

GRADIENT DESCENT METHODS IN MODERN  
MACHINE LEARNING PROBLEMS: PROVABLE  
GUARANTEES

by

Hanjing Zhu

Business Administration  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Jiaming Xu, Supervisor

\_\_\_\_\_  
Alex Belloni

\_\_\_\_\_  
Rong Ge

\_\_\_\_\_  
Yehua Wei

Dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy  
in Business Administration  
in the Graduate School of  
Duke University

2023

ABSTRACT

GRADIENT DESCENT METHODS IN MODERN  
MACHINE LEARNING PROBLEMS: PROVABLE  
GUARANTEES

by

Hanjing Zhu

Business Administration  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Jiaming Xu, Supervisor

\_\_\_\_\_  
Alex Belloni

\_\_\_\_\_  
Rong Ge

\_\_\_\_\_  
Yehua Wei

An abstract of a dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy  
in Business Administration  
in the Graduate School of  
Duke University

2023

Copyright © 2023 by Hanjing Zhu  
All rights reserved

# Abstract

Modern machine learning methods have demonstrated remarkable success in many industries. For instance, the famous ChatGPT relies on a machine learning model trained with a substantial volume of text and conversation data. To achieve optimal model performance, an efficient optimization algorithm is essential for learning the model parameters. Among optimization methods, gradient descent (GD) methods are the simplest ones. Traditionally, GD methods have shown excellent performance in conventional machine learning problems with nice objective functions and simple training paradigms where computation occurs on a single server storing all the data. However, the understanding of how GD methods perform in modern machine learning problems with non-convex and non-smooth objective functions or more complex training paradigm remains limited.

This thesis is dedicated to providing a theoretical understanding of why gradient descent methods excel in modern machine learning problems. In the first half of the thesis, we study stochastic gradient descent (SGD) in training multi-layer fully connected feedforward neural networks with Rectified Linear Unit (ReLU) activation. Since the loss function in training deep neural networks is non-convex and non-smooth, the standard convergence guarantees of GD for convex loss functions cannot be applied. Instead, through a kernel perspective, we demonstrate that when fresh data arrives in a stream, SGD ensures the exponential convergence of the average prediction error. In the second half, we investigate the utilization of GD methods in a new training paradigm, featuring a central parameter server (PS) and numerous clients storing data locally. Privacy constraints prevent the local data from being revealed to the PS, making this distributed learning setting particularly relevant in the current big-data era where data is often sensitive and too large to be stored on a single device. In practical applications, this distributed setting presents two major

challenges: data heterogeneity and adversarial attacks. To overcome these challenges and achieve accurate estimates of model parameters, we propose a GD-based algorithm and provide convergence guarantees for both strongly convex and non-convex loss functions.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>x</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Overview of the thesis . . . . .	3
<b>2 SGD in Learning Multi-Layer NN</b>	<b>5</b>
2.1 Introduction . . . . .	5
2.2 Previous works . . . . .	9
2.3 Problem Setup . . . . .	12
2.4 Main Result . . . . .	14
2.4.1 Concentration of NTK at Initialization . . . . .	14
2.4.2 Average Prediction Error under SGD . . . . .	19
2.5 Proof of Theorem 1 . . . . .	22
2.6 Bounding $\ H_t - H_0\ _\infty$ . . . . .	27
2.7 Proof of Theorem 2 . . . . .	31
2.8 Numerical Study . . . . .	38
2.8.1 Synthetic data . . . . .	38
2.8.2 Real data experiment . . . . .	41
<b>3 Proofs of Chapter 2 Intermediate Results</b>	<b>43</b>
3.1 Proofs in Section 2.5 . . . . .	43

3.1.1	Proof of Lemma 2.5.1 . . . . .	43
3.1.2	Proof of Lemma 2.5.2 . . . . .	49
3.1.3	Proof of Lemma 2.5.3 . . . . .	55
3.2	Proofs in Section 2.6 . . . . .	66
3.2.1	Proof of Proposition 2.6.1 . . . . .	66
3.2.2	Proof of Lemma 2.6.2 . . . . .	70
3.2.3	Proof of Lemma 2.6.3 . . . . .	76
3.2.4	Proof of Lemma 2.6.4 . . . . .	81
3.3	Proof of lemmas in Section 2.7 . . . . .	85
3.3.1	Proof of Lemma 2.7.2 . . . . .	85
3.3.2	Proof of Lemma 2.7.3 . . . . .	88
3.3.3	Proof of Lemma 2.7.4 . . . . .	102
3.4	Proof of Corollary 1 . . . . .	105
3.4.1	Proof of Theorem 3 . . . . .	107
<b>4</b>	<b>Robust GD in Federated Learning</b>	<b>109</b>
4.1	Introduction . . . . .	109
4.2	Problem Setup . . . . .	113
4.3	Algorithm and Theoretical Guarantees . . . . .	114
4.3.1	Theoretical Guarantees of Phase 1 . . . . .	116
4.3.2	Theoretical Guarantees of Phase 2 . . . . .	119
4.4	Byzantine Federated Learning Algorithm . . . . .	123
4.4.1	Phase 1: Coarse Estimation of Cluster Centers . . . . .	123
4.4.2	Phase 2: Clustering and Refinement of the Estimation . . . . .	127
4.5	Numerical Studies . . . . .	133

4.5.1	Comparison with Methods Ignoring Cluster Structure . . . . .	133
4.5.2	Comparison with Methods Utilizing Cluster Structure . . . . .	135
4.6	Further Discussions on SDP Relaxations . . . . .	139
4.7	Crucial Role of Initial Label in [GHYR19] . . . . .	140
<b>5</b>	<b>Proofs in Chapter 4</b>	<b>143</b>
5.1	Proof of Theorem 4 . . . . .	143
5.1.1	Proof of Lemma 5.1.1 . . . . .	144
5.1.2	Proof of Lemma 5.1.2 . . . . .	145
5.2	Proof of Theorem 5 . . . . .	156
5.3	Proof of Theorem 6 . . . . .	159
5.3.1	Proof of Theorem 8 and 9 . . . . .	159
5.3.2	Proof of Lemma 4.4.2 . . . . .	163
5.4	Proof of Lemma 4.4.1 . . . . .	165
5.5	Proof of Intermediate Results in Section 5.1 . . . . .	166
5.5.1	Proof of Proposition 5.1.5 . . . . .	166
5.5.2	Proof of Proposition 5.1.6 . . . . .	171
<b>6</b>	<b>Conclusion and Future Directions</b>	<b>177</b>
<b>A</b>	<b>Auxiliary Result</b>	<b>179</b>
A.1	Concentration Inequalities . . . . .	179
A.2	VC Dimension . . . . .	180
A.3	Kernel . . . . .	182
A.4	Others . . . . .	184
	<b>Bibliography</b>	<b>187</b>
	<b>Biography</b>	<b>194</b>

# List of Tables

2.1	Summary of related works . . . . .	9
4.1	Comparison of misclassification errors under Gaussian attack . . . . .	136

# List of Figures

1.1	Visual illustration of FL . . . . .	3
2.1	Comparison of GD/SGD dynamic versus different characterizations . . . . .	7
2.2	Illustration of Multi-Layer Neural Network $f(x; \mathbf{W})$ . . . . .	13
2.3	Normalized prediction error for different $f^*$ . . . . .	39
2.4	Average prediction error and the characterization based on $\Phi$ . . . . .	40
2.5	Evolution of weight and sign flips for a 4-layer teacher neural network $f^*$ . . . . .	41
2.6	Normalized training loss for the first 2000 iterations . . . . .	42
3.1	Diagram of the proof structure in Section 3.1 . . . . .	43
3.2	Illustration of the key idea in showing $ \mathcal{D}_k  \leq m^d  \mathcal{D}_{k-1} $ . . . . .	50
4.1	Solution to (4.3) with the adjacency matrix $\mathbf{A}$ replaced by the negative pairwise distance matrix $-\mathbf{D}$ . . . . .	125
4.2	Weighted population loss with varying cluster separation. . . . .	135
4.3	Comparison of estimation errors $\ \widehat{\theta}_k^{(t)} - \theta_k\ _2$ under different Byzantine attacks. . . . .	137
4.4	Comparison of estimation error $\sum_{k=1}^K \ \widehat{\theta}_k - \theta_k\ _2$ for various $\lambda$ . . . . .	138
4.5	Solution to Peng and Wei’s relaxation with the presence of outliers . . . . .	140
4.6	An example illustrating the limitation of [GHYR19] in handling unbalanced data . . . . .	142

# Acknowledgements

This thesis will not be possible without the help from many people.

