{l-1}| \leq m^{p(l-1)}$. Then, there exists a disjoint partition of $\mathcal{X}$ such that $\mathcal{X} = \bigcup_{j\in[|\mathcal{D}^{l-1}|]} U_j$, where $U_i \cap U_j = \varnothing$ for $i \neq j$ and the tuple $(\mathbf{D}_0^1(x), \ldots, \mathbf{D}_0^{l-1}(x)) \in \mathbb{R}^{(l-1)\times m\times m}$ is a fixed matrix sequence on each U_j. Therefore, $o_0^{l-1}(x) = \sqrt{\frac{2}{m}}\mathbf{D}_0^{l-1}(x)\mathbf{W}^{l-1}(0)\cdots\sqrt{\frac{2}{m}}\mathbf{D}_0^1(x)\mathbf{W}^1(0)x$ lies in a p-dimensional subspace of $\mathbb{R}^m$ on each U_j. Let V_j and V be the VC-dimension of the classes $\mathcal{F}_j = \{\mathbb{I}\{|\langle\,\cdot\,, o^{l-1}(x)\rangle_2| \leq R'\} : x \in U_j\}$ and $\mathcal{F}$, respectively. By Theorem 9.5 in [Györfi et al. 2006], the VC-dimension of the class of indicators of half spaces in $\mathbb{R}^p$ is $p+1$. Further, note that $o_0^{l-1}(x)$ lies in a p-dimensional subspace of $\mathbb{R}^m$ on each U_j and the indicator function $\mathbb{I}\{|\langle \mathbf{w}^l, o^{l-1}(x)\rangle_2| \leq R'\}$ can be written as the multiplication of two indicators of half space, i.e., $\mathbb{I}\{|\langle \mathbf{w}^l, o^{l-1}(x)\rangle_2| \leq R'\} = \mathbb{I}\{\langle \mathbf{w}^l, o^{l-1}(x)\rangle_2 \leq R'\}\mathbb{I}\{\langle \mathbf{w}^l, o^{l-1}(x)\rangle_2 \geq -R'\}$. Then, from Lemma 3.2.3 in [Blumer et al. 1989] with $s = 2$ we know $V_j \leq 10(p+1)$ for any j. By further applying Lemma 23 in [Xu and Zhu 2024] with $N = |\mathcal{D}^{l-1}| \leq m^{p(l-1)}$, there holds $V \lesssim \max(p\log(p), \log(|\mathcal{D}^{l-1}|)) \lesssim pl\log(m)$.

Now, we turn to control $\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{r=1}^m f(\mathbf{w}_r^l(0))$, which can be regarded as a fucntion on $(\mathbf{w}_1^l(0),\ldots,\mathbf{w}_m^l(0))$. One can check that the value of this function can change by at most $\frac{1}{m}$ under an arbitrary change of the r-th coordinate. Then, by McDiarmid's inequality, we know with probability at least $1-\exp(-2m^{\frac{1}{3}})$ over $\mathbf{W}^l(0)$, there holds

$$\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{r=1}^m f(\mathbf{w}_r^l(0)) \le m^{-\frac{1}{3}} + \mathbb{E}_{\mathbf{W}^l(0)}\Big[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{r=1}^m f(\mathbf{w}_r^l(0))\Big].$$

Now, we estimate the right hand side of the above inequality. There holds

$$\begin{aligned}
&\mathbb{E}_{\mathbf{W}^l(0)}\Big[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{r=1}^m f(\mathbf{w}_r^l(0))\Big]\\
&\le \mathbb{E}_{\mathbf{W}^l(0)}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{m}\sum_{r=1}^m f(\mathbf{w}_r^l(0)) - \mathbb{E}[f(\mathbf{w}^l(0))]\Big|\Big] + \sup_{f\in\mathcal{F}}\mathbb{E}[f(\mathbf{w}^l(0))]\\
&\le \sqrt{\frac{V}{m}} + \sup_{f\in\mathcal{F}}\mathbb{E}[f(\mathbf{w}_r^l(0))] \le \sqrt{\frac{pl\log(m)}{m}} + \sup_x \mathbb{E}\big[\mathbb{I}\big\{|\langle \mathbf{w}^l(0), o_0^{l-1}(x)\rangle_2| \le R'\big\}\big]\\
&\le \sqrt{\frac{pl\log(m)}{m}} + \sup_x \int_{-R'/\|o_0^{l-1}(x)\|_2}^{R'/\|o_0^{l-1}(x)\|_2}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \le \sqrt{\frac{pl\log(m)}{m}} + \sup_x \frac{\sqrt{2}R'}{\sqrt{\pi}\|o_0^{l-1}(x)\|_2},
\end{aligned}$$

where the second inequality is according to Theorem 8.3.23 in [Vershynin 2018], the third inequality follows from $V \le pl\log(m)$, in the last second inequality we have used $\langle \mathbf{w}^l(0), o_0^{l-1}(x)/\|o_0^{l-1}(x)\|_2\rangle_2 \sim \mathcal{N}(0,1)$, and in the last inequality we have used $e^{-\frac{t^2}{2}} \le 1$. It remains to estimate $\sup_x \frac{\sqrt{2}R'}{\sqrt{\pi}\|o_0^l(x)\|_2}$. For the case $l=1$, there holds $\|o^0(x)\|_2 = \|x\|_2 = 1$. For the case $l \ge 2$, for any $x\in\mathcal{X}$, from part (a) of Lemma 19 we have

$$\|o_0^{l-1}(x)\|_2 = 1-\big(1-\|o_0^{l-1}(x)\|_2\big) \ge 1-\big|\|o_0^{l-1}(x)\|_2-1\big| \ge 1-CL\sqrt{\frac{pL\log(m/\delta)}{m}} \ge \frac{1}{2},$$

where the last inequality follows from the condition $m \gtrsim L^{22}p^3R_{op}^2\log^3(m/\delta)$.

Combining the above estimates we obtain

$$\sup_x \frac{1}{m}\|\mathbf{E}^l(x)\|_0 = \sup_x \frac{1}{m}\sum_{r=1}^m \mathbb{I}\big\{\big|\langle \mathbf{w}_r^l(0), o_0^l(x)\rangle_2\big| \le R'\big\} \lesssim \frac{1}{m^{\frac{1}{3}}} + \sqrt{\frac{pl\log(m)}{m}} + R'.$$

Further, combining the estimates of $\sup_x \left\|(\mathbf{D}^l(x)-\mathbf{D}_0^l(x))(\mathbf{I}-\mathbf{E}^l(x))\right\|_0$ and $\sup_x \|\mathbf{E}^l(x)\|_0$, there holds

$$\sup_x \left\|\mathbf{D}^l(x)-\mathbf{D}_0^l(x)\right\|_0 \lesssim \frac{L^2R_{op}^2}{(R')^2}+R'm+2m^{\frac{2}{3}}+\sqrt{mpl\log(m)}.$$

Setting $R' \asymp (LR_{op})^{\frac{2}{3}}m^{-\frac{1}{3}}$. Noting that $m \gtrsim L^{22}p^3R_{op}^2\log^3(m/\delta)$ and $R_{op}\geq 1$, we have

$$\sup_x \left\|\mathbf{D}^l(x)-\mathbf{D}_0^l(x)\right\|_0 \lesssim \big(LmR_{op}\big)^{\frac{2}{3}}+2m^{\frac{2}{3}}+\sqrt{mpl\log(m)} \lesssim \big(LmR_{op}\big)^{\frac{2}{3}}.$$

The proof of the lemma is completed. □

Recall that

$$\mathbf{V}_{L,0}^l(x)=\sqrt{\frac{2}{m}}\mathbf{D}_0^L(x)\mathbf{W}^L(0)\cdots\sqrt{\frac{2}{m}}\mathbf{D}_0^{l+1}(x)\mathbf{W}^{l+1}(0)\sqrt{\frac{2}{m}}\mathbf{D}_0^l(x).$$

For any $l\in[L]$, let $\widehat{\mathbf{W}}^l$ and the diagonal matrix $\widehat{\mathbf{D}}^l$ be the matrices with the same size of $\mathbf{W}^l(0)$ and $\mathbf{D}_0^l(x)$, respectively. Define, for $k\in[L-1]$ and $l<k$,

$$\begin{aligned}\widehat{\mathbf{V}}_k^l(x)=&\sqrt{\frac{2}{m}}\big(\mathbf{D}_0^k(x)+\widehat{\mathbf{D}}^k\big)\big(\mathbf{W}^k(0)+\widehat{\mathbf{W}}^k\big)\cdots\\&\times\sqrt{\frac{2}{m}}\big(\mathbf{D}_0^{l+1}(x)+\widehat{\mathbf{D}}^{l+1}\big)\big(\mathbf{W}^{l+1}(0)+\widehat{\mathbf{W}}^{l+1}\big)\sqrt{\frac{2}{m}}\big(\mathbf{D}_0^l(x)+\widehat{\mathbf{D}}^l\big)\end{aligned}\tag{4.23}$$

and $\widehat{\mathbf{V}}_l^l(x)=\sqrt{\frac{2}{m}}\big(\mathbf{D}_0^l(x)+\widehat{\mathbf{D}}^l\big)$ for all $l\in[L]$.

LEMMA 23. *Let $\delta\in(0,1)$ and $\widehat{\mathbf{V}}_k^l(x)$ with $k\in[L]$ and $l<k$ be the matrix defined in* (4.23). *Let $R_{op}\geq 1$ and $s\in[m]$. Assume $m\geq CL^6\max\{s\log(m),R_{op}^2\}$ and $s\geq CpL\log(m/\delta)$. Then, with probability at least $1-\delta$ over initialization $\mathbf{W}(0)$, for any matrices satisfying $\|\widehat{\mathbf{W}}\|_{op,\infty}\leq R_{op}$ and diagonal matrices satisfying $\|\widehat{\mathbf{D}}^l\|_0\leq s$ and $\widehat{\mathbf{D}}^l,\mathbf{D}_0^l(x)+\widehat{\mathbf{D}}^l\in[-1,1]^{m\times m}$ for all $l\in[L]$ and $x\in\mathcal{X}$, there holds*

$$\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{V}}_k^l(x)\|_{op}\leq\frac{CL}{\sqrt{m}}.$$

PROOF. Similar to the proof of part (b) of Lemma 19, we know there exists a finite subset $\mathcal{D} \subset \mathcal{X}$ with $|\mathcal{D}| \leq m^{pL}$ such that

$$\sup_{x\in\mathcal{X}} \|\widehat{\mathbf{V}}_k^l(x)\|_{op} = \sup_{x\in\mathcal{D}} \|\widehat{\mathbf{V}}_k^l(x)\|_{op} \text{ for all } 1 \leq l < k \leq L.$$

Then, part (c) of Lemma 18 with $u = m^{pL}$ implies that

$$\begin{aligned}\sup_{x\in\mathcal{D}} \|\widehat{\mathbf{V}}_k^l(x)\|_{op} &\leq \Big\| \prod_{h=l+1}^{k} \sqrt{\frac{2}{m}}\big(\mathbf{D}_0^h(x) + \widehat{\mathbf{D}}^h\big)\big(\mathbf{W}^h(0) + \widehat{\mathbf{W}}^h\big)\Big\|_{op} \Big\|\sqrt{\frac{2}{m}}\big(\mathbf{D}_0^l(x) + \widehat{\mathbf{D}}^l\big)\Big\|_{op} \\ &\leq \frac{CL}{\sqrt{m}}.\end{aligned}$$

This completes the proof. □

LEMMA 24. *Let $\delta \in (0,1)$ and $\widehat{\mathbf{V}}_L^l(x)$ with $l \in [L]$ be the matrix defined in* (4.23)*. Let $R_{op} \geq 1$ and $s \in [m]$. Assume $\|\widehat{\mathbf{W}}\|_{op,\infty} = \max_{l\in[L]} \|\widehat{\mathbf{W}}^l\|_{op} \leq R_{op}$ and $\sup_{l\in[L]} \|\widehat{\mathbf{D}}^l\|_0 \leq s$ and $\widehat{\mathbf{D}}^l, \mathbf{D}_0^l(x) + \widehat{\mathbf{D}}^l \in [-1,1]^{m\times m}$ for all $x \in \mathcal{X}$. Suppose $m \geq CL^6 \max\{s\log(m), R_{op}^2\}$ and $s \geq CpL\log(m/\delta)$. Then, with probability at least $1-\delta$ over the random choice of the initialization $\mathbf{W}(0)$, there holds for all $l \in [L]$*

$$\sup_{x\in\mathcal{X}} \big\|\mathbf{a}^\top\big(\widehat{\mathbf{V}}_L^l(x) - \mathbf{V}_{L,0}^l(x)\big)\big\|_2 \lesssim \frac{L\big(\sqrt{s} + R_{op}\big)}{\sqrt{m}}.$$

PROOF. For the case $l = L$, according to definitions of $\widehat{\mathbf{V}}_L^l(x)$ and $\mathbf{V}_{L,0}^l(x)$ we know

$$\|\mathbf{a}^\top(\widehat{\mathbf{V}}_L^L(x) - \mathbf{V}_{L,0}^L(x))\|_2 = \sqrt{\frac{2}{m}}\|\mathbf{a}^\top\widehat{\mathbf{D}}^L\|_2 \lesssim \frac{\sqrt{s}}{\sqrt{m}}, \tag{4.24}$$

where the inequality is due to $a_r \in \{-1,1\}$ for $r \in [m]$ and $\|\widehat{\mathbf{D}}^L\|_0 \leq s$. This completes the proof of the case $l = L$.

For the case $l \in [L-1]$, noting that $\mathbf{a}^\top\sqrt{\frac{2}{m}}\mathbf{D}_0^L(x)\mathbf{W}^L(0) = 0$ (see Lemma 16), we know

$$\begin{aligned}&\mathbf{a}^\top\big(\widehat{\mathbf{V}}_L^l(x) - \mathbf{V}_{L,0}^l(x)\big) \\ &= \mathbf{a}^\top\Big(\sqrt{\frac{2}{m}}\big(\mathbf{D}_0^L(x) + \widehat{\mathbf{D}}^L(x)\big)\big(\mathbf{W}^L(0) + \widehat{\mathbf{W}}^L\big)\widehat{\mathbf{V}}_{L-1}^l(x) - \sqrt{\frac{2}{m}}\mathbf{D}_0^L(x)\mathbf{W}^L(0)\mathbf{V}_{L-1,0}^l(x)\Big)\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{a}^\top \sqrt{\frac{2}{m}} \mathbf{D}_0^L(x) \mathbf{W}^L(0) \big(\widehat{\mathbf{V}}_{L-1}^l(x) - \mathbf{V}_{L-1,0}^l(x)\big) + \mathbf{a}^\top \sqrt{\frac{2}{m}} \widehat{\mathbf{D}}^L(x) \big(\mathbf{W}^L(0) + \widehat{\mathbf{W}}^L\big) \widehat{\mathbf{V}}_{L-1}^l(x) \\
&\quad + \mathbf{a}^\top \sqrt{\frac{2}{m}} \mathbf{D}_0^L(x) \widehat{\mathbf{W}}^L \widehat{\mathbf{V}}_{L-1}^l(x) \\
&= \mathbf{a}^\top \sqrt{\frac{2}{m}} \widehat{\mathbf{D}}^L(x) \big(\mathbf{W}^L(0) + \widehat{\mathbf{W}}^L\big) \widehat{\mathbf{V}}_{L-1}^l(x) + \mathbf{a}^\top \sqrt{\frac{2}{m}} \mathbf{D}_0^L(x) \widehat{\mathbf{W}}^L \widehat{\mathbf{V}}_{L-1}^l(x).
\end{aligned}
$$

According to Lemma 23, we know $\sup_{x \in \mathcal{X}} \|\widehat{\mathbf{V}}_L^l(x)\|_{op} \leq \frac{CL}{\sqrt{m}}$. Then, there holds

$$
\begin{aligned}
&\big\|\mathbf{a}^\top \big(\widehat{\mathbf{V}}_L^l(x) - \mathbf{V}_{L,0}^l(x)\big)\big\|_2 \\
&\leq \sqrt{\frac{2}{m}} \big\|\mathbf{a}^\top \widehat{\mathbf{D}}^L(x)\big\|_2 \big\|\mathbf{W}^L(0) + \widehat{\mathbf{W}}^L\big\|_{op} \big\|\widehat{\mathbf{V}}_{L-1}^l(x)\big\|_{op} + \sqrt{\frac{2}{m}} \|\mathbf{a}\|_2 \big\|\mathbf{D}_0^L(x)\big\|_{op} \big\|\widehat{\mathbf{W}}^L\big\|_{op} \big\|\widehat{\mathbf{V}}_{L-1}^l(x)\big\|_{op} \\
&\leq \big\|\mathbf{a}^\top \widehat{\mathbf{D}}^L(x)\big\|_2 \frac{\sqrt{2}(c_0\sqrt{m} + R_{op})}{\sqrt{m}} \frac{CL}{\sqrt{m}} + \sqrt{2} R_{op} \frac{CL}{\sqrt{m}} \\
&\lesssim \frac{L\big(\sqrt{s} + R_{op}\big)}{\sqrt{m}},
\end{aligned}
$$

where the second inequality used the assumption $\|\widehat{\mathbf{W}}\|_{op,\infty} \leq R_{op}$ and $\|\mathbf{W}^L(0)\|_{op} \leq c_0\sqrt{m}$, and the last inequality used (4.24) and $R_{op} \leq \sqrt{m}$ by noting $m \geq CL^6 R_{op}^2$. This completes the proof of the lemma. □

LEMMA 25 (Claim 11.2 and Proposition 11.3 in [Allen-Zhu et al. 2019b]). *For any $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$ and $x \in \mathcal{X}$. There exist a series of diagonal matrices $\{(\mathbf{D}'')^l \in \mathbb{R}^{m \times m}\}_{l \in [L]}$ with entries in $[-1, 1]$ such that for any $l \in [L]$, there holds*

(a) $o^l(x) - \tilde{o}^l(x) = \sum_{h=1}^l \Big[\prod_{j=h+1}^l \sqrt{\frac{2}{m}} \big(\widetilde{\mathbf{D}}^j(x) + (\mathbf{D}'')^j\big) \widetilde{\mathbf{W}}^j\Big] \sqrt{\frac{2}{m}} \big(\widetilde{\mathbf{D}}^h(x) + (\mathbf{D}'')^h\big) \big(\mathbf{W}^h - \widetilde{\mathbf{W}}^h\big) o^{h-1}(x)$.

(b) $\big\|(\mathbf{D}'')^l\big\|_0 \leq \big\|\mathbf{D}^l(x) - \widetilde{\mathbf{D}}^l(x)\big\|_0$ *and* $\widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l$ *has entries in* $[0, 1]$.

The following lemma shows that the neural network is almost linear in terms of its weights and the loss is locally almost smooth near the initialization.

LEMMA 26. *Assume $R_{op} \geq 1$ and $m \geq CL^{22} p^3 R_{op}^2 \log^3(m/\delta)$. For any $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$, with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ over initialization $(\mathbf{a}, \mathbf{W}(0))$,*

for any $z = (x, y) \in \mathcal{Z}$, there holds

$$\Big|f_{\widetilde{\mathbf{W}}}(x) - f_{\mathbf{W}}(x) - \Big\langle \frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W} \Big\rangle_2\Big| \lesssim L^{\frac{7}{3}} \|\widetilde{\mathbf{W}} - \mathbf{W}\|_{op,\infty} R_{op}^{\frac{1}{3}} m^{-\frac{1}{6}}, \tag{4.25}$$

$$l(\widetilde{\mathbf{W}}; z) - \ell(\mathbf{W}; z) \geq \Big\langle \frac{\partial \ell(\mathbf{W}; z)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W} \Big\rangle_2 - |f_{\mathbf{W}}(x) - y| \cdot \epsilon, \tag{4.26}$$

with $\epsilon \lesssim L^{\frac{7}{3}} \|\widetilde{\mathbf{W}} - \mathbf{W}\|_{op,\infty} R_{op}^{\frac{1}{3}} m^{-\frac{1}{6}}$, and

$$\Big\| \frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}^l} - \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^l(0)} \Big\|_2 \lesssim L^{\frac{4}{3}} R_{op}^{\frac{1}{3}} m^{-\frac{1}{6}}. \tag{4.27}$$

PROOF. We first prove that the neural network f is almost linear in terms of its weights near the initialization. From the definition of f, we know

$$\begin{aligned}
&\Big|f_{\widetilde{\mathbf{W}}}(x) - f_{\mathbf{W}}(x) - \Big\langle \frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W} \Big\rangle_2\Big| \\
&= \Big|\mathbf{a}^\top \tilde{o}^L(x) - \mathbf{a}^\top o^L(x) - \sum_{l=1}^{L} \mathbf{a}^\top \Big[\prod_{h=l+1}^{L} \sqrt{\frac{2}{m}} \mathbf{D}^h(x) \mathbf{W}^h \Big] \sqrt{\frac{2}{m}} \mathbf{D}^l(x) \big(\widetilde{\mathbf{W}}^l - \mathbf{W}^l\big) o^{l-1}(x) \Big|,
\end{aligned}$$

where we used the conventional notation $\prod_{L+1}^{L} = \mathbf{I}$.