First of all, I thank my advisor, Prof. Jiaming Xu, for his advice and help. His knowledge and the way of thinking guided me throughout the five years at Duke. Whenever I encountered some difficulty, he was always willing to spend time to discuss with me and was able to give me constructive ideas to resolve the challenges. He was also patient in helping me with the writing and the presentations, which was of extreme importance to me.

I thank the entire Decision Science group at the Fuqua School of Business for a wonderful atmosphere. The rotation at the beginning of the PhD journey gave me the opportunity to explore various different fields. Group seminars allow me to learn more about cutting-edge research topics in operation research, machine learning, causal inference and etc. Peers within the Decision Science group are nice. It is always nice to chat about each other's research in the office area. All of these are extremely valuable experience to me.

In addition to Fuqua, I thank the department of Mathematics, Statistics and Computer Science at Duke University. The courses there prepared me with the tools to conduct research. I also thank the Pratt School of Engineering for offering me an opportunity of lecturing a decision model class. Having such a teaching experience is really valuable.

Outside Duke, I have received much support as well. My research is supported in part by the NSF Grant CCF-1856424 and an NSF CAREER award CCF-2144593. I thank Prof. Rong Chen at Rutgers University, Prof. Ruey Tsay and Prof. Weibiao Wu at University of Chicago for their support and encouragement at the very beginning when I decided to pursue a graduate study in the field of statistics.

In addition to the academic support, I am fortunate to have received so many support in my personal life. I want to thank my friends, at and outside Duke for their sincere support. I thank my friends Yue Li, Tianyu Wang and Kan Xu for their valuable feedbacks whenever I shared my concern or asked for help. I also thank my PhD peers at Duke, especially my cohort mates Jinwei, Minjun and Sophie for hangouts during the long pandemic shutdown.

I thank my office mate Benda Yin for his kind help throughout my last year of PhD study.

Finally, I want to send my deepest gratitude to my parents for giving me life, love and support throughout such a long journey. To give up a career in economics and start a new career in statistics and machine learning is never easy to me. Without the love and the support from my family, I wouldn't be brave enough to begin this journey and be able to devote myself in the study for all these years.

# Chapter 1

## Introduction

Machine learning is a field dedicated to the development of systems capable of learning and enhancing their prediction accuracy through data analysis. For example, hospitals employ CT images to develop model that detect tumors [HTW<sup>+</sup>18]. With more CT images fed to the model, the tumor detection accuracy improves. Additionally, machine learning finds extensive applications in various domains such as machine translation, autonomous driving, recommendation system, gaming AI and etc, where it plays a crucial role in making precise predictions.

A machine learning method consists of two components. First, it has a model taking data as input and generates predictions, with the accuracy of the prediction relying on the model parameters. Second, the method incorporates a loss function evaluating the model performance. A superior model, represented by parameters of higher quality, corresponds to a lower loss function value. The primary objective of the machine learning method is to attain parameters that yield a sufficiently low loss value.

While classical methods, such as ordinary least square, often offer explicit solutions for optimal model parameters, machine learning methods tend to be more intricate, lacking closed-form solutions. Thus, the learning of the optimal model parameters necessitates the use of optimization algorithms. Among optimization algorithms, gradient descent (GD) methods stands out as one of the simplest and the most widely used. Specifically, the vanilla gradient descent updates the model parameter  $\theta$  by the gradient as

$$\theta_{t+1} \leftarrow \theta_t - \eta \nabla L(\theta_t),$$

where  $\eta$  is the step size,  $\nabla L(\theta)$  is the gradient of the loss function  $L$  evaluated at  $\theta$ .

Though GD methods have proven effective in solving traditional machine learning prob-

lems with well-behaved loss functions and single-server computation, they face significant challenges in modern machine learning scenarios.

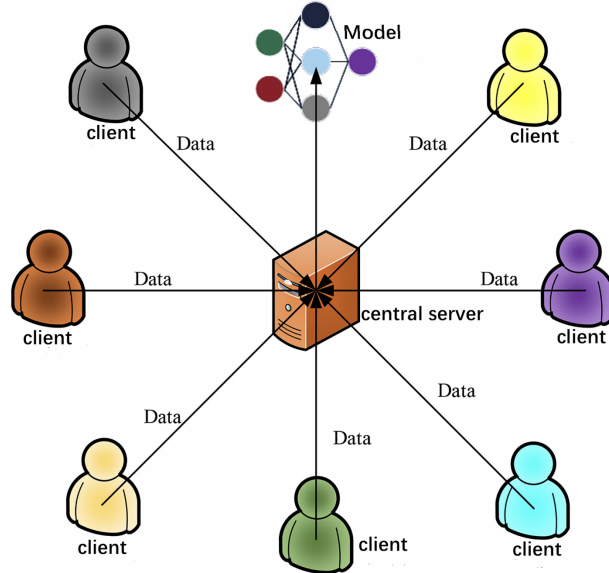
The first challenge arises from the increased complexity of loss functions, particularly from more intricate model architectures. Modern AI developments, like ChatGPT, rely on GD-optimized deep neural networks (NN) known as transformers, which take texts as input and generate predicted next words [WBZ<sup>+</sup>21]. Due to the complexities of these neural network architectures, the associated loss functions are often non-convex and non-smooth. This poses considerable difficulties in analyzing GD methods during the training of NNs, as existing works such as [RM51, Pol64, GFJ15, YLL16] have only proven the convergence guarantee of GD methods for convex and smooth loss functions.

In addition to coping with more complicated loss functions, GD in modern machine learning must navigate a new training paradigm. In the current big-data era, data is often too big to be stored in a single location. This large volume of data also raises privacy concerns, as the data source can potentially be deduced based on observations. To address these challenges, the distributed learning system illustrated in Figure 1.1 [MLL<sup>+</sup>22] is commonly utilized. This system consists of a central parameter server (PS) and numerous clients that store data. The goal of PS is to learn machine learning models through communications with clients while ensuring that during any communication round, clients cannot reveal the local data to PS due to privacy concern [MMR<sup>+</sup>17, KMA<sup>+</sup>21]. Such study receives large attention in practice, and is known as Federated Learning (FL).

Given the aforementioned two challenges, the thesis aims to address two primary questions:

- Why do GD methods perform well in deep learning despite dealing with non-convex and non-smooth loss functions?
- How GD algorithms should be employed in the novel training paradigm as FL?

By answering these questions, the thesis seeks to contribute to the understanding and optimization of gradient descent methods in modern machine learning.



**Figure 1.1:** Visual illustration of FL: Clients store data; Central parameter server (PS) communicates with each client to learn the desired machine learning model

## 1.1 Overview of the thesis

The thesis consists of two halves.

**SGD provably learns multi-layer neural networks** In the first half, we study SGD in training multi-layer neural networks. One way to analyze the dynamic of GD methods is through the lens of neural tangent kernel (NTK), which has led to significant progress in understanding why GD methods work well for training neural networks [DZPS19, DLL<sup>+</sup>19, ADH<sup>+</sup>19, SY19, CG19]. Yet there remain two significant gaps between theory and practice. First, the existing convergence theory only takes into account the contribution of the NTK from the last hidden layer, whereas in practice the intermediate layers also play an instrumental role. Second, most existing works assume that the training data are provided a priori in a batch, while less attention has been paid to the important setting where the training data arrive in a stream. In Chapter 2, we close these two gaps. We first show that with random initialization, the NTK function converges to some deterministic function uniformly for all layers as the number of neurons tends to infinity. Then we apply

the uniform convergence result to further prove that the prediction error of multi-layer neural networks under SGD converges in expectation in the streaming data setting. A key ingredient in our proof is to show the number of activation patterns of an  $L$ -layer neural network with width  $m$  is only polynomial in  $m$  although there are  $mL$  neurons in total. To ease the reading, we defer the proofs of technical lemmas to Chapter 3.

**Robust Gradient Descent Algorithm in FL with Adversarial Attacks** In the second half, we study GD methods in FL. In practice, FL encounters two challenges. First, with a system of large scale, machine failure and malicious attack can occur easily. Second, data among different clients are heterogeneous and unbalanced due to differences in client preference and activity. In Chapter 4, we address both challenges. Specifically, we assume that data points on clients follow one of  $K$  distributions based on a hidden cluster structure, and the local data volume varies across clients. Additionally, we consider the presence of a Byzantine adversary with complete knowledge of the system, capable of controlling up to  $\epsilon$  fraction of clients to behave adversarially. To overcome the challenges, we design a Byzantine-resilient algorithm that accurately recovers the underlying clusters and estimate the model parameters for each cluster. Our algorithm begins with graph clustering via semidefinite programming (SDP) on a few anchor clients (clients with a sufficient number of data points) to obtain a coarse estimation of model parameters. Then each client determines its cluster label utilizing this coarse estimation. Finally, within each output cluster, we apply batching and geometric median to robustly aggregate local gradients and iteratively refine the model estimation. We provide convergence analysis for both strongly convex and non-convex objectives. Notably, our results hold without any distributional assumptions and significantly improve upon prior work in cases of highly unbalanced local data volume. Our proof crucially relies on a novel deterministic analysis of SDP using a dual certificate argument, which could be of independent interest. To validate our theoretical findings, we present numerical experiments. The detailed proofs are provided in Chapter 5.

# Chapter 2

## SGD in Learning Multi-Layer NN

### 2.1 Introduction

In this chapter, we study gradient descent dynamics in training neural networks for minimizing non-convex and non-smooth loss functions. In recent years, researchers have proposed different ways to study the theoretical properties of GD dynamics in the training of neural networks. For example, the mean-field theory is used in [CCGZ20, MMN18, MMM19, SS22] to analyze the SGD of infinite-width feed-forward neural networks. Optimal transport theory is employed in [CB18] to study the gradient flow of neural networks and to show that the training error converges to the global optimum under some mild conditions. In addition, [HLLL19] connects the SGD of neural networks in training to the diffusion process.

A different line of research focuses on understanding the gradient descent of neural networks through kernels, in particular the neural tangent kernel (NTK). Specifically, given an  $L$ -layer neural network  $f(x; \mathbf{W})$  with input  $x$  and parameter  $\mathbf{W}$ , we define NTK for a sequence of weights  $\{\mathbf{W}(t)\}$  as

$$H_t(x, x') \triangleq \left\langle \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}}, \frac{\partial f(x'; \mathbf{W}(t))}{\partial \mathbf{W}} \right\rangle = \sum_{\ell=1}^L H_t^{(\ell)}(x, x'), \quad (2.1)$$

where

$$H_t^{(\ell)}(x, x') \triangleq \left\langle \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}, \frac{\partial f(x'; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} \right\rangle \quad (2.2)$$

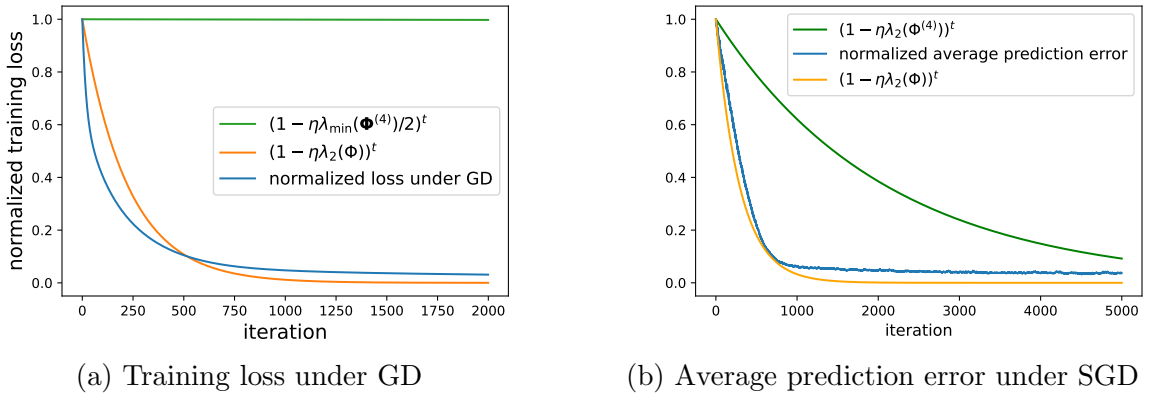
is the NTK from the  $\ell$ -th hidden layer. It is first introduced by [JGH18], which shows that gradient descent on infinite width neural networks can be viewed as learning through the NTK. Subsequent works [AZLS19a, DZPS19, SY19, ADH<sup>+</sup>19, DLL<sup>+</sup>19, ZCZG20] connect GD and SGD with the NTK, and show that with overparameterization and random initial-

ization, the training error converges to 0. Similar convergence results are also established in other types of neural networks beyond feed-forward neural networks [AZL20, AZL19b, AZL19a, AZLS19b, DWZ<sup>+</sup>18, LWY<sup>+</sup>19], such as convolutional neural networks (CNN) and residual neural networks (ResNet).

Despite these remarkable progresses, there remain two significant gaps. Firstly, the existing theory does not accurately characterize the convergence rate of GD. Specifically, given a batch  $\{(x_i, y_i)\}_{i=1}^n$ , [DLL<sup>+</sup>19] first shows that the NTK matrix from the last hidden layer  $\mathbf{H}_t^{(L)} = \left(\frac{1}{n}H_t^{(L)}(x_i, x_j)\right)$  is close to some deterministic kernel matrix  $\Phi^{(L)}$ . Based on this, the authors further show the training loss converges at a linear rate  $\left(1 - \frac{\eta}{2}\lambda_{\min}(\Phi^{(L)})\right)^t$  where  $\eta$  is the step size. However, such a characterization based on the last hidden layer is very loose for two reasons. First of all, the characterization only captures the contribution from the last hidden layer. Secondly, as pointed out by [SY19],  $\lambda_{\min}(\Phi^{(L)})$  goes to 0 as the batch size  $n$  goes to infinity. In contrast, as illustrated in Figure 2.1a, the actual GD dynamic converges much faster and can be more accurately characterized via the spectrum of the integral operator  $\Phi$  associated with some deterministic kernel function  $\Phi = \sum_{\ell=1}^L \Phi^{(\ell)}$ , which captures the contribution from all layers.