Lemma 25 with $l = L$ implies there exist a series of diagonal matrices $\{(\mathbf{D}'')^l \in \mathbb{R}^{m\times m}\}_{l\in[L]}$ with entries in $[-1, 1]$ such that

$$\begin{aligned}
o^L(x) - \tilde{o}^L(x) = &\sum_{l=1}^{L} \Big[\prod_{h=l+1}^{L} \sqrt{\frac{2}{m}} \big(\widetilde{\mathbf{D}}^h(x) + (\mathbf{D}'')^h\big) \widetilde{\mathbf{W}}^h \Big] \sqrt{\frac{2}{m}} \big(\widetilde{\mathbf{D}}^l(x) \\
&+ (\mathbf{D}'')^l\big) \big(\mathbf{W}^l - \widetilde{\mathbf{W}}^l\big) o^{l-1}(x).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\Big|f_{\widetilde{\mathbf{W}}}(x) - f_{\mathbf{W}}(x) - \Big\langle \frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W} \Big\rangle_2\Big| \\
&\leq \sum_{l=1}^{L} \Big|\mathbf{a}^\top \Big[\prod_{h=l+1}^{L} \sqrt{\frac{2}{m}} \big(\widetilde{\mathbf{D}}^h(x) + (\mathbf{D}'')^h\big) \widetilde{\mathbf{W}}^h \Big] \sqrt{\frac{2}{m}} \big(\widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l\big) \big(\widetilde{\mathbf{W}}^l - \mathbf{W}^l\big) o^{l-1}(x) \\
&\qquad - \mathbf{a}^\top \Big[\prod_{h=l+1}^{L} \sqrt{\frac{2}{m}} \mathbf{D}^h(x) \mathbf{W}^h \Big] \sqrt{\frac{2}{m}} \mathbf{D}^l(x) \big(\widetilde{\mathbf{W}}^l - \mathbf{W}^l\big) o^{l-1}(x) \Big|
\end{aligned}$$

$$=: \sum_{l=1}^{L} \Big| U_l^L(x)\big(\widetilde{\mathbf{W}}^l - \mathbf{W}^l\big) o^{l-1}(x) \Big| \leq \sum_{l=1}^{L} \|U_l^L(x)\|_2 \big\|\widetilde{\mathbf{W}}^l - \mathbf{W}^l\big\|_{op} \|o^{l-1}(x)\|_2, \quad (4.28)$$

where $U_l^L(x) = \mathbf{a}^\top \big[\prod_{h=l+1}^{L} \sqrt{\frac{2}{m}}\big(\widetilde{\mathbf{D}}^h(x) + (\mathbf{D}'')^h\big)\widetilde{\mathbf{W}}^h\big] \sqrt{\frac{2}{m}}\big(\widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l\big) - \mathbf{a}^\top \big[\prod_{h=l+1}^{L} \sqrt{\frac{2}{m}}\mathbf{D}^h(x)\mathbf{W}^h\big]\sqrt{\frac{2}{m}}\mathbf{D}^l(x)$.

We first consider estimating the term $\|U_l^L(x)\|_2$. We begin by showing that $\widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l$, $\widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l - \mathbf{D}_0^l(x) \in [-1,1]^{m\times m}$ for all $l \in [L]$ and $x \in \mathcal{X}$. Indeed, according to part (b) of Lemma 25, we know $\widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l \in [0,1]^{m\times m}$. Then, there holds $\widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l - \mathbf{D}_0^l(x) \in [-1,1]^{m\times m}$ by noting $\mathbf{D}_0^l(x) \in \{0,1\}^{m\times m}$.

Note $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$, then Lemma 22 implies that $\|\mathbf{D}^l(x) - \mathbf{D}_0^l(x)\|_0, \|\widetilde{\mathbf{D}}^l(x) - \mathbf{D}_0^l(x)\|_0 \lesssim (LmR_{op})^{\frac{2}{3}}$ with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ over initialization $(\mathbf{a}, \mathbf{W}(0))$ for all $l \in [L]$. Then, from part (b) of Lemma 25, we know

$$\begin{aligned}
&\big\|\widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^h - \mathbf{D}_0^l(x)\big\|_0 \\
&\leq \big\|\widetilde{\mathbf{D}}^l(x) - \mathbf{D}_0^l(x)\big\|_0 + \big\|\mathbf{D}''\big\|_0 \leq \big\|\widetilde{\mathbf{D}}^l(x) - \mathbf{D}_0^l(x)\big\|_0 + \big\|\widetilde{\mathbf{D}}^l(x) - \mathbf{D}^l(x)\big\|_0 \\
&\leq 2\|\widetilde{\mathbf{D}}^l(x) - \mathbf{D}_0^l(x)\big\|_0 + \|\mathbf{D}^l(x) - \mathbf{D}_0^l(x)\big\|_0 \\
&\lesssim (LmR_{op})^{\frac{2}{3}}.
\end{aligned}$$

Setting $s = (LmR_{op})^{\frac{2}{3}}$, the condition $m \gtrsim L^{22}p^3R_{op}^2\log^3(m/\delta)$ implies the conditions $m \gtrsim L^6\max\{s\log(m), R_{op}^2\}$ and $s \gtrsim pL\log(m/\delta)$ in Lemma 24. Then, by further noting that $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$, and $\widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l, \widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l - \mathbf{D}_0^l(x) \in [-1,1]^{m\times m}$ and $\mathbf{D}^l(x), \mathbf{D}^l(x) - \mathbf{D}_0^l(x) \in [-1,1]^{m\times m}$, we apply Lemma 24 twice with $s = (LmR_{op})^{\frac{2}{3}}$, $\widehat{\mathbf{W}}^l = \widetilde{\mathbf{W}}^l - \mathbf{W}^l(0), \widehat{\mathbf{D}}^l(x) = \widetilde{\mathbf{D}}^l(x) + (\mathbf{D}'')^l - \mathbf{D}_0^l(x)$ and $\widehat{\mathbf{W}}^l = \mathbf{W}^l - \mathbf{W}^l(0), \widehat{\mathbf{D}}^l(x) =$

$\mathbf{D}^l(x) - \mathbf{D}_0^l(x)$, respectively, and there holds

$$\begin{aligned}
\big\|U_l^L(x)\big\|_2 \leq{}& \bigg\|\mathbf{a}^\top\Big[\prod_{h=l+1}^{L}\sqrt{\frac{2}{m}}\big(\widetilde{\mathbf{D}}^h(x)+(\mathbf{D}'')^h\big)\widetilde{\mathbf{W}}^h\Big]\sqrt{\frac{2}{m}}\big(\widetilde{\mathbf{D}}^l(x)+(\mathbf{D}'')^l\big) - \mathbf{a}^\top V_{L,0}^l(x)\bigg\|_2 \\
&+ \bigg\|\mathbf{a}^\top V_{L,0}^l(x) - \mathbf{a}^\top\Big[\prod_{h=l+1}^{L}\sqrt{\frac{2}{m}}\mathbf{D}^h(x)\mathbf{W}^h\Big]\sqrt{\frac{2}{m}}\mathbf{D}^l(x)\bigg\|_2 \\
\lesssim{}& \frac{L\big((LmR_{op})^{\frac{1}{3}}+R_{op}\big)}{\sqrt{m}} \lesssim L^{\frac{4}{3}}R_{op}^{\frac{1}{3}}m^{-\frac{1}{6}},
\end{aligned}$$

where the last inequality used $R_{op} \leq (LmR_{op})^{\frac{1}{3}}$ by noting $m \gtrsim CL^{22}p^3R_{op}^2\log^3(m/\delta)$.

The term $\|o^l(x)\|_2$ can be controlled by using part (a) of Lemma 19, Lemma 21 and $m \gtrsim L^{22}p^3R_{op}^2\log^3(m/\delta)$ by

$$\|o^l(x)\|_2 \leq \|o^l(x) - o_0^l(x)\|_2 + \|o_0^l(x)\|_2 \lesssim \frac{lLR_{op}}{\sqrt{m}} + C \lesssim C.$$

Plugging the above two inequalities back into (4.4.2), we obtain

$$\begin{aligned}
&\Big|f_{\widetilde{\mathbf{W}}}(x) - f_{\mathbf{W}}(x) - \Big\langle\frac{\partial f_{\mathbf{W}}(x)}{\partial\mathbf{W}}, \widetilde{\mathbf{W}}-\mathbf{W}\Big\rangle_2\Big| \\
&\lesssim L^{\frac{4}{3}}\sum_{l=1}^{L}\|\widetilde{\mathbf{W}}^l - \mathbf{W}^l\|_{op}R_{op}^{\frac{1}{3}}m^{-\frac{1}{6}} \\
&\lesssim L^{\frac{7}{3}}\|\widetilde{\mathbf{W}} - \mathbf{W}\|_{op,\infty}R_{op}^{\frac{1}{3}}m^{-\frac{1}{6}}. \qquad (4.29)
\end{aligned}$$

The first part of the lemma is proved.

Now, we show the loss ℓ is locally almost smooth near the initialization. From the convexity of $\ell(\mathbf{W};z)$ (with respect to $f_{\mathbf{W}}$), we know

$$\begin{aligned}
\ell(\widetilde{\mathbf{W}};z) - \ell(\mathbf{W};z) &\geq \frac{\partial\ell(\mathbf{W};z)}{\partial f_{\mathbf{W}}}\big(f_{\widetilde{\mathbf{W}}}(x) - f_{\mathbf{W}}(x)\big) \\
&= \big(f_{\mathbf{W}}(x) - y\big)\big(f_{\widetilde{\mathbf{W}}}(x) - f_{\mathbf{W}}(x)\big).
\end{aligned}$$

Then, according to the chain rule, we get

$$\begin{aligned}
&\ell(\widetilde{\mathbf{W}}; z) - \ell(\mathbf{W}; z) \\
&\geq \big(f_{\mathbf{W}}(x) - y\big)\Big(f_{\widetilde{\mathbf{W}}}(x) - f_{\mathbf{W}}(x) - \Big\langle \frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W}\Big\rangle_2 + \Big\langle \frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W}\Big\rangle_2\Big) \\
&= \big(f_{\mathbf{W}}(x) - y\big)\Big\langle \frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W}\Big\rangle_2 + \big(f_{\mathbf{W}}(x) - y\big)\Big(f_{\widetilde{\mathbf{W}}}(x) - f_{\mathbf{W}}(x) \\
&\quad - \Big\langle \frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W}\Big\rangle_2\Big) \\
&\geq \Big\langle \frac{\partial \ell(\mathbf{W}; z)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W}\Big\rangle_2 - \big|f_{\mathbf{W}}(x) - y\big|\Big|f_{\widetilde{\mathbf{W}}}(x) - f_{\mathbf{W}}(x) - \Big\langle \frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W}\Big\rangle_2\Big|.
\end{aligned} \tag{4.30}$$

Plugging (4.29) back into (4.30), there holds

$$\ell(\widetilde{\mathbf{W}}; z) - \ell(\mathbf{W}; z) \geq \Big\langle \frac{\partial \ell(\mathbf{W}; z)}{\partial \mathbf{W}}, \widetilde{\mathbf{W}} - \mathbf{W}\Big\rangle_2 - |f_{\mathbf{W}}(x) - y| \cdot \epsilon$$

with $\epsilon \lesssim L^{\frac{7}{3}} \|\widetilde{\mathbf{W}} - \mathbf{W}\|_{op,\infty} R_{op}^{\frac{1}{3}} m^{-\frac{1}{6}}$. The second part of the lemma is proved.

Finally, we turn to prove the last part of the lemma. From the above estimates we already know $\|o^l(x) - o_0^l(x)\|_2 \lesssim lLR_{op}m^{-\frac{1}{2}}$, $\|o^l(x)\|_2 \lesssim C$ and

$$\Big\|\mathbf{a}^\top V_{L,0}^l(x) - \mathbf{a}^\top [\prod_{h=l+1}^{L} \sqrt{\frac{2}{m}}\mathbf{D}^h(x)\mathbf{W}^h]\sqrt{\frac{2}{m}}\mathbf{D}^l(x)\Big\|_2 \lesssim L^{\frac{4}{3}} R_{op}^{\frac{1}{3}} m^{-\frac{1}{6}}$$

for all $l \in [L]$ and $x \in \mathcal{X}$. Then, combining these estimates with Lemma 23, there holds

$$\begin{aligned}
&\Big\|\frac{\partial f_{\mathbf{W}}(x)}{\partial \mathbf{W}^l} - \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^l(0)}\Big\|_2 \\
&= \Big\|o^{l-1}(x)\mathbf{a}^\top\Big[\prod_{h=l+1}^{L} \sqrt{\frac{2}{m}}\mathbf{D}^h(x)\mathbf{W}^h\Big]\sqrt{\frac{2}{m}}\mathbf{D}^l(x) - o_0^{l-1}(x)\mathbf{a}^\top V_{L,0}^l(x)\Big\|_2 \\
&\leq \big\|o^{l-1}(x)\big\|_2\Big\|\mathbf{a}^\top\Big[\prod_{h=l+1}^{L}\sqrt{\frac{2}{m}}\mathbf{D}^h(x)\mathbf{W}^h\Big]\sqrt{\frac{2}{m}}\mathbf{D}^l(x) - \mathbf{a}^\top V_{L,0}^l(x)\Big\|_2 + \big\|o^{l-1}(x) - o_0^{l-1}(x)\big\|_2\big\|\mathbf{a}^\top V_{L,0}^l(x)\big\|_2 \\
&\lesssim L^{\frac{4}{3}} R_{op}^{\frac{1}{3}} m^{-\frac{1}{6}} + lL^2 R_{op} m^{-\frac{1}{2}} \lesssim L^{\frac{4}{3}} R_{op}^{\frac{1}{3}} m^{-\frac{1}{6}},
\end{aligned}$$

where the last inequality used $lL^2 R_{op} m^{-\frac{1}{2}} \lesssim L^{\frac{4}{3}} R_{op}^{\frac{1}{3}} m^{-\frac{1}{6}}$ by noting $m \gtrsim L^{22} p^3 R_{op}^2 \log^3(m/\delta)$. The proof is completed. □

4.4.3 Proofs for Gradient Descent

For notational convenience, define $\boldsymbol{f}_{\mathbf{W}(k)} = (f_{\mathbf{W}(k)}(x_1), \ldots, f_{\mathbf{W}(k)}(x_n))^\top \in \mathbb{R}^n$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. The following lemma shows that the trajectory of GD during the training process is always near the initialization. Note that we make no assumption on the data distribution and the NTK Gram matrix.

LEMMA 27. *Let $\delta \in (0,1)$ and $\{\mathbf{W}(k)\}$ be produced by* (4.3) *with $\eta \leq 1/5$. Assume* (4.10) *holds. Then, with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ over initialization $(\mathbf{a}, \mathbf{W}(0))$, for any $k \in [T]$, there holds*

$$\big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_{op,\infty}^2 \leq \big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_2^2 \leq 4\eta k$$

$$\|\boldsymbol{f}_{\mathbf{W}(k)} - \mathbf{y}\|_2 \leq 2\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2.$$

PROOF. The lemma is proved by induction. It's obvious that $\|\mathbf{W}(k) - \mathbf{W}(0)\|_2^2 \leq 0$ and $\|\boldsymbol{f}_{\mathbf{W}(k)} - \mathbf{y}\|_2 \leq 2\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2$ hold with $k = 0$. Assume, for all $t \in [k]$ with $k \leq T-1$, $\|\mathbf{W}(t) - \mathbf{W}(0)\|_2^2 \leq 4\eta t$ and $\|\boldsymbol{f}_{\mathbf{W}(t)} - \mathbf{y}\|_2 \leq 2\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2$ hold. We will show that $\|\mathbf{W}(k+1) - \mathbf{W}(0)\|_2^2 \leq 4\eta(k+1)$ and $\|\boldsymbol{f}_{\mathbf{W}(k+1)} - \mathbf{y}\|_2 \leq 2\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2$.

From the update rule (4.3), we know

$$\begin{aligned}
\big\|\mathbf{W}(k+1) - \mathbf{W}(0)\big\|_2^2 &= \Big\|\mathbf{W}(k) - \mathbf{W}(0) - \frac{\eta}{n}\sum_{i=1}^n \frac{\partial \ell(\mathbf{W}(k); z_i)}{\partial \mathbf{W}(k)}\Big\|_2^2 \\
&\leq \big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_2^2 + \eta^2\Big(\frac{1}{n}\sum_{i=1}^n \Big\|(f_{\mathbf{W}(k)}(x_i) - y_i)\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial \mathbf{W}(k)}\Big\|_2\Big)^2 \\
&\quad + \frac{2\eta}{n}\Big\langle \mathbf{W}(0) - \mathbf{W}(k), \sum_{i=1}^n \frac{\partial \ell(\mathbf{W}(k); z_i)}{\partial \mathbf{W}(k)}\Big\rangle_2 \\
&\leq \big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_2^2 + \frac{2\eta^2 \mathcal{E}_z(\mathbf{W}(k))}{n}\sum_{i=1}^n \Big\|\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial \mathbf{W}(k)}\Big\|_2^2 \\
&\quad + \frac{2\eta}{n}\Big\langle \mathbf{W}(0) - \mathbf{W}(k), \sum_{i=1}^n \frac{\partial \ell(\mathbf{W}(k); z_i)}{\partial \mathbf{W}(k)}\Big\rangle_2, \qquad (4.31)
\end{aligned}$$

where in the last inequality we have used $\frac{\|\boldsymbol{f}_{\mathbf{W}(k)} - \mathbf{y}\|_2^2}{n} = 2\mathcal{E}_z(\mathbf{W}(k))$.

Now, we turn to control $\|\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial \mathbf{W}(k)}\|_2$ and $\frac{2\eta}{n}\big\langle \mathbf{W}(0)-\mathbf{W}(k), \sum_{i=1}^n \frac{\partial \ell(\mathbf{W}(k);z_i)}{\partial \mathbf{W}(k)}\big\rangle_2$. Setting $R_{op}=2\sqrt{\eta T}$. By the induction assumption, there holds $\mathbf{W}(k)\in\mathcal{B}_{R_{op}}(\mathbf{W}(0))$. Then from Lemma 16 (if $l<L$) and part (c) of Lemma 19 (if $l=L$) and (4.27) in Lemma 26 with $R_{op}=2\sqrt{\eta T}$ and $\mathbf{W}=\mathbf{W}(k)$, there holds

$$\begin{aligned}\Big\|\frac{\partial f_{\mathbf{W}(k)}(\mathbf{x}_i)}{\partial \mathbf{W}(k)}\Big\|_2 &\le \Big\|\frac{\partial f_{\mathbf{W}(k)}(\mathbf{x}_i)}{\partial \mathbf{W}(k)}-\frac{\partial f_{\mathbf{W}(0)}(\mathbf{x}_i)}{\partial \mathbf{W}(0)}\Big\|_2+\Big\|\frac{\partial f_{\mathbf{W}(0)}(\mathbf{x}_i)}{\partial \mathbf{W}^L(0)}\Big\|_2\\ &\le \sqrt{L}\max_{\ell\in[L]}\Big\|\frac{\partial f_{\mathbf{W}(k)}(\mathbf{x}_i)}{\partial \mathbf{W}^\ell(k)}-\frac{\partial f_{\mathbf{W}(0)}(\mathbf{x}_i)}{\partial \mathbf{W}^\ell(0)}\Big\|_2+2\\ &\le \epsilon_3+2 \end{aligned} \tag{4.32}$$

with $\epsilon_3\lesssim L^{\frac{7}{3}}(\eta T)^{\frac{1}{6}}m^{-\frac{1}{6}}$.