Secondly, most existing works study GD in the batch setting where the training data is provided a priori in a batch. It remains unclear how SGD performs in the streaming setting where the data arrives in a stream. The streaming data arises in a variety of fields such as finance, news organization, and information technology [OMM<sup>+</sup>02, AZL19b, ILG07]. Such streaming data is usually inspected once and archived afterwards immediately without being examined again. Apart from vast sources of naturally generated streaming data, there are ubiquitous situations where the streaming data is preferred even though batches of samples can be obtained. For instance, [OMM<sup>+</sup>02] points out that in medical or marketing data mining, the volume of data is so large that only one pass over data is allowed due to computational constraints. Moreover, [FIM<sup>+</sup>01, Mut05] argues that the streaming data is useful in privacy-preserving data mining, where the data is kept confidential by users and analyzed via a single pass.

It is challenging to close these two gaps. The analysis of the NTK from intermediate



**Figure 2.1:** Comparison of GD/SGD dynamic versus different characterizations. The actual estimation errors are shown in blue. The characterization based on the last layer is shown in green while the characterization based on all layers is shown in orange. The data is generated according to  $y = f^*(x) + u$ , where  $f^*$  is a linear function,  $u \sim \mathcal{N}(0, 0.01)$ , and  $x$  is generated uniformly over the unit sphere. We learn a 4-layer neural network with  $m = 1000$  neurons in each hidden layer using step size  $\eta = 0.2$ . We use the symmetric initialization introduced in Section 2.4. For both GD and SGD, we normalize the error by the error at initialization. According to Corollary 1, under the symmetric initialization,  $\lambda_2(\Phi)$  provides a good characterization for linear  $f^*$ .

layers is significantly harder than that from only the last hidden layer. To see this, the analysis of the last hidden layer reduces to the one hidden layer analysis by treating the output from the second-to-last hidden layer as the input [DZPS19]. In particular, conditioning on the output from the second-to-last hidden layer, the NTK from the last hidden layer can be written as a sum of independent random variables. In contrast, the NTKs from intermediate layers not only depend on weights from previous layers but also subsequent layers. Thus, there is no similar conditional independence structure like the last hidden layer for us to utilize. As a result, a completely new method is needed to analyze the intermediate layers.

To see why it is hard to study SGD in the streaming setting, note that a critical step in existing analysis of GD on the batch setting [DLL<sup>+</sup>19] is to obtain the concentration of the finite-dimensional NTK matrix. There, pointwise concentration and union bounds suffice to obtain the desired convergence. However, in the analysis of SGD under the streaming setting, we need to show the uniform concentration of the infinite-dimensional

kernel function. Existing analysis techniques tailored to finite-dimensional kernel matrices are not enough to obtain the uniform convergence.

To overcome the above challenges, we first show the uniform concentration of NTK  $H_0$  defined in (2.1) at initialization to some deterministic kernel function and apply this to obtain the convergence of the average prediction error under SGD in the streaming setting.

In summary, we show

- For an  $L$ -layer fully connected feed-forward neural network  $f(x; \mathbf{W})$  with  $m$  neurons in each layer, we show that under Gaussian initialization, with high probability as  $m \rightarrow \infty$ , the NTK function from the  $\ell$ -th hidden layer  $H_0^{(\ell)}(x, x')$  defined in (2.2) at initialization concentrates on some deterministic function  $\Phi^{(\ell)}(x, x')$  uniformly for all  $x, x' \in \mathbb{S}^{d-1}$  and all layers  $1 \leq \ell \leq L$ ;
- We further apply the uniform concentration of NTK to show that with high probability as  $m \rightarrow \infty$ , the average prediction error under SGD at iteration  $T$  in the streaming setting is upper bounded by

$$\inf_{\ell \geq 1} \left\{ \prod_{t=0}^{T-1} (1 - \eta_t \lambda_\ell) \|\Delta_0\|_2 + \mathcal{R}(\Delta_0, \ell) \right\} + \mathfrak{Y},$$

where  $\eta_t$  is the step size at iteration  $t$ ,  $\lambda_\ell$  is the  $\ell$ -th eigenvalue of the integral operator  $\Phi$  associated with function  $\Phi = \sum_{\ell=1}^L \Phi^{(\ell)}$ ,  $\|\Delta_0\|_2$  is the prediction error at initialization, and  $\mathfrak{Y}$  is the error term capturing the approximation error from the non-linearity of ReLU function and the noise from stochastic gradients. Particularly, for an arbitrary small but fixed constant  $\epsilon > 0$ , by choosing an appropriate step size, we have  $\mathfrak{Y} < \epsilon$ , yielding a small average prediction error. In contrast to the characterization based on the NTK from only the last hidden layer, our analysis captures the contribution from all layers and provides a much tighter characterization of the average prediction error under SGD, as depicted in Figure 2.1b;

- On the technical front, to prove the convergence of the infinite-dimensional kernel function, one key step is to bound the number of activation patterns, that is, the

sign patterns of the ReLU units in all layers when varying the network input  $x$  while fixing the weights  $\mathbf{W}$ . Leveraging the recursive structure of the network in layers, we show that the number of sign patterns grows multiplicatively by a factor of  $m^d$  in every layer. This immediately implies that there are at most  $m^{dL}$  different activation patterns, despite that the network has  $mL$  neurons in total.

## 2.2 Previous works

There is a vast literature on overparametrized neural networks, and here we can only hope to cover a fraction of them we see the most relevant. A summary of the mostly related works on NTK is given in Table 2.1.

**Table 2.1:** Summary of related works

Literature	Error	Setting	Layer	Activation	Problem
[DZPS19] [SY19]	Training	Batch+GD	Single	ReLU	Regression
[DLL <sup>+</sup> 19]			Multi	Analytic	
[ADH <sup>+</sup> 19]	Generalization	Batch+GD	Single	ReLU	Regression
[CG19]		Stream+SGD	Multi		Classification
this chapter					Regression

To facilitate the discussion and better differentiate the algorithms, we use GD to denote the gradient descent algorithm where the entire batch is used to compute the gradient at each iteration, i.e., for the given batch  $\{(x_i, y_i)\}_{i=1}^n$  and a loss function  $\mathcal{L}(\cdot, \cdot)$ ,

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \frac{\eta t}{n} \sum_{i=1}^n \nabla_{\mathbf{w}} \mathcal{L}(f(x_i; \mathbf{W}(t)), y_i),$$

where  $\mathbf{W}(t)$  is the weight matrix at iteration  $t$ ,  $f(x; \mathbf{W}(t))$  is the neural network with parameter  $\mathbf{W}(t)$ . In contrast, our study focuses on the one-pass SGD, abbreviated as SGD, which draws a *single* fresh sample from the true data distribution to compute the gradient

at each iteration. In particular,

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \eta_t \nabla_{\mathbf{W}} \mathcal{L}(f(x_t; \mathbf{W}(t)), y_t), \quad (2.3)$$

where  $(x_t, y_t)$  is a freshly drawn sample at the  $t$ -th iteration from some unknown distribution  $\mu$ . The drawn sample  $(x_t, y_t)$  is then archived and not used any more.

**Training error with batch learning** For single-layer neural networks with a given batch  $\{(x_i, y_i)\}_{i=1}^n$ , it is shown in [DZPS19] that the NTK matrix  $\mathbf{H}_t = (\frac{1}{n} H_t(x_i, x_j))$ , concentrates on the deterministic matrix  $\Phi = (\frac{1}{n} \Phi(x_i, x_j))$  as the number of neurons goes to infinity. Then they utilize the spectrum of  $\Phi$  to prove that the training error of over-parametrized neural networks under GD converges at a linear rate  $[1 - \frac{\eta}{2} \lambda_{\min}(\Phi)]^t$ , where  $t$  is the number of iterations and  $\eta$  is the step size. The follow-up work [SY19] proves that as the sample size  $n$  grows,  $\lambda_{\min}(\Phi)$  decreases to 0 and hence the convergence rate can be very close to 0. Instead, they provide a different characterization showing that the training error under GD is upper bounded by

$$\left[1 - \frac{3\eta}{4} \lambda_r(\Phi)\right]^t + 2\sqrt{2}\mathcal{R}(\Delta_0, r) + \Theta\left(\frac{1}{\sqrt{n}}\right),$$

where  $\lambda_r(\Phi)$  is the  $r$ -th largest eigenvalue of the integral operator  $\Phi$  associated with the kernel function  $\Phi(x, x')$ , and  $\mathcal{R}(\Delta_0, r)$  is the  $L_2$  norm of the projection of  $\Delta_0 = f^*(x) - f(x; \mathbf{W}(0))$  onto the eigenspaces of kernel  $\Phi$  associated with  $\{\lambda_i(\Phi)\}_{i=r+1}^{\infty}$ . In addition, [DLL<sup>+</sup>19] extends the result of [DZPS19] to multi-layer neural networks with analytic activation functions. In particular, they first show the NTK matrix from the last hidden layer concentrates on some deterministic matrix  $\Phi^{(L)}$  and then characterize the GD dynamic utilizing the spectrum of  $\Phi^{(L)}$ . However, their analysis crucially utilizes the analytic property of the activation function, which does not cover the widely used ReLU-activated neural networks.

**Generalization error with batch learning** Apart from the training error, the generalization error which measures the accuracy of the model’s prediction on unseen data is also of wide interest. Following [DZPS19], [ADH<sup>+</sup>19] derives an upper bound of the generalization error of over-parameterized single layer neural networks under GD as

$$\mathbb{E}_{(X,y)\sim\mu} [\mathcal{L}(f(X; \mathbf{W}(t)), y)] \leq \frac{2y^\top \Phi^{-1}y}{n} + O\left(\sqrt{\frac{\log(n/\lambda_{\min}(\Phi))}{n}}\right),$$

where  $y = (y_1, y_2, \dots, y_n)^\top \in \mathbb{R}^n$  is the label of the i.i.d. sample  $\{(X_i, y_i)\}_{i=1}^n$  drawn from the distribution  $\mu$ . As mentioned above,  $\lambda_{\min}(\Phi)$  decreases to 0 and hence the generalization error can potentially blow up to infinity as  $n$  grows.

**Generalization error with streaming data** To learn a neural network in streaming setting, one way is to use SGD shown in (2.3). One work studying SGD with streaming data is [CG19] which focuses on the classification problem with the hinge loss function. Technically this work applies the online-to-batch conversion proposed in [CBCG04] to bound the generalization error  $\frac{1}{T} \sum_{s=1}^T \mathbb{E}_{(X,y)} [\mathbf{1}_{\{yf(X;W(s))<0\}}]$  from above by the empirical loss  $\frac{1}{T} \sum_{s=1}^T \mathcal{L}(y_s f(x_s; W(s)))$  with the hinge loss function  $\mathcal{L}(z) = \log(1 + \exp(-z))$ . Note that the online-to-batch conversion follows from an application of martingale concentration inequalities. It does not fully resolve the problem of bounding the generalization error as one still needs to bound the empirical loss. Indeed the authors bound the cumulative loss following a similar analysis of [DZPS19] and obtain an upper bound of the generalization error as

$$\frac{1}{T} \sum_{s=1}^T \mathbb{E}_{(X,y)} [\mathbf{1}_{\{yf(X;W(s))<0\}}] = O\left(\sqrt{\frac{y^\top (\Phi^{(L)})^{-1} y}{T}}\right) + O\left(\sqrt{1/T}\right),$$

where  $y = (y_1, \dots, y_T)^\top$ . However, as  $T$  increases,  $\lambda_{\min}(\Phi^{(L)})$  decreases to 0 and hence the upper bound may blow up.

## 2.3 Problem Setup

Suppose the data  $(X, y)$  is given by  $y = f^*(X) + u$ , where  $f^*$  is the underlying true function,  $X \in \mathbb{R}^d$  is the feature vector generated according to some distribution  $\mu$  on the unit sphere  $\mathbb{S}^{d-1}$ , and  $u$  is the bounded noise independent of  $X$  with mean 0, variance  $\tau^2$ . Denote  $\gamma \triangleq \max\{\|f^*\|_\infty, |u|\}$  which is independent of  $m$ .

We consider the following  $L$ -layer neural network, as illustrated in Figure 2.2:

$$f(x; \mathbf{W}) = a^\top \frac{1}{\sqrt{m}} \mathbf{D}^{(L)}(x) \mathbf{W}^{(L)} \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(1)}(x) \mathbf{W}^{(1)} x, \quad (2.4)$$

where  $a \in \mathbb{R}^{n_L}$  is the outer weight,  $\mathbf{W}^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$  is the weight of the  $\ell$ -th hidden layer whose  $i$ -th row is denoted as  $w_i^{(\ell)}$ ,

$$\mathbf{D}^{(\ell)}(x) = \text{diag} \left\{ \mathbf{1}_{\{\langle w_i^{(\ell)}, o^{(\ell-1)}(x) \rangle \geq 0\}} \right\} \in \mathbb{R}^{n_\ell \times n_\ell}, \quad (2.5)$$

and  $o^{(\ell)}(x)$  is the output of the  $\ell$ -th layer given by

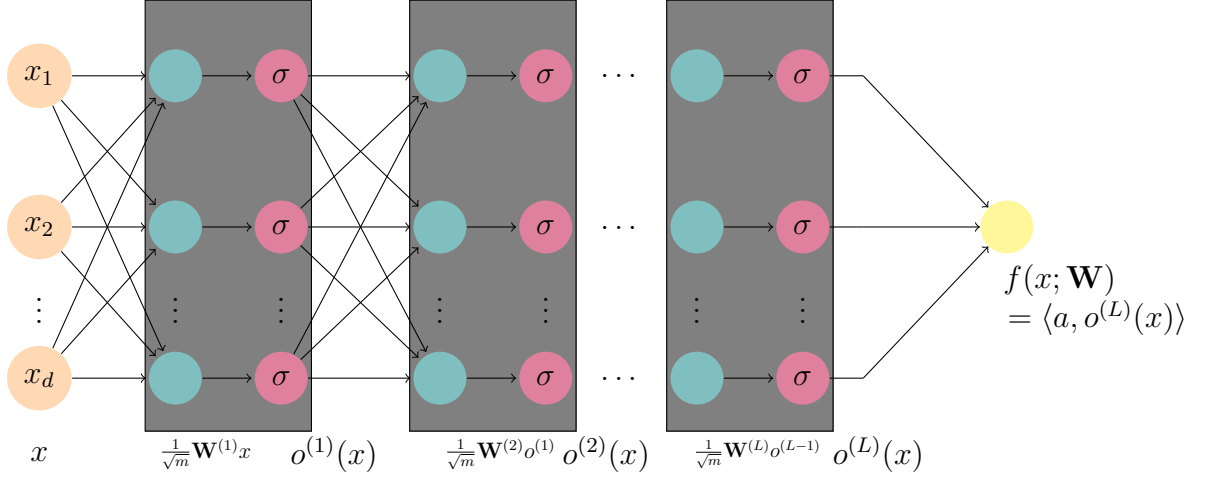
$$o^{(\ell)}(x) = \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x) \mathbf{W}^{(\ell)} \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(1)}(x) \mathbf{W}^{(1)} x \quad (2.6)$$

with  $o^{(0)}(x) = x$ .