According to (4.26) with $R_{op}=2\sqrt{\eta T}$, $\mathbf{W}=\mathbf{W}(k)$, $\widetilde{\mathbf{W}}=\mathbf{W}(0)$ and $\|\mathbf{W}(k)-\mathbf{W}(0)\|_{op,\infty}\le 2\sqrt{\eta T}$, we know

$$\frac{2\eta}{n}\Big\langle \mathbf{W}(0)-\mathbf{W}(k), \sum_{i=1}^n\frac{\partial \ell(\mathbf{W}(k);z_i)}{\partial \mathbf{W}(k)}\Big\rangle_2 \le 2\eta\big(\mathcal{E}_z(\mathbf{W}(0))-\mathcal{E}_z(\mathbf{W}(k))\big)+2\eta\epsilon_2\sum_{i=1}^n\frac{|f_{\mathbf{W}(k)}(x_i)-y_i|}{n},$$

with $\epsilon_2\lesssim L^{\frac{7}{3}}(\eta T)^{\frac{2}{3}}m^{-\frac{1}{6}}$.

Plugging the above two estimates back into (4.31), we get

$$
\begin{aligned}
&\big\|\mathbf{W}(k+1)-\mathbf{W}(0)\big\|_2^2\\
&\le \big\|\mathbf{W}(k)-\mathbf{W}(0)\big\|_2^2+2\eta^2(\epsilon_3+2)^2\mathcal{E}_z(\mathbf{W}(k))+2\eta\big(\mathcal{E}_z(\mathbf{W}(0))-\mathcal{E}_z(\mathbf{W}(k))\big)\\
&\qquad+2\eta\epsilon_2\sum_{i=1}^n\frac{1}{n}|f_{\mathbf{W}(k)}(x_i)-y_i|\\
&\le \big\|\mathbf{W}(k)-\mathbf{W}(0)\big\|_2^2+2\eta^2(\epsilon_3+2)^2\mathcal{E}_z(\mathbf{W}(k))+2\eta\big(\mathcal{E}_z(\mathbf{W}(0))-\mathcal{E}_z(\mathbf{W}(k))\big)\\
&\qquad+2\eta\epsilon_2\frac{1}{\sqrt{n}}\|\boldsymbol{f}_{\mathbf{W}(k)}-\mathbf{y}\|_2\\
&\le \big\|\mathbf{W}(k)-\mathbf{W}(0)\big\|_2^2+2\eta^2(\epsilon_3+2)^2\mathcal{E}_z(\mathbf{W}(k))+2\eta\big(\mathcal{E}_z(\mathbf{W}(0))-\mathcal{E}_z(\mathbf{W}(k))\big)+4\eta\epsilon_2\\
&\le \big\|\mathbf{W}(k)-\mathbf{W}(0)\big\|_2^2+2\eta\mathcal{E}_z(\mathbf{W}(k))\big(\eta(\epsilon_3+2)^2-1\big)+\eta+4\eta\epsilon_2\\
&\le \big\|\mathbf{W}(k)-\mathbf{W}(0)\big\|_2^2+2\eta\mathcal{E}_z(\mathbf{W}(k))\big(5\eta-1\big)+\eta+2\eta\\
&\le 4\eta k+\eta+2\eta<4\eta(k+1),
\end{aligned}
$$

where in the second inequality we have used Cauchy-Schwarz inequality, and in the third inequality we have used the induction assumption $\|\boldsymbol{f}_{\mathbf{W}(k)}-\mathbf{y}\|_2\le 2\|\boldsymbol{f}_{\mathbf{W}(0)}-\mathbf{y}\|_2=2\|\mathbf{y}\|_2\le 2\sqrt{n}$ by noting $f_{\mathbf{W}(0)}=0$, and in the last third inequality we have used $\mathcal{E}_z(\mathbf{W}(0))=\frac{1}{2n}\sum_{i=1}^n y_i^2\le\frac{1}{2}$, and the last second inequality used $\epsilon_2\le\frac{1}{2}$ and $\epsilon_3\le\sqrt{5}-2$ by condition (4.10), and the last inequality follows from the induction assumption and $\eta\le\frac{1}{5}$ and $\mathcal{E}_z(\mathbf{W}(k))\ge 0$.

Now, we turn to estimate $\|\boldsymbol{f}_{\mathbf{W}(k+1)}-\mathbf{y}\|_2$. Let

$$
\xi_i(k)=f_{\mathbf{W}(k+1)}(x_i)-f_{\mathbf{W}(k)}(x_i)-\Big\langle\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial\mathbf{W}(k)},\mathbf{W}(k+1)-\mathbf{W}(k)\Big\rangle_2,
$$

there holds for all $i\in[n]$ that

$$
\begin{aligned}
&f_{\mathbf{W}(k+1)}(x_i)-y_i=f_{\mathbf{W}(k+1)}(x_i)-f_{\mathbf{W}(k)}(x_i)+f_{\mathbf{W}(k)}(x_i)-y_i\\
&=\Big\langle\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial\mathbf{W}(k)},\mathbf{W}(k+1)-\mathbf{W}(k)\Big\rangle_2+\xi_i(k)+f_{\mathbf{W}(k)}(x_i)-y_i\\
&=-\frac{\eta}{n}\sum_{j=1}^n\Big\langle\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial\mathbf{W}(k)},\frac{\partial f_{\mathbf{W}(k)}(x_j)}{\partial\mathbf{W}(k)}\Big\rangle_2(f_{\mathbf{W}(k)}(x_j)-y_j)+\xi_i(k)+f_{\mathbf{W}(k)}(x_i)-y_i,
\end{aligned}
$$

where in the last equality we used the update rule (4.3). Define the matrix $\mathbf{H}(k) \in \mathbb{R}^{n\times n}$ by $(\mathbf{H}(k))_{i,j} := \big\langle \frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial \mathbf{W}(k)}, \frac{\partial f_{\mathbf{W}(k)}(x_j)}{\partial \mathbf{W}(k)} \big\rangle_2$. Denote $\boldsymbol{\xi}(k) = (\xi_1(k), \ldots, \xi_n(k))^\top \in \mathbb{R}^n$. Then, the above observation implies

$$\begin{aligned} \boldsymbol{f}_{\mathbf{W}(k+1)} - \mathbf{y} &= \boldsymbol{f}_{\mathbf{W}(k)} - \mathbf{y} - \frac{\eta}{n}\mathbf{H}(k)(\boldsymbol{f}_{\mathbf{W}(k)} - \mathbf{y}) + \boldsymbol{\xi}(k) \\ &= \Big(\mathbf{I} - \frac{\eta}{n}\mathbf{H}(k)\Big)(\boldsymbol{f}_{\mathbf{W}(k)} - \mathbf{y}) + \boldsymbol{\xi}(k). \end{aligned}$$

Applying the above equality recursively, we get

$$\boldsymbol{f}_{\mathbf{W}(k+1)} - \mathbf{y} = \sum_{s=0}^{k} \prod_{u=s+1}^{k} \Big(\mathbf{I} - \frac{\eta}{n}\mathbf{H}(u)\Big)\boldsymbol{\xi}(s) - \prod_{s=0}^{k}\Big(\mathbf{I} - \frac{\eta}{n}\mathbf{H}(s)\Big)\mathbf{y},$$

where we used the conventional notation $\prod_k^{k-1} = \mathbf{I}$. Then, there holds

$$\|\boldsymbol{f}_{\mathbf{W}(k+1)} - \mathbf{y}\|_2 \leq \sum_{s=0}^{k} \prod_{u=s+1}^{k} \Big\|\mathbf{I} - \frac{\eta}{n}\mathbf{H}(u)\Big\|_{op} \|\boldsymbol{\xi}(s)\|_2 + \prod_{s=0}^{k}\Big\|\mathbf{I} - \frac{\eta}{n}\mathbf{H}(s)\Big\|_{op}\|\mathbf{y}\|_2. \quad (4.33)$$

Now, we turn to estimate $\|\mathbf{I} - \frac{\eta}{n}\mathbf{H}(s)\|_{op}$ and $\|\boldsymbol{\xi}(k)\|_2$. According to (4.32) and (4.10), there holds $\epsilon_3 + 2 \leq \sqrt{5}$. By further noting that $\eta \leq \frac{1}{5}$, we know for all $s \in [k]$

$$\begin{aligned} \Big\|\frac{\eta}{n}\mathbf{H}(s)\Big\|_{op}^2 \leq \Big\|\frac{\eta}{n}\mathbf{H}(s)\Big\|_2^2 &= \frac{\eta^2}{n^2}\sum_{i,j=1}^{n}\Big\langle \frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial \mathbf{W}(k)}, \frac{\partial f_{\mathbf{W}(k)}(x_j)}{\partial \mathbf{W}(k)}\Big\rangle_2^2 \\ &\leq \frac{\eta^2}{n^2}\sum_{i,j=1}^{n}\Big\|\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial \mathbf{W}(k)}\Big\|_2^2\Big\|\frac{\partial f_{\mathbf{W}(k)}(x_j)}{\partial \mathbf{W}(k)}\Big\|_2^2 \leq 25\eta^2 \leq 1. \end{aligned}$$

Since $\frac{\eta}{n}\mathbf{H}(s)$ is a PSD matrix whose operator norm is not larger than 1, then $\|\mathbf{I} - \frac{\eta}{n}\mathbf{H}(s)\|_{op} \leq 1$.

Note we already showed that $\mathbf{W}(s+1) \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$ with $R_{op} = 2\sqrt{\eta T}$ for all $s \leq k$. From (4.25) in Lemma 26, we get

$$\begin{aligned}
\|\boldsymbol{\xi}(s)\|_2 = \Big(\sum_{i=1}^n \xi_i(s)^2\Big)^{\frac{1}{2}} &\lesssim L^{\frac{7}{3}}\sqrt{n}\big\|\mathbf{W}(s+1) - \mathbf{W}(s)\big\|_{op,\infty}(\eta T)^{\frac{1}{6}} m^{-\frac{1}{6}} \\
&= L^{\frac{7}{3}}\Big\|\frac{\eta}{\sqrt{n}}\sum_{i=1}^n \frac{\partial f_{\mathbf{W}(s)}(x_i)}{\partial \mathbf{W}(s)}(f_{\mathbf{W}(s)}(x_i) - y_i)\Big\|_{op,\infty}(\eta T)^{\frac{1}{6}} m^{-\frac{1}{6}} \\
&\lesssim L^{\frac{7}{3}}\Big[\sup_{l\in[L]}\frac{\eta}{\sqrt{n}}\sum_{i=1}^n \Big\|\frac{\partial f_{\mathbf{W}(s)}(x_i)}{\partial \mathbf{W}^l(s)}\Big\|_{op}|f_{\mathbf{W}(s)}(x_i) - y_i|\Big](\eta T)^{\frac{1}{6}} m^{-\frac{1}{6}} \\
&\lesssim L^{\frac{7}{3}}\eta\|\boldsymbol{f}_{\mathbf{W}(s)} - \mathbf{y}\|_2 \sup_{l\in[L],i\in[n]}\Big\|\frac{\partial f_{\mathbf{W}(s)}(x_i)}{\partial \mathbf{W}^l(s)}\Big\|_{op}(\eta T)^{\frac{1}{6}} m^{-\frac{1}{6}} \\
&\lesssim L^{\frac{7}{3}}\eta\|\boldsymbol{f}_{\mathbf{W}(s)} - \mathbf{y}\|_2(\eta T)^{\frac{1}{6}} m^{-\frac{1}{6}} \\
&\lesssim L^{\frac{7}{3}}\eta\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2(\eta T)^{\frac{1}{6}} m^{-\frac{1}{6}},
\end{aligned}$$

where the last third inequality used Cauchy-Schwarz inequality, and in the last second inequality we have used (4.32) with $\epsilon_3 + 2 \lesssim C$ by noting (4.10), and in the last inequality we have used the induction assumption $\|\boldsymbol{f}_{\mathbf{W}(s)} - \mathbf{y}\|_2 \leq 2\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2$.

Plugging the estimates $\|\mathbf{I} - \frac{\eta}{n}\mathbf{H}(s)\|_{op} \leq 1$ and the above inequality back into (4.33), and noting the condition (4.10) and $f_{\mathbf{W}(0)} = 0$, there holds

$$\begin{aligned}
\|f_{\mathbf{W}(k+1)} - \mathbf{y}\|_2 &\leq CL^{\frac{7}{3}}\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2(\eta T)^{\frac{7}{6}} m^{-\frac{1}{6}} + \|\mathbf{y}\|_2 \\
&= CL^{\frac{7}{3}}\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2(\eta T)^{\frac{7}{6}} m^{-\frac{1}{6}} + \|f_{\mathbf{W}(0)} - \mathbf{y}\|_2 \leq 2\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2.
\end{aligned}$$

This completes the proof of the lemma. □

Based on Lemma 27, we present the proof of Proposition 7 as follows.

PROOF OF PROPOSITION 7. Setting $R_{op} = 2\sqrt{\eta T}$. From Lemma 27, we know with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ over initialization $(\mathbf{a}, \mathbf{W}(0))$ that $\mathbf{W}(k) \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$. Then, (4.25) in Lemma 26 with $\widetilde{\mathbf{W}} = \mathbf{W}(k)$ and $\mathbf{W} = \mathbf{W}(0)$ implies

$$\Big|f_{\mathbf{W}(k)}(x) - f_{\mathbf{W}(0)}(x) - \Big\langle\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}(0)}, \mathbf{W}(k) - \mathbf{W}(0)\Big\rangle_2\Big| \lesssim L^{\frac{7}{3}}(\eta T)^{\frac{2}{3}} m^{-\frac{1}{6}}.$$

Then, there holds

$$\left\|f_{\mathbf{W}(T)}-f^{\text{lin}}_{\mathbf{W}(T)}\right\|_\rho^2\leq\left\|f_{\mathbf{W}(T)}-f^{\text{lin}}_{\mathbf{W}(T)}\right\|_\infty^2\lesssim\frac{L^{\frac{14}{3}}(\eta T)^{\frac{4}{3}}}{m^{\frac{1}{3}}}.$$

Since we assume that we are under the event $\{\|\mathbf{W}(0)\|_{op,\infty}\leq c_0\sqrt{m}\}$, whose probability is at least $1-L\exp(-Cm)$ according to Lemma 17. By further noting that $1-L\exp(\mathcal{O}(pL\log(m))-\Omega(m^{\frac{1}{3}}))-\delta\leq 1-L\exp(-Cm)$, the proof is completed. □

Let H be a separable Hilbert space. For $f\in H$, we define the operator $f\otimes f:H\to H$ by $(f\otimes f)g=\langle f,g\rangle_H f$. To estimate $\left\|f^{lin}_{\mathbf{W}(T)}-\mathbf{S}_m g_T^m\right\|_\rho^2$, we introduce the following useful lemma.

LEMMA 28 (Lemma 3 in [Carratino et al. 2018]). *Let $\lambda>0$, $\Gamma\in\mathbb{N}$ and $\delta\in(0,1)$. Let $\zeta_1,\ldots,\zeta_\Gamma$ be independent and identically distributed random vectors bounded by $\kappa>0$. Let $Q_\Gamma=\frac{1}{\Gamma}\sum_{i=1}^\Gamma\zeta_i\otimes\zeta_i$ and Q be the expectation of Q_Γ. Then, for any $\lambda\geq\frac{9\kappa^2}{\Gamma}\log\frac{\Gamma}{\delta}$, with probability at least $1-\delta$ over sampling, there holds*

$$\|(Q+\lambda\mathbf{I})^{1/2}(Q_\Gamma+\lambda\mathbf{I})^{-1/2}\|_{op}^2=\|(Q_\Gamma+\lambda\mathbf{I})^{-1/2}(Q+\lambda\mathbf{I})^{1/2}\|_{op}^2\leq 2.$$

Now, we give the estimate of the second term $\left\|f^{lin}_{\mathbf{W}(T)}-\mathbf{S}_m g_T^m\right\|_\rho^2$ as follows.

PROOF OF PROPOSITION 8. According to Lemma 17, we know $\|\mathbf{W}(0)\|_{op,\infty}\leq c_0\sqrt{m}$ holds with probability at least $1-L\exp(-Cm)$ over the random choice of $\mathbf{W}(0)$.

Let $F_k=f^{\text{lin}}_{\mathbf{W}(k)}-g_k^m\in\mathcal{H}_m$ and $\epsilon_k^1=f^{\text{lin}}_{\mathbf{W}(k+1)}-f^{\text{lin}}_{\mathbf{W}(k)}+\frac{\eta}{n}\sum_{i=1}^n(f^{\text{lin}}_{\mathbf{W}(k)}(x_i)-y_i)K_{x_i}^m\in\mathcal{H}_m$. Define the self-adjoint positive operator $\widehat{\mathbf{\Sigma}}_m=\frac{1}{n}\sum_{i=1}^n K_{x_i}^m\otimes K_{x_i}^m:\mathcal{H}_m\to\mathcal{H}_m$. From the update rule of g_k^m (4.7), we know

$$\begin{aligned}F_{k+1}&=\Big(f^{\text{lin}}_{\mathbf{W}(k)}-\frac{\eta}{n}\sum_{i=1}^n(f^{\text{lin}}_{\mathbf{W}(k)}(x_i)-y_i)K_{x_i}^m+\epsilon_k^1\Big)-\Big(g_k^m-\frac{\eta}{n}\sum_{i=1}^n(g_k^m(x_i)-y_i)K_{x_i}^m\Big)\\&=\big(f^{\text{lin}}_{\mathbf{W}(k)}-g_k^m\big)-\frac{\eta}{n}\sum_{i=1}^n\big(f^{\text{lin}}_{\mathbf{W}(k)}(x_i)-g_k^m(x_i)\big)K_{x_i}^m+\epsilon_k^1\\&=F_k-\frac{\eta}{n}\sum_{i=1}^n\langle F_k,K_{x_i}^m\rangle_{\mathcal{H}_m}K_{x_i}^m+\epsilon_k^1=\big(\mathbf{I}-\eta\widehat{\mathbf{\Sigma}}_m\big)F_k+\epsilon_k^1,\end{aligned}\tag{4.34}$$

where the last second equality follows from the fact $F_k = f^{\text{lin}}_{\mathbf{W}(k)} - g^m_k \in \mathcal{H}_m$ and the reproducing kernel property that $f^{\text{lin}}_{\mathbf{W}(k)}(x_i) - g^m_k(x_i) = \langle f^{\text{lin}}_{\mathbf{W}(k)} - g^m_k, K^m_{x_i}\rangle_{\mathcal{H}_m} = \langle F_k, K^m_{x_i}\rangle_{\mathcal{H}_m}$.