The neural network is trained by running the stochastic gradient descent on the streaming data in one pass. In particular, given the initialization  $\{\mathbf{W}^{(\ell)}(0)\}_{\ell=1}^L$  and outer weight  $a$ , the  $\ell$ -th layer weight matrix at the  $t$ -th iteration is updated as

$$\begin{aligned} \mathbf{W}^{(\ell)}(t+1) &= \mathbf{W}^{(\ell)}(t) - \eta_t \frac{\partial \mathcal{L}(y_t, f(X_t; \mathbf{W}(t)))}{\partial \mathbf{W}^{(\ell)}} \\ &= \mathbf{W}^{(\ell)}(t) + \eta_t (y_t - f(X_t; \mathbf{W}(t))) \frac{\partial f(X_t; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}, \end{aligned} \quad (2.7)$$

where  $\eta_t$  is the step size,  $\mathcal{L}(y, \hat{y}) = \frac{1}{2} (y - \hat{y})^2$  is the quadratic loss function, and  $(X_t, y_t)$  is the freshly drawn data that is independent and identically distributed as  $(X, y)$ .



**Figure 2.2:** Illustration of Multi-Layer Neural Network  $f(x; \mathbf{W})$

To derive  $\frac{\partial f(x; \mathbf{W})}{\partial \mathbf{W}^{(\ell)}}$ , recall from (2.4) and (2.6) that

$$\begin{aligned}
 f(x; \mathbf{W}) &= a^\top \frac{1}{\sqrt{m}} \mathbf{D}^{(L)}(x) \mathbf{W}^{(L)} \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x) \mathbf{W}^{(\ell)} o^{(\ell-1)}(x) \\
 &= a^\top \left[ \mathbf{V}_L^{(\ell)}(x) \right]^\top \mathbf{W}^{(\ell)} o^{(\ell-1)}(x) \\
 &= \left\langle \mathbf{V}_L^{(\ell)}(x) a \left[ o^{(\ell-1)}(x) \right]^\top, \mathbf{W}^{(\ell)} \right\rangle,
 \end{aligned}$$

where

$$\left[ \mathbf{V}_L^{(\ell)}(x) \right]^\top \triangleq \frac{1}{\sqrt{m}} \mathbf{D}^{(L)}(x) \mathbf{W}^{(L)} \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell+1)}(x) \mathbf{W}^{(\ell+1)} \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x). \quad (2.8)$$

Thus, we get<sup>1</sup>

$$\frac{\partial f(x; \mathbf{W})}{\partial \mathbf{W}^{(\ell)}} = \mathbf{V}_L^{(\ell)}(x) a \left[ o^{(\ell-1)}(x) \right]^\top. \quad (2.9)$$

Plugging (2.9) into (2.7), we have

$$\mathbf{W}^{(\ell)}(t+1) = \mathbf{W}^{(\ell)}(t) + \eta_t (y_t - f(X_t; \mathbf{W}(t))) \mathbf{V}_{L,t}^{(\ell)}(x) a \left[ o_t^{(\ell-1)}(x) \right]^\top, \quad (2.10)$$

<sup>1</sup>Note that  $\{\mathbf{D}^{(k)}, k \geq \ell\}$  all depend on  $\mathbf{W}^{(\ell)}$ . However, each entry of  $\mathbf{D}^{(k)}$  only takes value 0 or 1, and hence does not change with  $\mathbf{W}^{(\ell)}$  once its value is fixed to be either 0 and 1.

where  $\mathbf{V}_{L,t}^{(\ell)}(x)$  is defined as  $\mathbf{V}_L^{(\ell)}(x)$  with  $\mathbf{W}$  replaced by  $\mathbf{W}(t)$ .

At  $t = 0$ , we initialize each weight matrix  $\mathbf{W}^{(\ell)}(0)$  as Gaussian random matrix with *i.i.d.* standard normal entry. We also generate outer weight  $a$  to be Rademacher (symmetric Bernoulli) random variable with equal probability to be  $-1$  or  $1$  which will be fixed throughout the training. This initialization is widely used in existing literature such as [DZPS19, ADH<sup>+</sup>19, SY19]. Furthermore, it has been shown in [JGH18] that the training dynamic of gradient descent method under this initialization is governed by the NTK defined in (2.1).

For ease of presentation, we assume  $\gamma = O(1)$ , the step size  $\eta_t \leq \frac{\theta}{t+1}$  for some  $\theta$  independent of  $d$  and  $m$  and  $n_1 = n_2 = \dots = n_L = m$ , i.e., all hidden layers have the same width, and consider the overparameterized regime where  $m$  tends to  $\infty$ . Such overparameterized neural networks have been the focus in the literature of NTK [AZLS19a, DLL<sup>+</sup>19].

## 2.4 Main Result

In Section 2.4.1, we show the uniform concentration of NTK. In Section 2.4.2, we apply the uniform concentration to derive an upper bound of the average prediction error under one-pass SGD.

### 2.4.1 Concentration of NTK at Initialization

In this section, we show the concentration of NTK at initialization. For notation simplicity, we abbreviate  $H_0$  as  $H$ ,  $H_0^{(\ell)}$  as  $H^{(\ell)}$ ,  $\mathbf{W}^{(\ell)}(0)$  as  $\mathbf{W}^{(\ell)}$ ,  $\mathbf{D}_0^{(\ell)}$  as  $\mathbf{D}^{(\ell)}$  and  $o_0^{(\ell)}$  as  $o^{(\ell)}$  for all  $\ell$  throughout this section and Section 2.5.

Note that the kernel function  $H$  is a sum of  $L$  kernel functions where  $H^{(\ell)}$  represents the contribution from the  $\ell$ -th hidden layer. To show the concentration of the kernel function  $H$ , it is sufficient to show the concentration of  $H^{(\ell)}$  for each  $1 \leq \ell \leq L$ .

To obtain the closed-form expression of  $H^{(\ell)}$ , we plug (2.9) into (2.2) and get

$$H^{(\ell)}(x, x') = \underbrace{\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x') \rangle}_{\text{(I)}} \times \underbrace{a^\top \mathbf{G}_L^{(\ell)}(x, x') a}_{\text{(II)}}, \quad (2.11)$$

where

$$\mathbf{G}_L^{(\ell)}(x, x') \triangleq \left[ \mathbf{V}_L^{(\ell)}(x) \right]^\top \mathbf{V}_L^{(\ell)}(x') \quad (2.12)$$

with  $\mathbf{V}_L^{(\ell)}(x)$  defined in (2.8).

Here, we provide a heuristic on obtaining the limiting function  $\Phi$ . Consider  $H^{(\ell)}$  in (2.11). For term (I), by the definition of  $o^{(\ell)}$ , we have the following recursion:

$$\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x') \rangle = \frac{1}{m} \sum_{i=1}^m \sigma(\langle w_i^{(\ell-1)}, o^{(\ell-2)}(x) \rangle) \sigma(\langle w_i^{(\ell-1)}, o^{(\ell-2)}(x') \rangle). \quad (2.13)$$

Conditioning on  $o^{(\ell-2)}$ , since  $w_i^{(\ell-1)}$  are i.i.d. Gaussian random vectors across  $i$ , we expect  $\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x') \rangle$  concentrates on its conditional mean, i.e.,

$$\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x') \rangle \rightarrow \mathbb{E}_{w \sim \mathcal{N}(0, \mathbf{I})} \left[ \sigma(\langle w, o^{(\ell-2)}(x) \rangle) \sigma(\langle w, o^{(\ell-2)}(x') \rangle) \right] \quad (2.14)$$

where

$$\left( \langle w, o^{(\ell-2)}(x) \rangle, \langle w, o^{(\ell-2)}(x') \rangle \right) \sim \mathcal{N} \left( 0, \begin{pmatrix} \|o^{(\ell-2)}(x)\|_2^2 & \langle o^{(\ell-2)}(x), o^{(\ell-2)}(x') \rangle \\ \langle o^{(\ell-2)}(x), o^{(\ell-2)}(x') \rangle & \|o^{(\ell-2)}(x')\|_2^2 \end{pmatrix} \right).$$