Applying the above equality recursively, we get

$$F_{k+1} = \sum_{s=0}^{k} \big(\mathbf{I} - \eta\widehat{\mathbf{\Sigma}}_m\big)^s \epsilon^1_{k-s}. \tag{4.35}$$

Define the mean of $\widehat{\mathbf{\Sigma}}_m$ by $\mathbf{\Sigma}_m := \mathbb{E}[\widehat{\mathbf{\Sigma}}_m] = \int_{\mathcal{X}} K^m_x \otimes K^m_x d\rho_x(x) : \mathcal{H}_m \to \mathcal{H}_m$. Mercer's Theorem [Steinwart and Christmann 2008] implies $\|\mathbf{S}_m f\|_\rho = \|\mathbf{\Sigma}_m^{\frac{1}{2}} f\|_{\mathcal{H}_m}$ for any $f \in \mathcal{H}_m$. From Lemmas 16 and 19 we know $\|K^m_{x_i}\|_{\mathcal{H}_m} = \sqrt{K^m(x_i, x_i)} \le \sqrt{\|K^m\|_\infty} \le \sup_{x\in\mathcal{X}} \|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\|_2 \le 2$. Therefore, Lemma 28 with $\zeta_i = K^m_{x_i}$, $\Gamma = n$ and $\kappa = 2$ yields $\|(\mathbf{\Sigma}_m + \lambda\mathbf{I})^{\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_m + \lambda\mathbf{I})^{-\frac{1}{2}}\|_{op} \le 2$ with probability at least $1 - \delta/2$ over the sampling if $\lambda > \frac{36}{n}\log(\frac{2n}{\delta})$. Then, according to (4.35), we get

$$\begin{aligned}
&\|\mathbf{S}_m F_k\|_\rho = \big\|\mathbf{\Sigma}_m^{\frac{1}{2}} F_k\big\|_{\mathcal{H}_m} \le \big\|(\mathbf{\Sigma}_m + \lambda\mathbf{I})^{\frac{1}{2}} F_k\big\|_{\mathcal{H}_m} \\
&\le \big\|(\mathbf{\Sigma}_m + \lambda\mathbf{I})^{\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_m + \lambda\mathbf{I})^{-\frac{1}{2}}\big\|_{op}\big\|(\widehat{\mathbf{\Sigma}}_m + \lambda\mathbf{I})^{\frac{1}{2}} F_k\big\|_{\mathcal{H}_m} \le 2\big\|\widehat{\mathbf{\Sigma}}_m^{\frac{1}{2}} F_k\big\|_{\mathcal{H}_m} + 2\sqrt{\lambda}\big\|F_k\big\|_{\mathcal{H}_m} \\
&= 2\eta^{-\frac{1}{2}}\bigg\|\sum_{s=0}^{k-1}\big(\eta\widehat{\mathbf{\Sigma}}_m\big)^{\frac{1}{2}}\big(\mathbf{I} - \eta\widehat{\mathbf{\Sigma}}_m\big)^s \epsilon^1_{k-s-1}\bigg\|_{\mathcal{H}_m} + 2\sqrt{\lambda}\bigg\|\sum_{s=0}^{k-1}\big(\mathbf{I} - \eta\widehat{\mathbf{\Sigma}}_m\big)^s \epsilon^1_{k-s-1}\bigg\|_{\mathcal{H}_m} \\
&\le 2\eta^{-\frac{1}{2}}\sum_{s=0}^{k-1}\Big\|\big(\eta\widehat{\mathbf{\Sigma}}_m\big)^{\frac{1}{2}}\big(\mathbf{I} - \eta\widehat{\mathbf{\Sigma}}_m\big)^s\Big\|_{op}\big\|\epsilon^1_{k-s-1}\big\|_{\mathcal{H}_m} + 2\sqrt{\lambda}\sum_{s=0}^{k-1}\Big\|\big(\mathbf{I} - \eta\widehat{\mathbf{\Sigma}}_m\big)^s\Big\|_{op}\big\|\epsilon^1_{k-s-1}\big\|_{\mathcal{H}_m}.
\end{aligned} \tag{4.36}$$

For any $a \in [0, 1)$ and any $s \in \mathbb{N}$, it can be easily computed that $\sup_{t\in[0,1]} t^a(1-t)^s \le \big(\frac{a}{a+s}\big)^a$. Here, we take notation $0^0 = 1$. From (4.41) and $\eta \le 1/5$ we know $\eta\|\widehat{\mathbf{\Sigma}}_m\|_{op} \le \frac{\eta}{n}\sum_{j=1}^n \|K^m_{x_i} \otimes K^m_{x_i}\|_{op} = \frac{\eta}{n}\sum_{j=1}^n \|K^m_{x_i}\|^2_{\mathcal{H}_m} \le \eta\|K^m\|_\infty \le \eta \sup_{x\in\mathcal{X}} \|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\|_2^2 \le 1$.

Then, there holds

$$\sum_{s=0}^{k-1}\big\|(\eta\widehat{\mathbf{\Sigma}}_m)^a(\mathbf{I}-\eta\widehat{\mathbf{\Sigma}}_m)^s\big\|_{op}\le\sum_{s=0}^{k-1}\sup_{t\in[0,1]}t^a(1-t)^s\le\sum_{s=0}^{k-1}\Big(\frac{a}{a+s}\Big)^a=1+a^a\sum_{s=1}^{k-1}\Big(\frac{1}{a+s}\Big)^a$$
$$\le 1+a^a\sum_{s=1}^{k-1}\int_{s-1}^{s}\Big(\frac{1}{a+x}\Big)^a dx=1+a^a\int_0^{k-1}\Big(\frac{1}{a+x}\Big)^a dx$$
$$\le 1+\frac{a^a}{1-a}\big((k+a-1)^{1-a}-a^{1-a}\big)\le 1+\frac{(k+a-1)^{1-a}}{1-a}.$$

Combining (4.36) and the above inequality with $a=\frac{1}{2}$ and $a=0$, respectively, we have

$$\big\|\mathbf{S}_m F_k\big\|_\rho\le\bigg(2\eta^{-\frac{1}{2}}\sum_{s=0}^{k-1}\Big\|\big(\eta\widehat{\mathbf{\Sigma}}_m\big)^{\frac{1}{2}}\big(\mathbf{I}-\eta\widehat{\mathbf{\Sigma}}_m\big)^s\Big\|_{op}+2\sqrt{\lambda}\sum_{s=0}^{k-1}\Big\|\big(\mathbf{I}-\eta\widehat{\mathbf{\Sigma}}_m\big)^s\Big\|_{op}\bigg)\max_{s\in[k-1]}\|\epsilon_s^1\|_{\mathcal{H}_m}$$
$$\le\Big(2\eta^{-\frac{1}{2}}\big(1+2\sqrt{k}\big)+2\sqrt{\lambda}k\Big)\max_{s\in[k-1]}\|\epsilon_s^1\|_{\mathcal{H}_m}. \tag{4.37}$$

It remains to estimate $\|\epsilon_k^1\|_{\mathcal{H}_m}$. From the definition of $f_{\mathbf{W}}^{\text{lin}}$ and the update rule of GD (4.3),

$$\epsilon_k^1(x)=f_{\mathbf{W}(k+1)}^{\text{lin}}(x)-f_{\mathbf{W}(k)}^{\text{lin}}(x)+\frac{\eta}{n}\sum_{i=1}^{n}\big(f_{\mathbf{W}(k)}^{\text{lin}}(x_i)-y_i\big)K_{x_i}^m(x)$$
$$=\Big\langle\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}(0)},\mathbf{W}(k+1)-\mathbf{W}(k)\Big\rangle_2+\frac{\eta}{n}\sum_{i=1}^{n}\big(f_{\mathbf{W}(k)}^{\text{lin}}(x_i)-y_i\big)\Big\langle\frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial\mathbf{W}(0)},\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}(0)}\Big\rangle_2$$
$$=\Big\langle\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}^L(0)},\mathbf{W}^L(k+1)-\mathbf{W}^L(k)\Big\rangle_2+\frac{\eta}{n}\sum_{i=1}^{n}\big(f_{\mathbf{W}(k)}^{\text{lin}}(x_i)-y_i\big)\Big\langle\frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial\mathbf{W}^L(0)},\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}^L(0)}\Big\rangle_2$$
$$=\frac{\eta}{n}\sum_{i=1}^{n}\bigg[-\big(f_{\mathbf{W}(k)}(x_i)-y_i\big)\Big\langle\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial\mathbf{W}^L(k)},\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}^L(0)}\Big\rangle_2$$
$$+\big(f_{\mathbf{W}(k)}^{\text{lin}}(x_i)-y_i\big)\Big\langle\frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial\mathbf{W}^L(0)},\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}^L(0)}\Big\rangle_2\bigg]$$
$$=\bigg\langle\frac{\eta}{n}\sum_{i=1}^{n}\Big[\big(y_i-f_{\mathbf{W}(k)}(x_i)\big)\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial\mathbf{W}^L(k)}+\big(f_{\mathbf{W}(k)}^{\text{lin}}(x_i)-y_i\big)\frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial\mathbf{W}^L(0)}\Big],\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}^L(0)}\bigg\rangle_2$$
$$=:\Big\langle\Delta(k),\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}^L(0)}\Big\rangle_2,$$

where the second equality is due to $K_{x_i}^m(x)=K^m(x_i,x)$, the third equality used $\frac{\partial f_{\mathbf{W}(0)}}{\partial\mathbf{W}^l(0)}=0$ for $l\in[L-1]$ according to Lemma 16, the fourth equality is according to the update rule

(4.3), and in the last equality we define

$$\Delta(k) := \frac{\eta}{n}\sum_{i=1}^{n}\Big[\big(y_i - f_{\mathbf{W}(k)}(x_i)\big)\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial \mathbf{W}^L(k)} + \big(f^{\text{lin}}_{\mathbf{W}(k)}(x_i) - y_i\big)\frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial \mathbf{W}^L(0)}\Big]$$
$$= \frac{\eta}{n}\sum_{i=1}^{n}\Big[\big(y_i - f_{\mathbf{W}(k)}(x_i)\big)\Big(\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial \mathbf{W}^L(k)} - \frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial \mathbf{W}^L(0)}\Big) + \big(f^{\text{lin}}_{\mathbf{W}(k)}(x_i) - f_{\mathbf{W}(k)}(x_i)\big)\frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial \mathbf{W}^L(0)}\Big].$$

Let $\mathbf{\Delta}(k) = (0, \ldots, 0, \Delta(k)) \in \mathcal{W}$, then we know $\epsilon_k^1(x) = \langle \mathbf{\Delta}(k), \Phi_m(x)\rangle_2$. Note that for any $f \in \mathcal{H}_m$, $\|f\|^2_{\mathcal{H}_m} = \inf\big\{\sum_{l=1}^{L}\|\mathbf{W}^l\|_2^2 : \mathbf{W} \in \mathcal{W} \text{ with } f(x) = \big\langle \mathbf{W}, \Phi_m(x)\big\rangle_2\big\}$. We control $\|\epsilon_k^1\|^2_{\mathcal{H}_m}$ as follows

$$\begin{aligned}
&\|\epsilon_k^1\|_{\mathcal{H}_m} \le \|\Delta(k)\|_2 \\
&\le \frac{\eta}{n}\sum_{i=1}^{n}\big|y_i - f_{\mathbf{W}(k)}(x_i)\big|\Big\|\frac{\partial f_{\mathbf{W}(k)}(x_i)}{\partial \mathbf{W}^L(k)} - \frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial \mathbf{W}^L(0)}\Big\|_2 + \big|f^{\text{lin}}_{\mathbf{W}(k)}(x_i) - f_{\mathbf{W}(k)}(x_i)\big|\Big\|\frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial \mathbf{W}^L(0)}\Big\|_2 \\
&\le \eta\Big(\frac{\|\boldsymbol{f}_{\mathbf{W}(k)} - \mathbf{y}\|_2}{\sqrt{n}}\sup_{x\in\mathcal{X}}\Big\|\frac{\partial f_{\mathbf{W}(k)}(x)}{\partial \mathbf{W}^L(k)} - \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\|_2 + \sum_{i=1}^{n}\frac{\big|f^{\text{lin}}_{\mathbf{W}(k)}(x_i) - f_{\mathbf{W}(k)}(x_i)\big|}{n}\Big\|\frac{\partial f_{\mathbf{W}(0)}(x_i)}{\partial \mathbf{W}^L(0)}\Big\|_2\Big) \\
&\le \eta\Big(\frac{\|\boldsymbol{f}_{\mathbf{W}(k)} - \mathbf{y}\|_2}{\sqrt{n}}\sup_{x\in\mathcal{X}}\Big\|\frac{\partial f_{\mathbf{W}(k)}(x)}{\partial \mathbf{W}^L(k)} - \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\|_2 + \big\|f^{\text{lin}}_{\mathbf{W}(k)} - f_{\mathbf{W}(k)}\big\|_\infty \sup_{x\in\mathcal{X}}\Big\|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\|_2\Big),
\end{aligned} \tag{4.38}$$

where in the last second inequality we have used Cauchy-Schwarz inequality.

From part (c) of Lemma 19, we know $\sup_{x\in\mathcal{X}}\|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\|_2 \le 2$. Note we choose $R_{op} = 2\sqrt{\eta T}$. From Lemma 27 we know $\|\boldsymbol{f}_{\mathbf{W}(k)} - \mathbf{y}\|_2 \le 2\|\boldsymbol{f}_{\mathbf{W}(0)} - \mathbf{y}\|_2 = 2\|\mathbf{y}\|_2 \le 2\sqrt{n}$ and $\mathbf{W}(k) \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$ for any $k \in [T]$. Combining this with Lemma 26, there holds

$$\Big\|\frac{\partial f_{\mathbf{W}(k)}(x)}{\partial \mathbf{W}^L(k)} - \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\|_2 \lesssim L^{\frac{4}{3}}\Big(\frac{\eta T}{m}\Big)^{\frac{1}{6}}.$$

According to Proposition 7, we know $\|f_{\mathbf{W}(k)} - f^{\text{lin}}_{\mathbf{W}(k)}\|_\infty \lesssim L^{\frac{7}{3}}(\eta T)^{\frac{2}{3}}m^{-\frac{1}{6}}$. Plugging the above observations back into (4.38), we know with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta/2$ over initialization $(\mathbf{a}, \mathbf{W}(0))$, there holds $\|\epsilon_k^1\|_{\mathcal{H}_m} \lesssim \frac{L^{\frac{7}{3}}\eta^{\frac{5}{3}}T^{\frac{2}{3}}}{m^{\frac{1}{6}}}$.

Putting the estimate of $\|\epsilon_s^1\|_{\mathcal{H}_m}$ back into (4.37) and setting $\lambda = (\eta T)^{-1}$, with a little abuse of notation (we regard $f^{\text{lin}}_{\mathbf{W}(k)}$ as a function in $\mathcal{L}^2_{\rho_x}$ in the following first term), the following inequality holds with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ over the

initialization $(\mathbf{a}, \mathbf{W}(0))$ and sampling

$$\|f^{\text{lin}}_{\mathbf{W}(k)} - \mathbf{S}_m g_k^m\|_\rho = \|\mathbf{S}_m F_k\|_\rho \leq 7\eta^{-\frac{1}{2}}\sqrt{T}\max_{s\in[T-1]}\|\epsilon_s^1\|_{\mathcal{H}_m} \lesssim \frac{L^{\frac{7}{3}}(\eta T)^{\frac{7}{6}}}{m^{\frac{1}{6}}}.$$

Since we assume that we are under the event $\{\|\mathbf{W}(0)\|_{op,\infty} \leq c_0\sqrt{m}\}$, whose probability is at least $1 - L\exp(-Cm)$ according to Lemma 17. Squaring both sides of the above inequality and combining all the high probability events complete the proof of the proposition. □

We now introduce some notations for our further analysis. For any $k \in \mathbb{N}$, we denote $\mathbf{G}_k^m = (g_k^m(x_1), \ldots, g_k^m(x_n))^\top \in \mathbb{R}^n$, $\mathbf{G}_k = (g_k(x_1), \ldots, g_k(x_n))^\top \in \mathbb{R}^n$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Recall that $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^n$ and $\mathbf{K}^m = (K^m(x_i, x_j))_{i,j=1}^n$ are the Gram matrices with kernels K and K^m, respectively. The following lemma shows that $\|\mathbf{G}_k^m - \mathbf{G}_k\|_2 \to 0$ as $m \to \infty$ for any $k \in [T]$.

LEMMA 29. *Let $\delta \in (0,1)$. Assume $m \gtrsim pL^3\log(m/\delta)$ and $\eta \leq 1/4$. Then, with probability at least $1-\delta$ over the initialization $(\mathbf{a}, \mathbf{W}(0))$, for any $k \in [T]$ there holds*

$$\left\|\mathbf{G}_k^m - \mathbf{G}_k\right\|_2 \leq \eta T\sqrt{n}\|K^m - K\|_\infty.$$

PROOF. According to (4.7) and (4.8), we know for any $k \in [T-1]$,

$$\mathbf{G}_{k+1}^m = \mathbf{G}_k^m - \frac{\eta}{n}\mathbf{K}^m(\mathbf{G}_k^m - \mathbf{y}) \text{ and } \mathbf{G}_{k+1} = \mathbf{G}_k - \frac{\eta}{n}\mathbf{K}(\mathbf{G}_k - \mathbf{y}). \tag{4.39}$$

Then, there holds

$$\begin{aligned}
\mathbf{G}_{k+1}^m - \mathbf{G}_{k+1} &= \mathbf{G}_k^m - \mathbf{G}_k - \frac{\eta}{n}\big(\mathbf{K}^m(\mathbf{G}_k^m - \mathbf{y}) - \mathbf{K}(\mathbf{G}_k - \mathbf{y})\big) \\
&= \mathbf{G}_k^m - \mathbf{G}_k - \frac{\eta}{n}\big(\mathbf{K}^m(\mathbf{G}_k^m - \mathbf{G}_k) - (\mathbf{K} - \mathbf{K}^m)(\mathbf{G}_k - \mathbf{y})\big) \\
&= \Big(\mathbf{I} - \frac{\eta}{n}\mathbf{K}^m\Big)(\mathbf{G}_k^m - \mathbf{G}_k) + \frac{\eta}{n}(\mathbf{K} - \mathbf{K}^m)(\mathbf{G}_k - \mathbf{y}).
\end{aligned}$$

Applying the above equality recursively, we have

$$\begin{aligned}\big\|\mathbf{G}_{k+1}^m - \mathbf{G}_{k+1}\big\|_2 &= \Big\|\frac{\eta}{n}\sum_{s=0}^{k}\Big(\mathbf{I}-\frac{\eta}{n}\mathbf{K}^m\Big)^s(\mathbf{K}-\mathbf{K}^m)(\mathbf{G}_{k-s}-\mathbf{y})\Big\|_2 \\ &\le \frac{\eta}{n}\sum_{s=0}^{k}\Big\|\mathbf{I}-\frac{\eta}{n}\mathbf{K}^m\Big\|_{op}^s\|\mathbf{K}-\mathbf{K}^m\|_{op}\|\mathbf{G}_{k-s}-\mathbf{y}\|_2. \end{aligned} \tag{4.40}$$

From Lemma 16 we know that

$$\|K^m\|_\infty = \sup_{x,x'\in\mathcal{X}}\Big|\Big\langle\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}^L(0)},\frac{\partial f_{\mathbf{W}(0)}(x')}{\partial\mathbf{W}^L(0)}\Big\rangle_2\Big| \le \sup_{x\in\mathcal{X}}\Big\|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial\mathbf{W}^L(0)}\Big\|_2^2 \le 4. \tag{4.41}$$

where the last inequality used Lemma 19 and condition $m \gtrsim pL^3\log(m/\delta)$. Then, for any $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)^\top \in \mathbb{R}^n$ with $\|\boldsymbol{\alpha}\|_2 = 1$, there holds $\boldsymbol{\alpha}^\top\mathbf{K}^m\boldsymbol{\alpha} = \|\sum_{i=1}^n \alpha_i K_{x_i}^m\|_{\mathcal{H}_m}^2 \le (\sum_{i=1}^n|\alpha_i|\|K_{x_i}^m\|_{\mathcal{H}_m})^2 \le (\sum_{i=1}^n|\alpha_i|\|K^m\|_\infty^{\frac{1}{2}})^2 \le 4(\sum_{i=1}^n|\alpha_i|)^2 \le 4n$. This implies that $\|\mathbf{K}^m\|_{op}\le 4n$. Since $\eta\le 1/4$ and $\mathbf{K}^m$ is PSD, we know $\|\mathbf{I}-\frac{\eta}{n}\mathbf{K}^m\|_{op}\le 1$.