Analogous to (2.14), we show the covariance matrix on the right hand side of the above displayed equation concentrates on

$$\begin{pmatrix} \mathbb{E} [\sigma^2(\langle w, o^{(\ell-3)}(x) \rangle)] & \mathbb{E} [\sigma(\langle w, o^{(\ell-3)}(x) \rangle) \sigma(\langle w, o^{(\ell-3)}(x') \rangle)] \\ \mathbb{E} [\sigma(\langle w, o^{(\ell-3)}(x) \rangle) \sigma(\langle w, o^{(\ell-3)}(x') \rangle)] & \mathbb{E} [\sigma^2(\langle w, o^{(\ell-3)}(x') \rangle)] \end{pmatrix}.$$

In view of this recursive relation of  $(\langle w, o^{(\ell-2)}(x) \rangle, \langle w, o^{(\ell-2)}(x') \rangle)$ , we can approximate  $(\langle w, o^{(\ell-2)}(x) \rangle, \langle w, o^{(\ell-2)}(x') \rangle)$  by a pair of bivariate normal random variables. In particular,

we define  $(U^{(\ell-1)}(x), U^{(\ell-1)}(x'))$  such that

$$\begin{aligned} (U^{(\ell-1)}(x), U^{(\ell-1)}(x')) &\sim \mathcal{N}\left(0, \Sigma^{(\ell-2)}(x, x')\right) \\ \Sigma^{(\ell-2)}(x, x') &\triangleq \begin{pmatrix} \mathbb{E}[\sigma^2(U^{(\ell-2)}(x))] & \mathbb{E}[\sigma(U^{(\ell-2)}(x))\sigma(U^{(\ell-2)}(x'))] \\ \mathbb{E}[\sigma(U^{(\ell-2)}(x))\sigma(U^{(\ell-2)}(x'))] & \mathbb{E}[\sigma^2(U^{(\ell-2)}(x'))] \end{pmatrix} \end{aligned} \quad (2.15)$$

with  $\Sigma^{(0)}(x, x') = \begin{pmatrix} 1 & \langle x, x' \rangle \\ \langle x, x' \rangle & 1 \end{pmatrix}$ , and show that

$$(I) \rightarrow \mathbb{E} \left[ \sigma \left( U^{(\ell-1)}(x) \right) \sigma \left( U^{(\ell-1)}(x') \right) \right]. \quad (2.16)$$

For (II), conditioning on weight matrices  $\{\mathbf{W}^{(k)}\}_{k=1}^L$ , we have

$$(II) \rightarrow \mathbb{E}_a \left[ a^\top \mathbf{G}_L^{(\ell)} a \right] = \text{Tr}(\mathbf{G}_L^{(\ell)}).$$

Moreover, crucially  $\text{Tr}(\mathbf{G}_L^{(\ell)})$  approximately satisfies a recursion. In particular, by the definition of  $\mathbf{G}_L^{(\ell)}$ , for any fixed  $\ell \leq L$ ,

$$\begin{aligned} &\text{Tr} \left( \mathbf{G}_{L+1}^{(\ell)}(x, x') \right) \\ &= \text{Tr} \left( \mathbf{D}^{(L+1)}(x) \mathbf{W}^{(L+1)} \mathbf{G}_L^{(\ell)}(x, x') \left[ \mathbf{W}^{(L+1)} \right]^\top \mathbf{D}^{(L+1)}(x') \right) \\ &= \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(L+1)}, o^{(L)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(L+1)}, o^{(L)}(x') \rangle \geq 0\}} \left[ w_i^{(L+1)\top} \mathbf{G}_L^{(\ell)}(x, x') w_i^{(L+1)} \right] \\ &\rightarrow \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(L+1)}, o^{(L)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(L+1)}, o^{(L)}(x') \rangle \geq 0\}} \text{Tr} \left( \mathbf{G}_L^{(\ell)}(x, x') \right), \end{aligned} \quad (2.17)$$

where the last assertion holds because  $w^\top \mathbf{G}_L^{(\ell)}(x, x') w$  concentrates on its mean

$\text{Tr} \left( \mathbf{G}_L^{(\ell)}(x, x') \right)$ .

When  $\ell = L + 1$ , we know  $\mathbf{V}_{L+1}^{(L+1)}(x) = \frac{1}{\sqrt{m}} \mathbf{D}^{(L+1)}(x)$  in view of (2.8). From (2.12), we

get  $\mathbf{G}_{L+1}^{(L+1)}(x, x') = \frac{1}{m} \mathbf{D}^{(L+1)}(x) \mathbf{D}^{(L+1)}(x')$  and

$$\mathrm{Tr} \left( \mathbf{G}_{L+1}^{(L+1)}(x, x') \right) = \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(L+1)}, o^{(L)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(L+1)}, o^{(L)}(x') \rangle \geq 0\}}.$$

Furthermore,

$$\begin{aligned} & \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(L+1)}, o^{(L)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(L+1)}, o^{(L)}(x') \rangle \geq 0\}} \\ & \rightarrow \mathbb{E}_{w \sim \mathcal{N}(0, \mathbf{I})} \left[ \mathbf{1}_{\{\langle w, o^{(L)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w, o^{(L)}(x') \rangle \geq 0\}} \right] \\ & \rightarrow \frac{\pi - \arccos \rho^{(L)}(x, x')}{2\pi}, \end{aligned} \tag{2.18}$$

where the first step holds by conditioning on  $o^{(L)}$ , and the last line follows as

$$\left\langle \frac{o^{(L)}(x)}{\|o^{(L)}(x)\|_2}, \frac{o^{(L)}(x')}{\|o^{(L)}(x')\|_2} \right\rangle \rightarrow \frac{\mathbb{E} [\sigma(U^{(L)}(x)) \sigma(U^{(L)}(x'))]}{\sqrt{\mathbb{E} [\sigma^2(U^{(L)}(x))]} \sqrt{\mathbb{E} [\sigma^2(U^{(L)}(x'))]}} \triangleq \rho^{(L)}(x, x').$$

Therefore, by defining

$$\begin{aligned} q_{L+1}^{(\ell)}(x, x') &= \frac{\pi - \arccos \rho^{(L)}(x, x')}{2\pi} q_L^{(\ell)}(x, x'), \quad \forall \ell \leq L, \\ q_{L+1}^{(L+1)}(x, x') &= \frac{\pi - \arccos \rho^{(L)}(x, x')}{2\pi}, \end{aligned} \tag{2.19}$$

we get that

$$\text{(II)} \rightarrow q_L^{(\ell)}(x, x'). \tag{2.20}$$

Combining (2.16) and (2.20), we get that

$$H^{(\ell)}(x, x') \rightarrow \mathbb{E} \left[ \sigma \left( U^{(\ell-1)}(x) \right) \sigma \left( U^{(\ell-1)}(x') \right) \right] q_L^{(\ell)}(x, x') \triangleq \Phi^{(\ell)}(x, x'). \tag{2.21}$$

It has been shown in [JGH18] that for fixed  $(x, x')$  and fixed  $\ell$ ,  $H^{(\ell)}(x, x')$  converges to  $\Phi^{(\ell)}(x, x')$  in probability. The following theorem strengthens their result, showing the

uniform convergence of  $H^{(\ell)}$  to  $\Phi^{(\ell)}$  for all  $\ell$  and characterizing the rate of the convergence.