Then, there holds

$$\begin{aligned}\|\mathbf{K}^m-\mathbf{K}\|_{op} &= \sup_{\|\boldsymbol{\alpha}\|_2=1}|\boldsymbol{\alpha}^\top(\mathbf{K}^m-\mathbf{K})\boldsymbol{\alpha}| = \sup_{\|\boldsymbol{\alpha}\|_2=1}\Big|\sum_{i,j=1}^n\alpha_i\alpha_j\big(K^m(x_i,x_j)-K(x_i,x_j)\big)\Big| \\ &\le \|K^m-K\|_\infty\sup_{\|\boldsymbol{\alpha}\|_2=1}\sum_{i,j=1}^n|\alpha_i\alpha_j| = \|K^m-K\|_\infty\sup_{\|\boldsymbol{\alpha}\|_2=1}\Big(\sum_{i=1}^n|\alpha_i|\Big)\Big(\sum_{j=1}^n|\alpha_j|\Big) \\ &\le n\|K^m-K\|_\infty.\end{aligned}$$

Further, from (4.39), we know $\mathbf{G}_k = (\mathbf{I}-\frac{\eta}{n}\mathbf{K})\mathbf{G}_{k-1}+\frac{\eta}{n}\mathbf{K}\mathbf{y}$. Recursively applying this equation, we get $\mathbf{G}_k = \frac{\eta}{n}\sum_{s=0}^{k-1}(\mathbf{I}-\frac{\eta}{n}\mathbf{K})^s\mathbf{K}\mathbf{y}$. Analogous to the estimate of $\|\mathbf{I}-\frac{\eta}{n}\mathbf{K}^m\|_{op}$, we can show that $\|\mathbf{K}\|_{op}\le n$ and $\|\mathbf{I}-\frac{\eta}{n}\mathbf{K}^m\|_{op}\le 1$ by noting $\|K\|_\infty\le 1$ (see Property 1). Then, there holds

$$\begin{aligned}\|\mathbf{G}_k\|_2 &\le \Big\|\sum_{s=0}^{k-1}\Big(\mathbf{I}-\frac{\eta}{n}\mathbf{K}\Big)^s\frac{\eta}{n}\mathbf{K}\Big\|_{op}\|\mathbf{y}\|_2 \le \sqrt{n}\sup_{t\in[0,1]}\Big|\sum_{s=0}^{k-1}(1-t)^s t\Big| \\ &= \sqrt{n}\sup_{t\in[0,1]}\big(1-(1-t)^k\big) \le \sqrt{n}.\end{aligned} \tag{4.42}$$

Plugging the above estimates back into (4.40), we have

$$\left\|\mathbf{G}_{k+1}^m - \mathbf{G}_{k+1}\right\|_2 \leq \frac{\eta}{n}\sum_{s=0}^{k}\|\mathbf{K}^m - \mathbf{K}\|_{op}\big(\|\mathbf{G}_{k-s}\|_2 + \|\mathbf{y}\|_2\big) \leq \eta T\sqrt{n}\|K^m - K\|_\infty,$$

which completes the proof. □

Based on the above lemma, we give the proof of Proposition 9 as follows.

PROOF OF PROPOSITION 9. For any $x \in \mathcal{X}$ and $k \in [T-1]$, from the definitions we know

$$\begin{aligned}
&|g_{k+1}^m(x) - g_{k+1}(x)| \\
&= \left|g_k^m(x) - g_k(x) - \frac{\eta}{n}\sum_{i=1}^{n}\big[\big(g_k^m(x_i) - g_k(x_i)\big)K^m(x_i,x) + \big(g_k(x_i) - y_i\big)\big(K^m(x_i,x) - K(x_i,x)\big)\big]\right| \\
&\leq \left|g_k^m(x) - g_k(x)\right| + \frac{\eta}{n}\sum_{i=1}^{n}\Big(\|K^m\|_\infty\left|g_k^m(x_i) - g_k(x_i)\right| + \|K^m - K\|_\infty\left|g_k(x_i) - y_i\right|\Big) \\
&\leq \left|g_k^m(x) - g_k(x)\right| + \frac{\eta}{\sqrt{n}}\big(\|K^m\|_\infty\|\mathbf{G}_k^m - \mathbf{G}_k\|_2 + \|K^m - K\|_\infty\|\mathbf{G}_k - \mathbf{y}\|_2\big),
\end{aligned}$$

where the last inequality used Cauchy-Schwarz inequality.

Combining Lemmas 29, (4.41) and (4.42) and with the above observation, we get

$$\|g_{k+1}^m - g_{k+1}\|_\infty \leq \|g_k^m - g_k\|_\infty + 6\eta^2 T\|K^m - K\|_\infty.$$

Applying the above inequality recursively and noting that $g_0^m = g_0$, we have

$$\|g_{k+1}^m - g_{k+1}\|_\infty \leq 6(\eta T)^2\|K^m - K\|_\infty.$$

From Lemma 20 and the condition (4.10) we know $\|K^m - K\|_\infty \lesssim \frac{\sqrt{L}}{m^{\frac{1}{6}}}$. Therefore, for any $k \in [T]$

$$\begin{aligned}
\left\|\mathbf{S}_m g_k^m - \mathbf{S}g_k\right\|_\rho^2 &= \int_{\mathcal{X}}|g_k^m(x) - g_k(x)|^2 d\rho_{\mathcal{X}}(x) \leq \|g_k^m - g_k\|_\infty^2 \\
&\leq 36(\eta T)^4\|K^m - K\|_\infty^2 \lesssim \frac{L(\eta T)^4}{m^{\frac{1}{3}}}.
\end{aligned}$$

The desired result is obtained by setting $k = T$. □

To estimate the last term $\|\mathbf{S}g_T - f_\rho\|_\rho^2$ in (4.9), we first introduce an intermediate term. Define the population iteration h_k on $\mathcal{H}_K$ as

$$h_{k+1} = h_k - \eta \int_{\mathcal{Z}} \big(\langle h_k, K_x\rangle_{\mathcal{H}_K} - y\big) K_x d\rho(z) \ \text{ with } \ h_0 = 0. \tag{4.43}$$

If we regard the population risk $\mathcal{E}(\cdot)$ as a functional on $\mathcal{H}_K$, then the population iteration h_k can be viewed as the GD of $\mathcal{E}(\cdot)$ initialized at $h_0 = 0$.

LEMMA 30. *Let $\mathcal{H}$ be the closure of $\mathcal{H}_K$ in $\mathcal{L}^2_{\rho_x}$. Then, Assumption 12 implies $f_\rho \in \mathcal{H}$.*

PROOF. Note that $\mathbf{L}$ has the eigen-decomposition $\mathbf{L}f = \sum_{i=1}^\infty \lambda_i \langle f, \Phi_i\rangle_{\mathcal{L}^2_{\rho_x}} \Phi_i$. According to Assumption 12, we know there exists a $g \in \mathcal{L}^2_{\rho_x}$ such that

$$f_\rho = \mathbf{L}^\beta g = \sum_{i=1}^\infty \lambda_i^\beta \langle g, \Phi_i\rangle_{\mathcal{L}^2_{\rho_x}} \Phi_i = \sum_{i:\lambda_i\neq 0}^\infty \lambda_i^\beta \langle g, \Phi_i\rangle_{\mathcal{L}^2_{\rho_x}} \Phi_i.$$

Since for any $\lambda_i \neq 0$, the associated eigenfunction $\Phi_i \in \mathcal{H}_K$ (see Chapter 4.5 in [Steinwart and Christmann 2008]), we conclude that $f_\rho \in \mathcal{H}$. □

LEMMA 31. *Suppose Assumptions 11 and 12 hold. Assume $\eta \leq 1$. For any $\delta_1, \delta_2 \in (0, 1/2)$, assume $\eta T \leq n(9\log(n/\delta_2))^{-1}$. Then, with probability at least $1 - \delta_1 - \delta_2$ over sampling, the following statements hold for all $k \in [T]$.*

(a) For the case $\beta \geq \frac{1}{2}$, there holds

$$\|\mathbf{S}g_k - \mathbf{S}h_k\|_\rho \leq 4(B+1)(12 + 4\log(k) + \sqrt{2}\eta)\Big(\frac{\sqrt{\eta k}}{n} + \sqrt{\frac{2c_\gamma(\eta k)^\gamma}{n}}\Big)\log\big(\frac{4}{\delta_1}\big).$$

(b) For the case $\beta \in (0, \frac{1}{2})$, there holds

$$\|\mathbf{S}g_T - \mathbf{S}h_T\|_\rho \leq (12 + 4\log(T) + \sqrt{2}\eta)\bigg(2(6+B)\Big(\frac{\sqrt{\eta T}}{n} + \sqrt{\frac{2c_\gamma(\eta T)^\gamma}{n}}\Big) + \frac{4B\big((\eta T)^{1-\beta} + 1\big)}{n}\bigg)\log\Big(\frac{3T}{\delta_1}\Big).$$

PROOF. The proof is derived from Theorem 5 in [Lin and Rosasco 2017], which provides upper bounds for $\|\mathcal{S}_\rho \nu_{k+1} - \mathcal{S}_\rho \mu_{k+1}\|_\rho$ with two iteration sequences $\{\nu_{k+1}\}$ and $\{\mu_{k+1}\}$. We

first show that their assumptions are satisfied in our setting, and then apply their results with our Lemma 28 by showing that $\mathcal{S}_\rho \nu_{k+1} - \mathcal{S}_\rho \mu_{k+1}$ is equivalent to $\mathbf{S}g_k - \mathbf{S}h_k$.

Since we assume $|y| \leq 1$, their Assumption 1 is satisfied with $M = v = 1$. Instead of using the notations $x, \langle x, x' \rangle_H$ and $\mathcal{S}_\rho$ in [Lin and Rosasco 2017] for any $x, x' \in \mathcal{X}$, we use K_x, $\langle K_x, K_{x'} \rangle_{\mathcal{H}_K}$ and $\mathbf{S}$ in our setting. Then, their H_ρ is the same as our $\mathcal{H}_K$. Since $f_\mathcal{H}$ in [Lin and Rosasco 2017] is the projection of f_ρ onto the closure of H_ρ in $\mathcal{L}^2_{\rho_x}$, from Lemma 30 we know their $f_\mathcal{H}$ is equivalent to our f_ρ. Hence, Assumption 2 in [Lin and Rosasco 2017] holds true with $\zeta = \beta$ and $R = B$ due to our Assumption 12. Further, their Assumption 3 is guaranteed by Assumption 11, their equation (3) holds true with $\kappa^2 = 1$ due to $\langle K_x, K_{x'} \rangle_{\mathcal{H}_K} = K(x, x') \leq \|K\|_\infty \leq 1$ (see Property 1). Their equation (47) is guaranteed by Lemma 28 with $\kappa = 1$, $\Gamma = n$, $\delta = \delta_2$, $\zeta_i = K_{x_i}$, $Q = \int_\mathcal{X} K_x \otimes K_x d\rho_x$. In addition, by taking the step-size $\eta_k = \eta$ for all $k \in [T]$, we know $\mathcal{S}_\rho \nu_{k+1} - \mathcal{S}_\rho \mu_{k+1}$ in [Lin and Rosasco 2017] is equivalent to our $\mathbf{S}g_k - \mathbf{S}h_k$.

Then, combining above observations and Theorem 5 in [Lin and Rosasco 2017] with $\eta_k = \eta$, $\theta = 0$, $\lambda = (\eta k)^{-1}$, $\kappa = 1$, $M = v = 1$, $R = B$, $\zeta = \beta$ and $m = n$, we get the desired results. □

LEMMA 32 (Proposition 2 in [Lin and Rosasco 2017]). *Suppose Assumption 12 holds. Let $\eta \in (0, 1]$ be the step size. For any $k \in \mathbb{N}$, there holds*

$$\|\mathbf{S}h_k - f_\rho\|_\rho \leq B\Big(\frac{\beta}{2\eta k}\Big)^\beta.$$

PROOF. In the proof of Lemma 31, we already showed that $\mathcal{S}_\rho \mu_{k+1}$ and $f_\mathcal{H}$ in [Lin and Rosasco 2017] are equivalent to our $\mathbf{S}h_k$ and f_ρ. Then, by applying Proposition 2 in [Lin and Rosasco 2017] with $\eta_k = \eta$, $\kappa = 1$, $R = B$ and $\zeta = \beta$, we get the desired results. □

Combining Lemma 31 and Lemma 32, we give the proof of Proposition 10.

PROOF OF PROPOSITION 10. Note that $\|\mathbf{S}g_T - f_\rho\|^2_\rho \lesssim \|\mathbf{S}g_T - \mathbf{S}h_T\|^2_\rho + \|\mathbf{S}h_T - f_\rho\|^2_\rho$. The desired results are obtained by combining Lemma 31 with $\delta_1 = \delta_2 = \frac{\delta}{2}$ and Lemma 32. □

Now, we give proofs for Theorem 9 and Corollary 1.

PROOF OF THEOREM 9. Combining Propositions 7, 8, 9 and 10 with δ replaced by $\frac{\delta}{4}$, with probability at least $1-L\exp\big(\mathcal{O}(pL\log(m))-\Omega(m^{\frac{1}{3}})\big)-\delta$ over initialization $(\mathbf{a},\mathbf{W}(0))$ and sampling, there holds

$$\mathcal{E}(f_{\mathbf{W}(T)})-\mathcal{E}(f_\rho)\lesssim\frac{L^{\frac{14}{3}}(\eta T)^4}{m^{\frac{1}{3}}}+\Big(\frac{\eta T}{n^2}+\frac{(\eta T)^\gamma+(\eta T)^{1-2\beta}}{n}\Big)\log^4\big(\frac{T}{\delta}\big)+(\eta T)^{-2\beta}.$$

The proof of the theorem is completed. □

PROOF OF COROLLARY 1. The proof is derived by Theorem 9 with δ replaced by $\delta/2$. We first prove that the condition $n\geq\frac{16}{\delta}\big(\frac{36(2\beta+\gamma)}{\beta}\big)^{\frac{2\beta+\gamma}{\beta}}$ implies $\eta T\leq\frac{n}{36\log(16n/\delta)}$. Since $\eta T\leq 2n^{\frac{1}{2\beta+\gamma}}$, the condition reduces to show $n^{\frac{2\beta}{2\beta+\gamma}}\geq 72\log(\frac{16n}{\delta})$, which is equivalent to showing $\big(\frac{16n}{\delta}\big)^{\frac{2\beta}{2\beta+\gamma}}\geq\frac{36(2\beta+\gamma)}{\beta}\big(\frac{16}{\delta}\big)^{\frac{2\beta}{2\beta+\gamma}}\log\big(\frac{16n}{\delta}\big)^{\frac{2\beta}{2\beta+\gamma}}$. From (9.17) and (9.18) in Györfi et al. 2006 we know $u>2c\log(c)$ implies $u>c\log(u)$ for any $c\geq e$. Setting $u=\big(\frac{16n}{\delta}\big)^{\frac{2\beta}{2\beta+\gamma}}$ and $c=\frac{36(2\beta+\gamma)}{\beta}\big(\frac{16}{\delta}\big)^{\frac{2\beta}{2\beta+\gamma}}$ and solving $u\geq c^2$, the desired result is obtained by noting $u\geq c^2>2c\log(c)$ for all $c\geq e$. Combining this with $T=\lceil n^{\frac{1}{2\beta+\gamma}}\rceil$, we know $n\geq\max\big\{\big(\frac{36(2\beta+\gamma)}{\beta}\big)^{\frac{2\beta+\gamma}{\beta}}\frac{16}{\delta},\eta^{-(2\beta+\gamma)}\big\}$ implies $1\leq\eta T\leq n(36\log(16n/\delta))^{-1}$. Similarly, setting $u=(m/\delta)^{\frac{1}{3}}$ and $c=3\big(L^{22}p^2(\eta T)^7/\delta\big)^{\frac{1}{3}}$, and noting $\eta T\asymp n^{\frac{1}{2\beta+\gamma}}$, we know $m\gtrsim L^{22}p^2n^{\frac{7}{2\beta+\gamma}}\log^3(npL/\delta)$ ensures condition (4.10) in Theorem 9.

Note $m\gtrsim L^{14}n^{\frac{6\beta+12}{2\beta+\gamma}}$ ensures $\frac{L^{\frac{14}{3}}(\eta T)^4}{m^{\frac{1}{3}}}\lesssim n^{-\frac{2\beta}{2\beta+\gamma}}$ and (4.10) implies $L\exp\big(\mathcal{O}(pL\log(m))-\Omega(m^{\frac{1}{3}})\big)\leq\delta/2$. Then, from Theorem 9 we know with probability at least $1-\delta$ over initialization $(\mathbf{a},\mathbf{W}(0))$ and sampling, there holds

$$\mathcal{E}(f_{\mathbf{W}(T)})-\mathcal{E}(f_\rho)\lesssim n^{-\frac{2\beta}{2\beta+\gamma}}+\Big(\frac{\eta T}{n^2}+\frac{(\eta T)^\gamma+(\eta T)^{1-2\beta}}{n}\Big)\log^2(T)\log^2\big(\frac{T}{\delta}\big)+(\eta T)^{-2\beta}.$$

In addition, since $2\beta+\gamma>1$ and $\eta T\geq 1$, there holds $(\eta T)^{1-2\beta}\leq(\eta T)^\gamma$. Plugging the choice of $\eta T\asymp n^{\frac{1}{2\beta+\gamma}}$ back into the above inequality, we get

$$\mathcal{E}(f_{\mathbf{W}(T)})-\mathcal{E}(f_\rho)\lesssim n^{-\frac{2\beta}{2\beta+\gamma}}\log^4(\frac{n}{\delta}).$$

The proof is completed. □

4.4.4 Proofs for Stochastic Gradient Descent

We first show that the trajectory of SGD with deep ReLU networks also falls inside local balls around the initialization $\mathbf{W}(0)$.

LEMMA 33. *Let* $\{\mathbf{W}(k)\}$ *be produced by* (4.4) *with* $\eta \leq 1/5$. *Assume* (4.13) *holds. Then, with probability at least* $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ *over initialization* $(\mathbf{a}, \mathbf{W}(0))$, *for any* $k \in [T]$, *there holds*

$$\big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_{op,\infty}^2 \leq \big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_2^2 \leq 4\eta k$$

and

$$|f_{\mathbf{W}(k)}(x) - y| \leq CL^2\sqrt{\eta k} + 1 \text{ for any } z = (x, y) \in \mathcal{Z}.$$

PROOF. The first part of the lemma is proved by induction. It's obvious that $\|\mathbf{W}(k) - \mathbf{W}(0)\|_2^2 \leq 0$ holds with $k = 0$. Assume, for all $t \in [k]$ with $k \leq T-1$, $\|\mathbf{W}(k) - \mathbf{W}(0)\|_2^2 \leq 4\eta k$ holds. We will show that $\|\mathbf{W}(k+1) - \mathbf{W}(0)\|_2^2 \leq 4\eta(k+1)$.

From the update rule (4.4), we know

$$\begin{aligned}
&\big\|\mathbf{W}(k+1) - \mathbf{W}(0)\big\|_2^2 = \Big\|\mathbf{W}(k) - \mathbf{W}(0) - \eta\frac{\partial \ell(\mathbf{W}(k); z_{i_k})}{\partial \mathbf{W}(k)}\Big\|_2^2 \\
&= \big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_2^2 + \eta^2\Big\|\frac{\partial \ell(\mathbf{W}(k); z_{i_k})}{\partial \mathbf{W}(k)}\Big\|_2^2 + 2\eta\Big\langle \mathbf{W}(0) - \mathbf{W}(k), \frac{\partial \ell(\mathbf{W}(k); z_{i_k})}{\partial \mathbf{W}(k)}\Big\rangle_2 \\
&= \big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_2^2 + 2\eta^2\ell(\mathbf{W}(k), z_{i_k})\Big\|\frac{\partial f_{\mathbf{W}(k)}(x_{i_k})}{\partial \mathbf{W}(k)}\Big\|_2^2 \\
&\quad + 2\eta\Big\langle \mathbf{W}(0) - \mathbf{W}(k), \frac{\partial \ell(\mathbf{W}(k); z_{i_k})}{\partial \mathbf{W}(k)}\Big\rangle_2,
\end{aligned} \tag{4.44}$$

where in the last inequality we have used $\frac{\partial \ell(\mathbf{W}(k); z_{i_k})}{\partial \mathbf{W}(k)} = (f_{\mathbf{W}(k)}(x_{i_k}) - y_{i_k})\frac{\partial f_{\mathbf{W}(k)}(x_{i_k})}{\partial \mathbf{W}(k)}$ and $(f_{\mathbf{W}(k)}(x_{i_k}) - y_{i_k})^2 = 2\ell(\mathbf{W}(k); z_{i_k})$.

Setting $R_{op} = 2\sqrt{\eta T}$. By the induction assumption, there holds $\mathbf{W}(k), \mathbf{W}(0) \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$. Then from Lemma 16 (if $l < L-1$) and part (c) of Lemma 19 (if $l = L$) and (4.27) in Lemma

26 with $\mathbf{W} = \mathbf{W}(k)$, we have

$$\Big\|\frac{\partial f_{\mathbf{W}(k)}(\mathbf{x}_{i_k})}{\partial \mathbf{W}(k)}\Big\|_2 \leq \Big\|\frac{\partial f_{\mathbf{W}(k)}(\mathbf{x}_{i_k})}{\partial \mathbf{W}(k)} - \frac{\partial f_{\mathbf{W}(0)}(\mathbf{x}_{i_k})}{\partial \mathbf{W}(0)}\Big\|_2 + \Big\|\frac{\partial f_{\mathbf{W}(0)}(\mathbf{x}_{i_k})}{\partial \mathbf{W}^L(0)}\Big\|_2 \tag{4.45}$$

$$\leq \sqrt{L}\max_{\ell\in[L]}\Big\|\frac{\partial f_{\mathbf{W}(k)}(\mathbf{x}_{i_k})}{\partial \mathbf{W}^\ell(k)} - \frac{\partial f_{\mathbf{W}(0)}(\mathbf{x}_i)}{\partial \mathbf{W}^\ell(0)}\Big\|_2 + 2$$

$$\leq \epsilon_3 + 2 \tag{4.46}$$

with $\epsilon_3 \lesssim L^{\frac{7}{3}}(\eta T)^{\frac{1}{6}} m^{-\frac{1}{6}}$.

Further, from the induction assumption we know $|f_{\mathbf{W}(k)}(x) - y| \leq CL^2 R_{op}$ and (4.26) in Lemma 26 with $\mathbf{W} = \mathbf{W}(k)$, $\widetilde{\mathbf{W}} = \mathbf{W}(0)$ and $\|\mathbf{W}(k) - \mathbf{W}(0)\|_{op,\infty} \leq 2\sqrt{\eta T}$ implies

$$2\eta\Big\langle \mathbf{W}(0) - \mathbf{W}(k), \frac{\partial \ell(\mathbf{W}(k); z_{i_k})}{\partial \mathbf{W}(k)}\Big\rangle_2 \leq 2\eta\big(\ell(\mathbf{W}(0), z_{i_k}) - \ell(\mathbf{W}(k), z_{i_k})\big) + 2\eta\epsilon_2$$

with $\epsilon_2 \lesssim L^{\frac{13}{3}}(\eta T)^{\frac{7}{6}} m^{-\frac{1}{6}}$.

Plugging the above two estimates back into (4.44), we get

$$\begin{aligned}
&\big\|\mathbf{W}(k+1) - \mathbf{W}(0)\big\|_2^2 \\
&\leq \big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_2^2 + 2\eta^2 \ell(\mathbf{W}(k), z_{i_k})(\epsilon_3 + 2)^2 + 2\eta\big(\ell(\mathbf{W}(0), z_{i_k}) - \ell(\mathbf{W}(k), z_{i_k})\big) + 2\eta\epsilon_2 \\
&\leq \big\|\mathbf{W}(k) - \mathbf{W}(0)\big\|_2^2 + 10\eta^2 \ell(\mathbf{W}(k), z_{i_k}) + 2\eta\big(\ell(\mathbf{W}(0), z_{i_k}) - \ell(\mathbf{W}(k), z_{i_k})\big) + 2\eta \\
&\leq \big\|\mathbf{W}^l(k) - \mathbf{W}(0)\big\|_2^2 + 2\eta\ell(\mathbf{W}(0), z_{i_k}) + 2\eta \\
&\leq 4\eta k + 3\eta \leq 4\eta(k+1),
\end{aligned}$$

where in the second inequality we have used $\epsilon_3 \leq 5 - \sqrt{2}$ and $\epsilon_2 \leq 1$ implied by (4.13), in the third inequality we have used $10\eta^2 \leq 2\eta$ by noting $\eta \leq 1/5$, in the last second inequality we have used $\ell(\mathbf{W}(0), z_{i_k}) \leq 1/2$ by observing $f_{\mathbf{W}(0)} = 0$ and the induction assumption $\|\mathbf{W}(k) - \mathbf{W}(0)\|_2^2 \leq 4\eta k$. The first part of the lemma is proved.

Combining Lemma 21 with $\|\mathbf{W}(k)-\mathbf{W}(0)\|_{op,\infty}^2 \le \|\mathbf{W}(k)-\mathbf{W}(0)\|_2^2 \le 4\eta T = R_{op}^2$, we know

$$\begin{aligned}|f_{\mathbf{W}(k)}(x)-y| &\le |f_{\mathbf{W}(k)}(x)-f_{\mathbf{W}(0)}(x)| + |f_{\mathbf{W}(0)}(x)-y| \le \|\mathbf{a}\|_2\|o_k^L(x)-o_0^L(x)\|_2+1\\ &\le CL^2\sqrt{\eta k}+1,\end{aligned}$$

which completes the second part of the lemma. □

The proof of Proposition 11 is presented as follows.

PROOF OF PROPOSITION 11. The proof is similar to that of Proposition 7. Setting $R_{op} = 2\sqrt{\eta T}$. Combining Lemma 33 and (4.25) in Lemma 26 with $\widetilde{\mathbf{W}} = \mathbf{W}(k)$ and $\mathbf{W} = \mathbf{W}(0)$, we get the desired results. □

Based on Lemma 33, we give the proof of Proposition 12 as follows.

PROOF OF PROPOSITION 12. Denote $\epsilon_k = f^{\text{lin}}_{\mathbf{W}(k+1)} - f^{\text{lin}}_{\mathbf{W}(k)} + \eta(f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - y_{i_k})K^m_{x_{i_k}} \in \mathcal{H}_m$. From the update rule of f_k^m (4.11), we know

$$\begin{aligned}f^{\text{lin}}_{\mathbf{W}(k+1)} - f^m_{k+1} &= \big(f^{\text{lin}}_{\mathbf{W}(k)} - f^m_k\big) - \eta\big(f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - f^m_k(x_{i_k})\big)K^m_{x_{i_k}} + \epsilon_k\\ &= \big(f^{\text{lin}}_{\mathbf{W}(k)} - f^m_k\big) - \eta\big\langle f^{\text{lin}}_{\mathbf{W}(k)} - f^m_k, K^m_{x_{i_k}}\big\rangle_{\mathcal{H}_m} K^m_{x_{i_k}} + \epsilon_k\\ &= \big(\mathbf{I} - \eta K^m_{x_{i_k}} \otimes K^m_{x_{i_k}}\big)\big(f^{\text{lin}}_{\mathbf{W}(k)} - f^m_k\big) + \epsilon_k,\end{aligned}$$

where the second equality follows from the fact $f^{\text{lin}}_{\mathbf{W}(k)} - f^m_k \in \mathcal{H}_m$ and the reproducing kernel property $f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - f^m_k(x_{i_k}) = \langle f^{\text{lin}}_{\mathbf{W}(k)} - f^m_k, K^m_{x_{i_k}}\rangle_{\mathcal{H}_m}$.

Applying the above equality recursively, we get

$$f^{\text{lin}}_{\mathbf{W}(k+1)} - f^m_{k+1} = \sum_{s=0}^{k}\prod_{a=s+1}^{k}\big(\mathbf{I} - \eta K^m_{x_{i_a}} \otimes K^m_{x_{i_a}}\big)\epsilon_s,$$

where we used the conventional notation $\prod_{k+1}^k = \mathbf{I}$ for any $k \in \mathbb{N}$. Note that for any $a \in [k]$ and i_a, $\eta K^m_{x_{i_a}} \otimes K^m_{x_{i_a}}$ is self-adjoint and positive, and from (4.41) we know $\eta\|K^m_{x_{i_a}} \otimes K^m_{x_{i_a}}\|_{op} = \eta\|K^m_{x_{i_a}}\|^2_{\mathcal{H}_m} \le \eta\|K^m\|_\infty \le 4\eta \le 1$. Then, $\|\mathbf{I} - \eta K^m_{x_{i_a}} \otimes K^m_{x_{i_a}}\|_{op} \le 1$.

According to the above inequality, we have

$$\begin{aligned}
&\big\|f^{\text{lin}}_{\mathbf{W}(k+1)} - f^m_{k+1}\big\|_\infty = \sup_{x\in\mathcal{X}} \big\langle f^{\text{lin}}_{\mathbf{W}(k+1)} - f^m_{k+1}, K^m_x\big\rangle_{\mathcal{H}_m} \\
&\leq \sup_{x\in\mathcal{X}} \big\|f^{\text{lin}}_{\mathbf{W}(k+1)} - f^m_{k+1}\big\|_{\mathcal{H}_m} \|K^m_x\|_{\mathcal{H}_m} \\
&\leq \big\|f^{\text{lin}}_{\mathbf{W}(k+1)} - f^m_{k+1}\big\|_{\mathcal{H}_m} \sqrt{\|K^m\|_\infty} \\
&\leq 2\sum_{s=0}^{k} \prod_{a=s+1}^{k} \big\|\mathbf{I} - \eta K^m_{x_{i_a}} \otimes K^m_{x_{i_a}}\big\|_{op} \big\|\epsilon_s\big\|_{\mathcal{H}_m} \leq 2\sum_{s=0}^{k} \big\|\epsilon_s\big\|_{\mathcal{H}_m},
\end{aligned} \tag{4.47}$$

where in the first equality we have used the reproducing kernel property and in the last second inequality we have used (4.41) with $\sqrt{\|K^m\|_\infty} \leq 2$.

Now, we turn to estimate $\|\epsilon_k\|_{\mathcal{H}_m}$. For any $k \in [T]$, from the definition of $f^{\text{lin}}_{\mathbf{W}}$ and the update rule of SGD (4.4), there holds

$$\begin{aligned}
&\epsilon_k(x) = f^{\text{lin}}_{\mathbf{W}(k+1)}(x) - f^{\text{lin}}_{\mathbf{W}(k)}(x) + \eta\big(f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - y_{i_k}\big) K^m_{x_{i_k}}(x) \\
&= \Big\langle \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}(0)}, \mathbf{W}(k+1) - \mathbf{W}(k)\Big\rangle_2 + \eta\big(f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - y_{i_k}\big)\Big\langle \frac{\partial f_{\mathbf{W}(0)}(x_{i_k})}{\partial \mathbf{W}(0)}, \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}(0)}\Big\rangle_2 \\
&= \Big\langle \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}, \mathbf{W}^L(k+1) - \mathbf{W}^L(k)\Big\rangle_2 + \eta\big(f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - y_{i_k}\big)\Big\langle \frac{\partial f_{\mathbf{W}(0)}(x_{i_k})}{\partial \mathbf{W}^L(0)}, \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\rangle_2 \\
&= \eta\Big[\big(y_{i_k} - f_{\mathbf{W}(k)}(x_{i_k})\big)\Big\langle \frac{\partial f_{\mathbf{W}(k)}(x_{i_k})}{\partial \mathbf{W}^L(k)}, \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\rangle_2 \\
&\qquad + \big(f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - y_{i_k}\big)\Big\langle \frac{\partial f_{\mathbf{W}(0)}(x_{i_k})}{\partial \mathbf{W}^L(0)}, \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\rangle_2\Big] \\
&= \Big\langle \eta\Big[\big(y_{i_k} - f_{\mathbf{W}(k)}(x_{i_k})\big)\frac{\partial f_{\mathbf{W}(k)}(x_{i_k})}{\partial \mathbf{W}^L(k)} + \big(f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - y_{i_k}\big)\frac{\partial f_{\mathbf{W}(0)}(x_{i_k})}{\partial \mathbf{W}^L(0)}\Big], \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\rangle_2 \\
&=: \Big\langle \Delta(k), \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\rangle_2,
\end{aligned}$$

where the second equality is due to $K^m_{x_{i_k}}(x) = K^m(x_{i_k}, x)$, the third equality used $\frac{\partial f_{\mathbf{W}(0)}}{\partial \mathbf{W}^l(0)} = 0$ for $l \in [L-1]$ according to Lemma 16, and the fourth equality is according to the update

rule (4.4), and in the last equality we define

$$\Delta(k) := \eta\Big[\big(y_{i_k} - f_{\mathbf{W}(k)}(x_{i_k})\big)\frac{\partial f_{\mathbf{W}(k)}(x_{i_k})}{\partial \mathbf{W}^L(k)} + \big(f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - y_{i_k}\big)\frac{\partial f_{\mathbf{W}(0)}(x_{i_k})}{\partial \mathbf{W}^L(0)}\Big]$$
$$= \eta\Big[\big(y_{i_k} - f_{\mathbf{W}(k)}(x_{i_k})\big)\Big(\frac{\partial f_{\mathbf{W}(k)}(x_{i_k})}{\partial \mathbf{W}^L(k)} - \frac{\partial f_{\mathbf{W}(0)}(x_{i_k})}{\partial \mathbf{W}^L(0)}\Big) + \big(f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - f_{\mathbf{W}(k)}(x_{i_k})\big)\frac{\partial f_{\mathbf{W}(0)}(x_{i_k})}{\partial \mathbf{W}^L(0)}\Big].$$

Let $\mathbf{\Delta}(k) = (0, \ldots, 0, \Delta(k)) \in \mathcal{W}$, then $\epsilon_k(x) = \langle \mathbf{\Delta}(k), \Phi_m(x)\rangle_2$. There holds

$$\begin{aligned}
&\|\epsilon_k\|_{H_m} \leq \|\Delta(k)\|_2 \\
&\leq \eta\Big[\big|y_{i_k} - f_{\mathbf{W}(k)}(x_{i_k})\big|\Big\|\frac{\partial f_{\mathbf{W}(k)}(x_{i_k})}{\partial \mathbf{W}^L(k)} - \frac{\partial f_{\mathbf{W}(0)}(x_{i_k})}{\partial \mathbf{W}^L(0)}\Big\|_2 + \big|f^{\text{lin}}_{\mathbf{W}(k)}(x_{i_k}) - f_{\mathbf{W}(k)}(x_{i_k})\big| \\
&\qquad \times \Big\|\frac{\partial f_{\mathbf{W}(0)}(x_{i_k})}{\partial \mathbf{W}^L(0)}\Big\|_2\Big] \\
&\leq \eta\Big(\sup_{z\in\mathcal{Z}}|f_{\mathbf{W}(k)}(x) - y| \sup_{x\in\mathcal{X}}\Big\|\frac{\partial f_{\mathbf{W}(k)}(x)}{\partial \mathbf{W}^L(k)} - \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\|_2 + \big\|f^{\text{lin}}_{\mathbf{W}(k)} - f_{\mathbf{W}(k)}\big\|_\infty \\
&\qquad \times \sup_{x\in\mathcal{X}}\Big\|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\|_2\Big). \qquad (4.48)
\end{aligned}$$

From part (c) in Lemma 19 we know $\sup_{x\in\mathcal{X}}\|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\|_2 \leq 2$. Setting $R_{op} = 2\sqrt{\eta T}$, from Lemma 33 we know $\sup_{z\in\mathcal{Z}}|f_{\mathbf{W}(k)}(x) - y| \leq CL^2\sqrt{\eta k} + 1$ and $\mathbf{W}(k) \in \mathcal{B}_{R_{op}}(\mathbf{W}(0))$ for any $k \in [T]$. Combining this and Lemma 26 with $R_{op} = 2\sqrt{\eta T}$, there holds

$$\sup_{z\in\mathcal{Z}}|f_{\mathbf{W}(k)}(x) - y| \sup_{x\in\mathcal{X}}\Big\|\frac{\partial f_{\mathbf{W}(k)}(x)}{\partial \mathbf{W}^L(k)} - \frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^L(0)}\Big\|_2 \lesssim \frac{L^{\frac{10}{3}}(\eta T)^{\frac{2}{3}}}{m^{\frac{1}{6}}}.$$

According to Proposition 11, we know $\|f_{\mathbf{W}(k)} - f^{\text{lin}}_{\mathbf{W}(k)}\|_\infty \lesssim L^{\frac{7}{3}}(\eta T)^{\frac{2}{3}}m^{-\frac{1}{6}}$. Plugging the above estimates back into (4.48), we know with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ over initialization $(\mathbf{a}, \mathbf{W}(0))$, there holds

$$\|\epsilon_k\|_{\mathcal{H}_m} \lesssim \frac{L^{\frac{10}{3}}\eta^{\frac{5}{3}}T^{\frac{2}{3}}}{m^{\frac{1}{6}}}.$$

Plugging the estimate of $\|\epsilon_s\|_{\mathcal{H}_m}$ back into (4.47), we know with probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})) - \delta$ over the random choice of $(\mathbf{a}, \mathbf{W}(0))$, there holds

$$\|f^{\text{lin}}_{\mathbf{W}(k)} - f^m_k\|_\infty \leq \sum_{s=0}^{k}\|\epsilon_s\|_{\mathcal{H}_m} \lesssim \frac{L^{\frac{10}{3}}(\eta T)^{\frac{5}{3}}}{m^{\frac{1}{6}}}.$$

This completes the proof of the proposition. □

To control $\|\mathbf{S}_m f_T^m - f_\rho\|_\rho^2$, we first control $\|\mathbf{S}_m f_T^m - \mathbf{S}_m g_T^m\|_\rho^2$, i.e., the distance between the SGD and GD on $\mathcal{H}_m$.

LEMMA 34. *Let $\delta \in (0,1)$ and $T \in \mathbb{N}$. Suppose $0 < \eta \leq \frac{1}{32(\log(T)+1)}$ and $\frac{1}{\eta T} \geq \frac{36}{n}\log\left(\frac{2n}{\delta}\right)$, and $m \gtrsim pL^3\log^3(m/\delta)$, Then, with probability at least $1 - L\exp\left(\mathcal{O}(d\log(m)) - \Omega(m^{\frac{1}{3}})\right) - \delta$ over initialization $(\mathbf{W}(0), \mathbf{a})$ and sampling, there holds*

$$\mathbb{E}_{\mathcal{A}}\left[\left\|\mathbf{S}_m(f_T^m - g_T^m)\right\|_\rho^2\right] \lesssim \eta(\log(T) \vee 1).$$

PROOF. The lemma is proved by using Proposition 6 in [Lin and Rosasco 2017], which provides upper bounds for $\|\mathcal{S}_\rho \omega_{T+1} - \mathcal{S}_\rho \nu_{T+1}\|_\rho$. We first show that their assumptions are satisfied in our setting, and then apply their results with our Lemma 28 by showing that $\mathcal{S}_\rho \nu_{T+1} - \mathcal{S}_\rho \mu_{T+1}$ is equivalent to $\mathbf{S}_m f_T^m - \mathbf{S}_m g_T^m$.

Note we assume $|y| \leq 1$, then their Assumption 1 is satisfied with $M = v = 1$. Instead of using the notations $x, \langle x, x'\rangle_H$ and $\mathcal{S}_\rho$ in [Lin and Rosasco 2017] for any $x, x' \in \mathcal{X}$, we use K_x^m, $\langle K_x^m, K_{x'}^m\rangle_{\mathcal{H}_m}$ and $\mathbf{S}_m$ in our setting. With probability at least $1 - \delta/2$ over the random choice of $\mathbf{W}(0)$, their equation (3) holds true with $\kappa^2 = 4$ due to $\langle K_x^m, K_{x'}^m\rangle_{\mathcal{H}_m} = K^m(x, x') \leq \|K^m\|_\infty \leq 4$ according to (4.41). Their equation (47) is guaranteed with probability at least $1 - \delta/2$ over sampling by Lemma 28 with $\kappa = 2$, $\Gamma = n$, $\zeta_i = K_{x_i}^m$, $Q = \int_{\mathcal{X}} K_x^m \otimes K_x^m d\rho_x$, and $\lambda = (\eta T)^{-1}$. In addition, by taking the batch-size $b = 1$ and the step size $\eta_k = \eta$ for all $k \in [T]$, we know $\mathcal{S}_\rho \nu_{T+1} - \mathcal{S}_\rho \mu_{T+1}$ in [Lin and Rosasco 2017] is equivalent to our $\mathbf{S}_m f_T^m - \mathbf{S}_m g_T^m$.

Then, combining above observations and Proposition 6 in [Lin and Rosasco 2017] with $\eta_k = \eta$, $\theta = 0$, $\lambda = (\eta T)^{-1}$, $\kappa = 2$, $M = v = 1$ and $b = 1$, we get the desired results. □

Now, we present the proof of Proposition 13.

PROOF OF PROPOSITION 13. Note that

$$\mathbb{E}_{\mathcal{A}}[\|\mathbf{S}_m f_T^m - f_\rho\|_\rho^2] \lesssim \mathbb{E}_{\mathcal{A}}[\|\mathbf{S}_m f_T^m - \mathbf{S}_m g_T^m\|_\rho^2] + \|\mathbf{S}_m g_T^m - \mathbf{S} g_T\|_\rho^2 + \|\mathbf{S} g_T - f_\rho\|_\rho^2.$$

Then, the desired results are obtained by combining Lemma 34 with δ replaced by $\delta/3$, Proposition 9with δ replaced by $\delta/3$, and Proposition 10 with δ replaced by $\delta/3$. □

PROOF OF THEOREM 10. Combining Propositions 11, 12 and 13 with δ replaced by $\delta/3$, the desired result is obtained. □

PROOF OF COROLLARY 2. It is easy to show that the inequality $\eta T \leq n(36\log(24n/\delta))^{-1}$ holds for $\eta = (72\log(24n/\delta))^{-1} n^{-\frac{2\beta}{2\beta+\gamma}}$ and $T = \lceil n^{\frac{2\beta+1}{2\beta+\gamma}} \rceil$. Similar to the proof of Corollary 1, one can check that $n \geq (72(2\beta+\gamma))^{2(2\beta+\gamma)}(\frac{24}{\delta})$ implies $\eta T \geq 1$ and $\eta \leq \frac{1}{32(\log(T)+1)}$. Note that the choices of η and T implies $\eta\log(T) + (\frac{\eta T}{n^2} + \frac{(\eta T)^\gamma + (\eta T)^{1-2\beta}}{n})\log^2(T)\log^2(\frac{T}{\delta}) + (\eta T)^{-2\beta} \lesssim n^{-\frac{2\beta}{2\beta+\gamma}}\log^2(n)\log^{2\beta}(\frac{n}{\delta})$. Further, according to the proof of Corollary 1, one can also show that $m \gtrsim L^{20}\max\{L^6 p^3 n^{\frac{7}{2\beta+\gamma}}\log^3(npL/\delta), n^{\frac{6\beta+12}{2\beta+\gamma}}\}$ indicates (4.13) and $\frac{L^{\frac{20}{3}}(\eta T)^4}{m^{\frac{1}{3}}} \lesssim n^{-\frac{2\beta}{2\beta+\gamma}}$. In addition, note that (4.13) implies $L\exp\big(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}})\big) \leq \delta/2$. Combining the above observations with Theorem 10 with δ replaced by $\delta/2$ yields the desired results. □

4.4.5 The NTK for Deep ReLU Networks with Non-Symmetric Initialization

In this section, we discuss the uniform concentration of the NTK with non-symmetric initialization, i.e., we consider the following initialization

$$\textit{for the first layer: } \mathbf{w}_r^1(0) \sim \mathcal{N}(0, \mathbf{I}_p), \ \textit{for } l = 2, \ldots, L: \ \mathbf{w}_r^l(0) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathbf{I}_m) \textit{ for all } r \in [m],$$
$$\textit{for the output layer: } a_r \overset{\text{i.i.d.}}{\sim} \textit{Unif}\,(\{-1, 1\}) \textit{ for } r \in [m]. \tag{4.49}$$

Indeed, the symmetric setting can be seen as a special case of the above general setting. We will show that for this general setting, the results of Lemma 20 still holds with the

same convergence rates. Note that K^m and K are different between the symmetric and non-symmetric settings. We first give their definitions as follows.

As discussed in [Jacot et al. 2018; Xu and Zhu 2024], the NTK $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for deep ReLU networks with initialization (4.49) is defined, for any $x, x' \in \mathcal{X}$, by

$$K(x, x') = \sum_{l=1}^{L} K^l(x, x') = \sum_{l=1}^{L} 2\mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))] \prod_{h=l}^{L} q^h(x, x'), \tag{4.50}$$

where $(U^l(x), U^l(x'))$ is a pair of bivariate normal variables defined iteratively by

$$(U^l(x), U^l(x')) \sim \mathcal{N}(0, \Sigma^{l-1}(x, x')) \tag{4.51}$$

with

$$\Sigma^{l-1}(x, x') = 2 \begin{pmatrix} \mathbb{E}[\sigma^2(U^{l-1}(x))] & \mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))] \\ \mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))] & \mathbb{E}[\sigma^2(U^{l-1}(x'))] \end{pmatrix}$$

and

$$\Sigma^0(x, x') = \begin{pmatrix} 1 & \langle x, x' \rangle_2 \\ \langle x, x' \rangle_2 & 1 \end{pmatrix},$$

and $q^l(x, x') = \frac{\pi - \arccos(p^{l-1}(x,x'))}{\pi}$ with

$$p^{l-1}(x, x') = \frac{\mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))]}{\sqrt{\mathbb{E}[\sigma^2(U^{l-1}(x))]}\sqrt{\mathbb{E}[\sigma^2(U^{l-1}(x'))]}}.$$

Note under the symmetric initialization (4.2), $K(x, x')$ degenerates to $K^L(x, x')$.

Similar to Lemma 19, the following results still hold under non-symmetric initialization.

LEMMA 35. *The following statements hold with probability at least* $1 - \delta$ *over initialization* $\mathbf{W}(0)$ *for all* $l \in [L]$.

(a) Assume $m \gtrsim pL \log(\frac{1}{\delta})$, *there holds* $\sup_{x \in \mathcal{X}} \left| \|o_0^l(x)\|_2 - 1 \right| \leq Cl\sqrt{\frac{pL \log(m/\delta)}{m}}$.
(b) Assume $m \gtrsim pL \log(\frac{m}{\delta})$, *there holds* $\sup_{x \in \mathcal{X}} \|\mathbf{V}_{L,0}^l(x)\|_{op} \leq \frac{CL}{\sqrt{m}}$.
(c) Assume $m \gtrsim pL^3 \log(\frac{m}{\delta})$, *there holds* $\sup_{x \in \mathcal{X}} \|\frac{\partial f_{\mathbf{W}(0)}(x)}{\partial \mathbf{W}^l(0)}\|_2 \leq CL$.

PROOF. The proofs of the first two parts are the same as those of Lemma 19. We only prove part (c) here. According to the first two parts, for any $l \in [L]$, there holds

$$\sup_{x\in\mathcal{X}} \Big\| \frac{\partial f_{\mathbf{W}}(0)(x)}{\partial \mathbf{W}^l(0)} \Big\|_2 = \sup_{x\in\mathcal{X}} \Big\| \mathbf{V}_{L,0}^l(x)\mathbf{a}\big(o_0^{l-1}(x)\big)^\top \Big\|_2 \leq \sqrt{m} \sup_{x\in\mathcal{X}} \|\mathbf{V}_{L,0}^l(x)\|_{op} \sup_{x\in\mathcal{X}} \|o_0^{l-1}(x)\|_2$$
$$\leq CL\Big(Cl\sqrt{\frac{pL\log(m/\delta)}{m}} + 1\Big) \leq CL,$$

where the last inequality follows from the condition $m \gtrsim pL^3 \log(\frac{m}{\delta})$. This completes the proof. □

Now, we give the concentration results of the general case. The proof is similar to that of Lemma 20. Recall the definition of K^l (see (4.50)), similarly we define $K^{m,l}(x,x') = \langle \frac{\partial f_{\mathbf{W}}(0)(x)}{\partial \mathbf{W}^l(0)}, \frac{\partial f_{\mathbf{W}}(0)(x')}{\partial \mathbf{W}^l(0)} \rangle_2$ for all $l \in [L]$ and $x, x' \in X$.

LEMMA 36. *Let $\delta \in (0,1)$. Assume $m \gtrsim pL^3 \log(\frac{m}{\delta})$. With probability at least $1 - L\exp(\mathcal{O}(pL\log(m)) - \Omega(m^{\frac{1}{3}}))) - \delta$ over initialization $(\mathbf{a}, \mathbf{W}(0))$, for all $l \in [L]$, there holds*

$$\|K^{m,l} - K^l\|_\infty \lesssim \sqrt{L} m^{-\frac{1}{6}} + \sqrt{pL\log(m) m^{-1}} + L^3 m^{-\frac{1}{3}}.$$

PROOF. For all $l \in [L]$, instead of using the estimates $\sup_{x\in\mathcal{X}} \|\mathbf{V}_{k,0}^l(x)\|_{op} \leq c_0^{k-l} m^{-\frac{1}{2}}$ in the proof of Lemma 33 in [Xu and Zhu 2024], we employ estimate $\sup_{x\in\mathcal{X}} \|\mathbf{V}_{k,0}^l(x)\|_{op} \leq CLm^{-\frac{1}{2}}$ in Lemma 35. Then, the term $|(\mathbf{V}_{L,0}^l(x)\mathbf{a})^\top \mathbf{V}_{L,0}^l(x')\mathbf{a}|$ can be controlled by CL^2.

Combining this with (4.22) yields that

$$
\begin{aligned}
&\|K^{m,l}-K^l\|_\infty \\
&=\sup_{x,x'\in\mathcal{X}}\Big|\langle o_0^{l-1}(x),o_0^{l-1}(x')\rangle_2(\mathbf{V}_{L,0}^l(x)\mathbf{a})^\top\mathbf{V}_{L,0}^l(x')\mathbf{a}-2\mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))]\prod_{h=l}^L q^h(x,x')\Big| \\
&\le \sup_{x,x'\in\mathcal{X}}\Big|\langle o_0^{l-1}(x),o_0^{l-1}(x')\rangle_2-2\mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))]\Big|\cdot\Big|(\mathbf{V}_{L,0}^l(x)\mathbf{a})^\top\mathbf{V}_{L,0}^l(x')\mathbf{a}\Big| \\
&\quad+\sup_{x,x'\in\mathcal{X}}\Big|2\mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))]\Big|\cdot\Big|(\mathbf{V}_{L,0}^l(x)\mathbf{a})^\top\mathbf{V}_{L,0}^l(x')\mathbf{a}-tr\big(\mathbf{V}_{L,0}^l(x)^\top\mathbf{V}_{L,0}^l(x')\big)\Big| \\
&\quad+\sup_{x,x'\in\mathcal{X}}\Big|2\mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))]\Big|\cdot\Big|tr\big(\mathbf{V}_{L,0}^l(x)^\top\mathbf{V}_{L,0}^l(x')\big)-\prod_{h=l}^L q^h(x,x')\Big| \\
&\lesssim L^2\sup_{x,x'\in\mathcal{X}}\Big|\langle o_0^{l-1}(x),o_0^{l-1}(x')\rangle_2-2\mathbb{E}[\sigma(U^{l-1}(x))\sigma(U^{l-1}(x'))]\Big| \\
&\quad+\sup_{x,x'\in\mathcal{X}}\Big|(\mathbf{V}_{L,0}^l(x)\mathbf{a})^\top\mathbf{V}_{L,0}^l(x')\mathbf{a}-tr\big(\mathbf{V}_{L,0}^l(x)^\top\mathbf{V}_{L,0}^l(x')\big)\Big| \\
&\quad+\sup_{x,x'\in\mathcal{X}}\Big|tr\big(\mathbf{V}_{L,0}^l(x)^\top\mathbf{V}_{L,0}^l(x')\big)-\prod_{h=l}^L q^h(x,x')\Big| \\
&=:\mathcal{E}_1^l+\mathcal{E}_2^l+\mathcal{E}_3^l,
\end{aligned}
$$

The estimates of the above three terms $\mathcal{E}_1,\mathcal{E}_2,\mathcal{E}_3$ are given as follows.

Estimate of $\mathcal{E}_1^l$: The estimate of $\mathcal{E}_1^l$ follows the same proof steps as in Lemma 6 in [Xu and Zhu 2024]. According to Lemma 6 in [Xu and Zhu 2024], one can get that $\mathcal{E}_1\lesssim LC^Lm^{-\frac{1}{3}}$. We improve this estimate from $LC^Lm^{-\frac{1}{3}}$ to $L^3m^{-\frac{1}{3}}$ by using more finer estimates of initialization terms. Specifically, instead of using their estimate $\sup_x\|o_0^l(x)\|_2\le c_0^l$ in Lemma 30 of [Xu and Zhu 2024], we apply the tight estimate $\sup_x\|o_0^l(x)\|_2\le C$ according to part (a) of Lemma 35. In addition, we set V_0 to be a $c_0^{-L}m^{-2}$-net of the S^{p-1} rather than a m^{-2}-net. Then, following the same steps of the proof of Lemma 6, with probability at least $1-L\exp(\mathcal{O}(pL\log(m))-\Omega(m^{\frac{1}{3}}))$ over initialization $\mathbf{W}(0)$, there holds

$$\mathcal{E}_1^l\lesssim L^3m^{-\frac{1}{3}}.$$

Estimates of $\mathcal{E}_2^l$: Similar to the proof of the estimate of $\mathcal{E}_1$, by using more finer estimates $\sup_x \|o_0^l(x)\|_2 \leq C$ and $\sup_{x\in\mathcal{X}} \|\mathbf{V}_{L,0}^l(x)\|_{op} \leq CLm^{-\frac{1}{2}}$, follows the same proof steps of Lemma 7 in [Xu and Zhu 2024], we can show that

$$\mathcal{E}_2^l \lesssim \frac{L^2}{m^{\frac{1}{3}}}.$$

Estimates of $\mathcal{E}_3^l$: Similar to the above arguments, we use the estimates $\sup_x \|o_0^l(x)\|_2 \leq C$ and $\sup_{x\in\mathcal{X}} \|\mathbf{V}_{L,0}^l(x)\|_{op} \leq CLm^{-\frac{1}{2}}$ to improve the proof of Lemma 8 in [Xu and Zhu 2024] and get

$$\mathcal{E}_3^l \lesssim \frac{\sqrt{L}}{m^{\frac{1}{6}}} + \sqrt{\frac{pL\log(m)}{m}} + \frac{L^2}{m^{\frac{1}{3}}}.$$

Combining the above estimates of $\mathcal{E}_1^l, \mathcal{E}_2^l, \mathcal{E}_3^l$ completes the proof of this lemma. □

CHAPTER 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we studied the learning ability of deep ReLU networks on two popular learning problems, i.e., pairwise learning tasks and gradient descent methods. For pairwise learning tasks, we provided comprehensive generalization analysis of pairwise learning problems by constructing structured deep ReLU neural networks as an approximation of the true metric. For gradient descent methods, we presented optimal rates of generalization bounds for both GD and SGD with deep ReLU networks by utilizing the power of overparameterization. Our main contributions can be summarized as follows.

- We provided generalization analysis of pairwise learning under general settings. Specifically, we established an oracle inequality of the empirical minimizer in the order of $O((\frac{\log(n)}{n})^{\frac{1}{2-\beta}})$ for a general hypothesis space with Lipschitz continuous and symmetric loss. Our general results greatly relax the conditions in previous works and can be widely applied to various learning problems. As an application, we applied our general results to conduct comprehensive generalization analysis for pairwise learning with structured deep ReLU networks. In particular, the excess generalization error bound of order $O(n^{-\frac{2r}{2r+p}})$ that matches the minimax lower bound is achieved for pairwise least squares regression.
- We presented comprehensive generalization analysis for metric and similarity learning with the hinge loss. By deriving an explicit structure of the true metric d_ρ with the hinge loss, we constructed a novel hypothesis space consisting of the structured deep ReLU networks and further established some excess generalization error bounds by

carefully estimating both the approximation error and the estimation error within the introduced structured hypothesis space. An optimal learning rate $O\big(n^{-\frac{(\theta+1)r}{p+(\theta+2)r}}\big)$ of the excess generalization error bound up to a logarithmic term is derived. We further revisited some regular properties of the problem setting and the true metric with a general loss.

- Under standard regularity assumptions on the regression function and capacity assumptions associated with the RKHS, we proved that both GD and SGD with deep ReLU networks can achieve the minimax-optimal rates $\mathcal{O}(n^{-\frac{2\beta}{2\beta+\gamma}})$ of the excess risk when the network width scales polynomially with respect to the number of layers L, the size of training dataset n and date dimension p. Our results indicate that gradient descent methods with deep ReLU networks can achieve generalization performance that is at least comparable to classical gradient methods in the kernel setting.

5.2 Future Work

Several promising directions for future research remain.

- For the pairwise learning tasks discussed in Chapters 2 and 3, our analysis primarily focused on the theoretical foundations without delving into specific algorithms for empirical risk minimization. However, in practical applications, effective learning often relies on algorithmic approaches, such as GD and SGD, to optimize the empirical risk. Understanding the generalization performance of these algorithms is crucial, as it directly impacts their real-world effectiveness. This includes analyzing their convergence rates, stability, and robustness, as well as the trade-offs between computational efficiency and generalization accuracy. Moreover, studying these algorithms in the context of pairwise learning presents unique challenges. Pairwise losses often involve dependencies between pairs of samples, leading to non-i.i.d. structures that complicate standard analyses. For instance, contrastive learning, metric learning, and ranking tasks typically involve objectives that differ significantly from the single-instance losses commonly studied in classical supervised learning.

Given these complexities, it would be of great interest to develop a comprehensive theoretical framework for analyzing the generalization behavior of GD and SGD in pairwise learning, potentially drawing connections to recent advances in deep learning theory, kernel methods, and statistical learning.

- For least squares regression with gradient descent methods discussed in Chapter 4, extending our current results to deep networks with smooth activation functions, beyond the commonly studied ReLU networks, would be a valuable direction for future research. Smooth activations, such as the hyperbolic tangent (tanh), sigmoid, or Gaussian error linear unit (GELU), often exhibit distinct optimization landscapes and generalization properties compared to ReLU, potentially leading to more refined theoretical insights.

  Additionally, it would be worthwhile to broaden the analysis to other widely-used network architectures, such as convolutional neural networks (CNNs) and residual networks (ResNets). CNNs, with their weight-sharing and local connectivity structures, are particularly effective in capturing spatial hierarchies in data, while ResNets, known for their skip connections, address the vanishing gradient problem and enable the training of extremely deep networks. Exploring these architectures in the context of least squares regression could provide more complete understanding of the interplay between network structure, optimization dynamics, and generalization performance. This line of inquiry may also reveal architecture-specific advantages, shedding light on why certain designs outperform others in practice.

Bibliography

Agarwal, Shivani and Partha Niyogi (2009). 'Generalization bounds for ranking algorithms via algorithmic stability'. In: *Journal of Machine Learning Research* 10.Feb, pp. 441–474.

Allen-Zhu, Zeyuan, Yuanzhi Li and Yingyu Liang (2019a). 'Learning and generalization in overparameterized neural networks, going beyond two layers'. In: *Advances in Neural Information Processing Systems*. Vol. 32.

Allen-Zhu, Zeyuan, Yuanzhi Li and Zhao Song (2019b). 'A convergence theory for deep learning via over-parameterization'. In: *International Conference on Machine Learning*. PMLR, pp. 242–252.

Amodei, Dario et al. (2016). 'Deep Speech 2: End-to-end Speech Recognition in English and Mandarin'. In: *International Conference on Machine Learning*. PMLR, pp. 173–182.

Arora, Sanjeev et al. (2019). 'Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks'. In: *International Conference on Machine Learning*. PMLR, pp. 322–332.

Bach, Francis (2017). 'Breaking the curse of dimensionality with convex neural networks'. In: *Journal of Machine Learning Research* 18.19, pp. 1–53.

Bahdanau, Dzmitry, Kyunghyun Cho and Yoshua Bengio (2014). 'Neural machine translation by jointly learning to align and translate'. In: *arXiv preprint arXiv:1409.0473*.

Bar-Hillel, Aharon et al. (2005). 'Learning a Mahalanobis metric from equivalence constraints.' In: *Journal of Machine Learning Research* 6.32, pp. 937–965.

Bartlett, P.L., O. Bousquet and S. Mendelson (2005). 'Local Rademacher complexities'. In: *Annals of Statistics* 33.4, pp. 1497–1537.

Bartlett, Peter, Michael Jordan and Jon McAuliffe (2006). 'Convexity, classification, and risk bounds'. In: *Journal of the American Statistical Association* 101.473, pp. 138–156.

Bartlett, Peter L, Dylan J Foster and Matus J Telgarsky (2017). 'Spectrally-normalized margin bounds for neural networks'. In: *Advances in Neural Information Processing Systems* 30.

Bartlett, Peter L and Shahar Mendelson (2006). 'Empirical minimization'. In: *Probability theory and related fields* 135.3, pp. 311–334.

Bartlett, Peter L et al. (2019). 'Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks'. In: *Journal of Machine Learning Research* 20.63, pp. 1–17.

Bietti, Alberto and Francis Bach (2021). 'Deep equals shallow for ReLU networks in kernel regimes'. In: *International Conference on Learning Representations*.

Bietti, Alberto and Julien Mairal (2019). 'On the inductive bias of neural tangent kernels'. In: *Advances in Neural Information Processing Systems*. Vol. 32.

Blumer, Anselm et al. (1989). 'Learnability and the Vapnik-Chervonenkis dimension'. In: *Journal of the ACM (JACM)* 36.4, pp. 929–965.

Boucheron, Stéphane, Olivier Bousquet and Gábor Lugosi (2005). 'Theory of classification: A survey of some recent advances'. In: *ESAIM: probability and statistics* 9, pp. 323–375.

Bousquet, Olivier (2002). 'Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms'. PhD thesis. École Polytechnique: Department of Applied Mathematics Paris, France.

Braun, Alina et al. (2024). 'Convergence rates for shallow neural networks learned by gradient descent'. In: *Bernoulli* 30.1, pp. 475–502.

Brutzkus, Alon et al. (2018). 'SGD learns over-parameterized networks that provably generalize on linearly separable data'. In: *International Conference on Learning Representations*.

Cao, Dinghao, Zheng-Chu Guo and Lei Shi (2024). 'Stochastic Gradient Descent for Two-layer Neural Networks'. In: *arXiv preprint arXiv:2407.07670*.

Cao, Qiong, Zheng-Chu Guo and Yiming Ying (2016). 'Generalization bounds for metric and similarity learning'. In: *Machine Learning* 102.1, pp. 115–132.

Cao, Yuan and Quanquan Gu (2019). 'Generalization bounds of stochastic gradient descent for wide and deep neural networks'. In: *Advances in Neural Information Processing Systems*. Vol. 32.

— (2020). ‘Generalization error bounds of gradient descent for learning over-parameterized deep relu networks’. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04, pp. 3349–3356.

Caponnetto, Andrea and Ernesto De Vito (2007). ‘Optimal rates for the regularized least-squares algorithm’. In: *Foundations of Computational Mathematics* 7.3, pp. 331–368.

Carratino, Luigi, Alessandro Rudi and Lorenzo Rosasco (2018). ‘Learning with sgd and random features’. In: *Advances in Neural Information Processing Systems*, pp. 10213–10224.

Cesa-Bianchi, Nicolo, Alex Conconi and Claudio Gentile (2004). ‘On the generalization ability of on-line learning algorithms’. In: *IEEE Transactions on Information Theory* 50.9, pp. 2050–2057.

Chechik, Gal et al. (2010). ‘Large scale online learning of image similarity through ranking.’ In: *Journal of Machine Learning Research* 11.36, pp. 1109–1135.

Chen, Zixiang et al. (2021a). ‘How much over-parameterization is sufficient to learn deep ReLU networks?’ In: *International Conference on Learning Representation*.

— (2021b). ‘How much over-parameterization is sufficient to learn deep ReLU networks?’ In: *International Conference on Learning Representations*.

Clémençon, Stéphan, Gabor Lugosi and Nicolas Vayatis (2008). ‘Ranking and empirical minimization of U-statistics’. In: *The Annals of Statistics* 36.2, pp. 844–874.

Cucker, Felipe and Steve Smale (2002). ‘On the mathematical foundations of learning’. In: *Bulletin of the American mathematical society* 39.1, pp. 1–49.

Cucker, Felipe and Ding-Xuan Zhou (2007). *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press.

Davis, Jason V et al. (2007). ‘Information-theoretic metric learning’. In: *Proceedings of the 24th international conference on Machine learning*, pp. 209–216.

De la Peña, V. and E. Giné (1999). *Decoupling: From Dependence to Independence*. New York: Springer-Verlag.

Devlin, Jacob et al. (2019). ‘Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding’. In: *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers), pp. 4171–4186.

Dieuleveut, Aymeric and Francis Bach (2016). 'Nonparametric stochastic approximation with large step-sizes'. In: *Annals of Statistics* 44.4, pp. 1363–1399.

Du, Simon et al. (2019). 'Gradient descent finds global minima of deep neural networks'. In: *International Conference on Machine Learning*. PMLR, pp. 1675–1685.

Du, Simon S et al. (2018). 'Gradient descent provably optimizes over-parameterized neural networks'. In: *International Conference on Learning Representations*.

Frei, Spencer, Niladri S Chatterji and Peter L Bartlett (2023). 'Random feature amplification: Feature learning and generalization in neural networks'. In: *Journal of Machine Learning Research* 24.303, pp. 1–49.

Glorot, Xavier, Antoine Bordes and Yoshua Bengio (2011). 'Deep sparse rectifier neural networks'. In: *International Conference on Artificial Intelligence and Statistics*, pp. 315–323.

Golowich, Noah, Alexander Rakhlin and Ohad Shamir (2018). 'Size-independent sample complexity of neural networks'. In: *Conference On Learning Theory*. PMLR, pp. 297–299.

Goodfellow, Ian, Yoshua Bengio and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Guo, Yinglong, Shaohan Li and Gilad Lerman (2024). 'The effect of Leaky ReLUs on the training and generalization of overparameterized networks'. In: *arXiv preprint arXiv:2402.11942*.

Guo, Zheng-Chu, Ting Hu and Lei Shi (2022). 'Distributed spectral pairwise ranking algorithms'. In: *Inverse Problems* 39.2, p. 025003.

Guo, Zheng-Chu and Lei Shi (2019). 'Fast and strong convergence of online learning algorithms'. In: *Advances in Computational Mathematics* 45, pp. 2745–2770.

Guo, Zheng-Chu and Yiming Ying (2014). 'Guaranteed classification via regularized similarity learning'. In: *Neural Computation* 26.3, pp. 497–522.

Györfi, László et al. (2006). *A Distribution-free Theory of Nonparametric Regression*. Springer Science & Business Media.

He, Kaiming et al. (2016). 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Hinton, Geoffrey et al. (2012). 'Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups'. In: *IEEE Signal Processing Magazine* 29.6, pp. 82–97.

Hoeffding, Wassily (1963). 'Probability inequalities for sums of bounded random variables'. In: *Journal of the American Statistical Association* 58.301, pp. 13–30.

Hu, Tianyang et al. (2021). 'Regularization matters: A nonparametric perspective on overparametrized neural network'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 829–837.

Huai, Mengdi et al. (2019). 'Deep metric learning: The Generalization Analysis and an Adaptive Algorithm.' In: *IJCAI*, pp. 2535–2541.

Huang, Shuo et al. (2023). 'Generalization analysis of pairwise learning for ranking with deep neural networks'. In: *Neural Computation* 35.6, pp. 1135–1158.

Jacot, Arthur, Franck Gabriel and Clément Hongler (2018). 'Neural tangent kernel: Convergence and generalization in neural networks'. In: *Advances in Neural Information Processing Systems* 31.

Ji, Ziwei and Matus Telgarsky (2020). 'Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks'. In: *International Conference on Learning Representations*.

Jin, Rong, Shijun Wang and Yang Zhou (2009). 'Regularized distance metric learning: Theory and algorithm'. In: *Advances in Neural Information Processing Systems*, pp. 862–870.

Kar, Purushottam and Prateek Jain (2011). 'Similarity-based learning via data driven embeddings'. In: *Advances in Neural Information Processing Systems* 24.

Kaya, Mahmut and Hasan Şakir Bilge (2019). 'Deep metric learning: A survey'. In: *Symmetry* 11.9, p. 1066.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton (2012). 'Imagenet classification with deep convolutional neural networks'. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E Hinton (2017). 'Imagenet classification with deep convolutional neural networks'. In: *Communications of the ACM* 60.6, pp. 84–90.

Kuzborskij, Ilja and Csaba Szepesvári (2022). 'Learning lipschitz functions by gd-trained shallow overparameterized relu neural networks'. In: *arXiv preprint arXiv:2212.13848*.

Lai, Jianfa et al. (2023). 'Generalization ability of wide neural networks on $\mathbb{R}$'. In: *arXiv preprint arXiv:2302.05933*.

LeCun, Yann, Yoshua Bengio and Geoffrey Hinton (2015). 'Deep learning'. In: *Nature* 521.7553, pp. 436–444.

Ledoux, Michel and Michel Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Vol. 23. Berlin: Springer.

Lei, Guanhang and Lei Shi (2024). 'Pairwise ranking with Gaussian kernel'. In: *Advances in Computational Mathematics* 50.4, p. 70.

Lei, Yunwen, Rong Jin and Yiming Ying (2022). 'Stability and generalization analysis of gradient methods for shallow neural networks'. In: *Advances in Neural Information Processing Systems*. Vol. 35. PMLR.

Lei, Yunwen and Yiming Ying (2016). 'Generalization analysis of multi-modal metric learning'. In: *Analysis and Applications* 14.04, pp. 503–521.

— (2020). 'Sharper generalization bounds for learning with gradient-dominated objective functions'. In: *International Conference on Learning Representations*.

Li, Yuanzhi and Yingyu Liang (2018). 'Learning overparameterized neural networks via stochastic gradient descent on structured data'. In: *Advances in Neural Information Processing Systems*. Vol. 31.

Lin, Junhong and Lorenzo Rosasco (2016). 'Optimal learning for multi-pass stochastic gradient methods'. In: *Advances in Neural Information Processing Systems*, pp. 4556–4564.

— (2017). 'Optimal rates for multi-pass stochastic gradient methods'. In: *Journal of Machine Learning Research* 18.1, pp. 3375–3421.

Maurer, Andreas (2008). 'Learning similarity with operator-valued large-margin classifiers'. In: *Journal of Machine Learning Research* 9, pp. 1049–1082.

Mendelson, Shahar (2003). 'A few notes on statistical learning theory'. In: *Advanced lectures on machine learning*. Springer, pp. 1–40.

Mücke, Nicole, Gergely Neu and Lorenzo Rosasco (2019). 'Beating SGD Saturation with Tail-Averaging and Minibatching'. In: *Advances in Neural Information Processing Systems*, pp. 12568–12577.

Nacson, Mor Shpigel et al. (2019). 'Convergence of gradient descent on separable data'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3420–3428.

Neyshabur, Behnam, Ryota Tomioka and Nathan Srebro (2015). 'Norm-based capacity control in neural networks'. In: *Conference on Learning Theory*. PMLR, pp. 1376–1401.

Nguyen, Mike and Nicole Mücke (2023). 'Random feature approximation for general spectral methods'. In: *arXiv preprint arXiv:2308.15434*.

— (2024). 'How many neurons do we need? A refined analysis for shallow networks trained with gradient descent'. In: *Journal of Statistical Planning and Inference*, p. 106169.

Nitanda, Atsushi, Geoffrey Chinot and Taiji Suzuki (2019). 'Gradient descent can learn less over-parameterized two-layer neural networks on classification problems'. In: *arXiv preprint arXiv:1905.09870*.

Nitanda, Atsushi and Suzuki Taiji (2021). 'Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime'. In: *International Conference on Learning Representations*.

Parhi, Rahul and Robert D Nowak (2022). 'Near-minimax optimal estimation with shallow ReLU neural networks'. In: *IEEE Transactions on Information Theory* 69.2, pp. 1125–1140.

Pillaud-Vivien, Loucas, Alessandro Rudi and Francis Bach (2018). 'Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes'. In: *Advances in Neural Information Processing Systems*, pp. 8114–8124.

Rahimi, Ali and Benjamin Recht (2008). 'Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning'. In: *Advances in Neural Information Processing Systems* 21.

Rejchel, Wojciech (2012). 'On ranking and generalization bounds'. In: *Journal of Machine Learning Research* 13.May, pp. 1373–1392.

Richards, Dominic and Ilja Kuzborskij (2021). 'Stability & Generalisation of Gradient Descent for Shallow Neural Networks without the Neural Tangent Kernel'. In: *Advances in Neural Information Processing Systems*. Vol. 34. PMLR.

Richards, Dominic and Mike Rabbat (2021). 'Learning with gradient descent and weakly convex losses'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1990–1998.

Rosasco, Lorenzo and Silvia Villa (2015). 'Learning with incremental iterative regularization'. In: *Advances in Neural Information Processing Systems*, pp. 1630–1638.

Roth, Karsten et al. (2020). 'Revisiting training strategies and generalization performance in deep metric learning'. In: *International Conference on Machine Learning*. PMLR, pp. 8242–8252.

Scetbon, Meyer and Zaid Harchaoui (2021). 'A spectral analysis of dot-product kernels'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3394–3402.

Schmidt-Hieber, Johannes (2020). 'Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function'. In: *The Annals of Statistics* 48.4, pp. 1875–1897.

Shalit, Uri, Daphna Weinshall and Gal Chechik (2010). 'Online learning in the manifold of low-rank matrices'. In: *Advances in Neural Information Processing Systems* 23.

Silver, David et al. (2016). 'Mastering the game of Go with deep neural networks and tree search'. In: *Nature* 529.7587, pp. 484–489.

Simonyan, Karen and Andrew Zisserman (2014). 'Very deep convolutional networks for large-scale image recognition'. In: *arXiv preprint arXiv:1409.1556*.

Smale, Steve and Ding-Xuan Zhou (2007). 'Learning theory estimates via integral operators and their approximations'. In: *Constructive Approximation* 26.2, pp. 153–172.

Steinwart, Ingo and Andreas Christmann (2008). *Support Vector Machines*. Springer Science & Business Media.

Su, Lili and Pengkun Yang (2019). 'On learning over-parameterized neural networks: A functional approximation perspective'. In: *Advances in Neural Information Processing Systems*. Vol. 32.

Suh, Namjoon, Hyunouk Ko and Xiaoming Huo (2021). 'A non-parametric regression viewpoint: Generalization of overparametrized deep ReLU network under noisy observations'. In: *International Conference on Learning Representations*.

Sutskever, Ilya, Oriol Vinyals and Quoc V Le (2014). 'Sequence to sequence learning with neural networks'. In: *Advances in Neural Information Processing Systems* 27.

Taheri, Hossein and Christos Thrampoulidis (2024). 'Generalization and Stability of Interpolating Neural Networks with Minimal Width'. In: *Journal of Machine Learning Research* 25.156, pp. 1–41.

Taheri, Hossein, Christos Thrampoulidis and Arya Mazumdar (2024). 'Sharper Guarantees for Learning Neural Network Classifiers with Gradient Methods'. In: *arXiv preprint arXiv:2410.10024*.

Vaart, Aad W. van der and Jon A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Vaswani, Ashish et al. (2017). 'Attention is All you Need'. In: *Advances in Neural Information Processing Systems* 30.

Vershynin, Roman (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Wainwright, Martin J (2019). *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Vol. 48. Cambridge University Press.

Wang, Puyu et al. (2025a). 'Generalization guarantees of gradient descent for shallow neural networks'. In: *Neural Computation* 37.2, pp. 344–402.

Wang, Puyu et al. (2025b). 'Optimal rates for gradient descent methods with two-Layer ReLU networks'. In: *Preprint*.

Xu, Jiaming and Hanjing Zhu (2024). 'Overparametrized multi-layer neural networks: uniform concentration of neural tangent kernel and convergence of stochastic gradient descent'. In: *Journal of Machine Learning Research* 25.94, pp. 1–83.

Yao, Yuan, Lorenzo Rosasco and Andrea Caponnetto (2007). 'On early stopping in gradient descent learning'. In: *Constructive Approximation* 26.2, pp. 289–315.

Yarotsky, Dmitry (2017). 'Error bounds for approximations with deep ReLU networks'. In: *Neural networks* 94, pp. 103–114.

Ye, Han-Jia, De-Chuan Zhan and Yuan Jiang (2019). 'Fast generalization rates for distance metric learning'. In: *Machine Learning* 108.2, pp. 267–295.

Ying, Yiming and Colin Campbell (2010). 'Rademacher chaos complexities for learning the kernel problem'. In: *Neural Computation* 22.11, pp. 2858–2886.

Ying, Yiming and Peng Li (2012). 'Distance metric learning with eigenvalue optimization'. In: *The Journal of Machine Learning Research* 13.1, pp. 1–26.

Ying, Yiming and Massimiliano Pontil (2008). 'Online gradient descent learning algorithms'. In: *Foundations of Computational Mathematics* 8.5, pp. 561–596.

Ying, Yiming and Ding-Xuan Zhou (2006). 'Online regularized classification algorithms'. In: *IEEE Transactions on Information Theory* 52.11, pp. 4775–4788.

— (2016). 'Online pairwise learning algorithms'. In: *Neural Computation* 28.4, pp. 743–777.

Zhang, Chiyuan et al. (2021a). 'Understanding deep learning (still) requires rethinking generalization'. In: *Communications of the ACM* 64.3, pp. 107–115.

— (2021b). 'Understanding deep learning (still) requires rethinking generalization'. In: *Communications of the ACM* 64.3, pp. 107–115.

Zhang, Tong (2004). 'Statistical behavior and consistency of classification methods based on convex risk minimization'. In: *The Annals of Statistics* 32.1, pp. 56–85.

— (2005). 'Learning bounds for kernel regression using effective data dimensionality'. In: *Neural Computation* 17.9, pp. 2077–2098.

Zhou, Ding-Xuan (2020). 'Universality of deep convolutional neural networks'. In: *Applied and Computational Harmonic Analysis* 48.2, pp. 787–794.

Zhou, Junyu, Puyu Wang and Ding-Xuan Zhou (2024a). 'Generalization analysis with deep ReLU networks for metric and similarity learning'. In: *arXiv preprint arXiv:2405.06415*.

Zhou, Junyu et al. (2024b). 'Fine-grained analysis of non-parametric estimation for pairwise learning'. In: *arXiv preprint arXiv:2305.19640*.

Zhou, Junyu et al. (2025). 'Optimal rates for generalization of gradient descent methods with deep neural networks'. In: *Preprint*.

Zhou, Tian-Yi and Xiaoming Huo (2024). 'Classification of data generated by Gaussian mixture models using deep ReLU networks'. In: *Journal of Machine Learning Research* 25.190, pp. 1–54.

Zhou, Tian-Yi et al. (2024c). ‘Optimal classification-based anomaly detection with neural networks: theory and practice’. In: *arXiv preprint arXiv:2409.08521*.

Zou, Difan and Quanquan Gu (2019). ‘An improved analysis of training over-parameterized deep neural networks’. In: *Advances in neural information processing systems*. Vol. 32.

Zou, Difan et al. (2018). ‘Stochastic gradient descent optimizes over-parameterized deep relu networks. arxiv e-prints, art’. In: *arXiv preprint arXiv:1811.08888*.

— (2020). ‘Gradient descent optimizes over-parameterized deep ReLU networks’. In: *Machine Learning* 109, pp. 467–492.