**Theorem 1.** *Under Gaussian initialization, For  $m \geq Cd^2 \exp(L^2)$  for some constant  $C$ , there exist constants  $C_1, C_2$  and  $C_3$  such that, with probability at least  $1 - \exp(-C_1 m^{1/3})$ ,*

$$\|H^{(\ell)} - \Phi^{(\ell)}\|_{\infty} \leq C_2 \left( \frac{C_3^L}{m^{1/6}} + \sqrt{\frac{dL \log m}{m}} \right), \quad \forall 1 \leq \ell \leq L. \quad (2.22)$$

**Remark 2.4.1.** *Theorem 1 significantly improves the concentration bounds in [DLL<sup>+</sup>19]. Specifically, [DLL<sup>+</sup>19] only establishes the concentration of the last hidden layer  $H^{(L)}(x_i, x_j)$  for a bounded number of data points  $\{x_i\}_{i=1}^n$ . In contrast, Theorem 1 establishes the concentration uniformly over all  $x \in \mathbb{S}^{d-1}$  and for all layers  $\ell \in [L]$ , which is much stronger and more challenging to obtain. To see why, note that a simple pointwise control and union bounds fall short of proving the uniform concentration over all  $x \in \mathbb{S}^{d-1}$ . More importantly, from the definition (2.12), we know*

$$H^{(L)} = \langle o^{(L-1)}(x), o^{(L-1)}(x') \rangle \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x') \rangle \geq 0\}}.$$

is a sum of independent random variables conditioning on  $o^{(L-1)}$  for which a simple concentration inequality can be applied to. In contrast, the intermediate layer  $H^{(\ell)}$  with  $\ell < L$  depends on not only previous hidden layers but also weight matrices and activation patterns of subsequent layers through  $\mathbf{G}_L^{(\ell)}$ . To overcome these challenges, we fix  $\mathbf{G}_L^{(\ell)}$  and view  $a^\top \mathbf{G}_L^{(\ell)}(x, x')a$  as a quadratic term. This allows us to apply Hanson-Wright inequality [Ver19, Theorem 6.2.1] and obtain the concentration of  $a^\top \mathbf{G}_L^{(\ell)}(x, x')a$  for any fixed  $x, x'$ . To upgrade this point-wise concentration to the uniform one, we utilize the critical observation that the number of different  $\mathbf{G}_L^{(\ell)}(x, x')$  for fixed  $\{\mathbf{W}^{(\ell)}\}_{\ell=1}^L$  depends on the number of activation patterns  $|\mathcal{D}_L|$  where  $\mathcal{D}_L \triangleq \{(\mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(L)}(x)) : x \in \mathbb{S}^{d-1}\}$ . We then show  $|\mathcal{D}_L| \leq m^{dL}$  through showing  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$ . See Lemma 2.5.2 for more details.

**Implications in batch setting** Theorem 1 implies that if  $m = \Omega(\exp(L^2) \text{poly}(d, \frac{1}{\epsilon}))$ , then  $\|H - \Phi\|_{\infty} < \epsilon$  with high probability. Interestingly, our uniform bounds enable us to

derive a sufficient condition on the over-parameterization  $m$  in the batch setting that is independent of the batch size. Specifically, in the batch setting with data points  $\{x_i\}_{i=1}^n$ , by defining kernel matrices  $\mathbf{H} = (\frac{1}{n}H(x_i, x_j)) \in \mathbb{R}^{n \times n}$  and  $\Phi = (\frac{1}{n}\Phi(x_i, x_j)) \in \mathbb{R}^{n \times n}$ , we can deduce that  $\|\mathbf{H} - \Phi\|_F \leq \|H - \Phi\|_\infty < \epsilon$ . In contrast, the previous works in the batch setting [DZPS19, DLL<sup>+</sup>19, SY19] require  $m$  to grow sufficiently fast in  $n$  to ensure  $\|\mathbf{H} - \Phi\|_F \leq \epsilon$ . For example, [DZPS19] requires that  $m = \Omega(n^6)$ .

In addition to the above application, Theorem 1 also plays an important role in the analysis of gradient descent dynamic. Existing work [DLL<sup>+</sup>19] shows the training error under GD decays at the rate of  $(1 - \eta\lambda_{\min}(\Phi^{(L)})/2)^t$  where  $\Phi^{(L)} = (\frac{1}{n}\Phi^{(L)}(x_i, x_j))$  is the limit of the NTK matrix from the last hidden layer as the number of neurons goes to infinity. With Theorem 1, we are able to show a tighter rate  $(1 - \eta\lambda_{\min}(\Phi)/2)^t$ .

Beyond the application in batch setting, Theorem 1 further enables us to characterize the convergence of the prediction error under SGD in the streaming data setting, as we shall present next.

## 2.4.2 Average Prediction Error under SGD

Define the prediction error  $\Delta_t(x) \triangleq f^*(x) - f(x; \mathbf{W}(t))$ . We aim to characterize the convergence of the average prediction error  $\|\Delta_t\|_2 \triangleq \sqrt{\mathbb{E}_X [\Delta_t^2(X)]}$ .

To analyze  $\|\Delta_t\|_2$ , we first show a linear approximation of  $\Delta_t$ :

$$\Delta_{t+1} = (\mathbf{I} - \eta_t \mathbf{H}_t) \Delta_t + v_t + \epsilon_t, \quad (2.23)$$

where  $\mathbf{I}$  is the identity operator,  $\mathbf{H}_t$  is the integral operator associated with the kernel function  $H_t(x, x')$ ,  $v_t$  is the noise from the stochastic gradient, and  $\epsilon_t$  is the approximation error.

Note that  $H_t$  depends on  $\{\mathbf{W}(s), s \leq t\}$  and hence further depends on the sample path  $\{X_s, y_s\}_{s=0}^{t-1}$ . To circumvent this dependency, we first show  $\mathbf{W}(t)$  stays relatively close to  $\mathbf{W}(0)$  in operator norm under the over-parameterized regime with large  $m$ . This further

allows us to show  $\|H_t - H\|_\infty$  is small. Applying the triangle inequality together with Theorem 1, we deduce that  $\|H_t - \Phi\|_\infty$  is small. It then follows from (2.23) that the prediction error under SGD can be approximated by a linear dynamic governed by  $\Phi$  for any sample path  $\{X_t, y_t\}$ :

$$\Delta_{t+1} = (1 - \eta_t \Phi) \Delta_t + \eta_t (\Phi - H_t) \Delta_t + v_t + \epsilon_t, \quad (2.24)$$

where  $\Phi$  is the integral operator associated with  $\Phi$  defined in (2.21). This recursion reveals that the evolution of  $\Delta_t$  is governed by the spectrum of  $\Phi$ .

More specifically, denote the eigenvalues of  $\Phi$  as  $\{\lambda_i\}_{i=1}^\infty$  with  $\lambda_1 \geq \lambda_2 \geq \dots$  and the corresponding eigen-functions  $\phi_i$ . For any function  $g \in L_2(\mu)$ , denote the residual projection error  $\mathcal{R}(g, r)$  as the  $L_2$  norm of the projection of  $g$  onto the space spanned by eigen-functions  $\{\phi_i\}_{i=r+1}^\infty$ , i.e.,

$$\mathcal{R}(g, r) = \sqrt{\sum_{i=r+1}^\infty \langle g, \phi_i \rangle^2}. \quad (2.25)$$

**Theorem 2.** *Given  $m \geq C_3 d^7 \exp(\theta C^L \log T)$  and  $\eta_t = \frac{\theta}{t+1}$  for  $\theta < \frac{9}{2\sqrt{44L}}$ , with probability at least  $1 - \exp(-C_4^{-L} m^{1/36})$  over the initialization, we have*

$$\mathbb{E} [\|\Delta_t\|_2] \leq \inf_\ell \left\{ \left( \prod_{s=0}^{t-1} (1 - \eta_s \lambda_\ell) \right) \|\Delta_0\|_2 + \mathcal{R}(\Delta_0, \ell) \right\} + 2c_2 \|\Delta_0\|_2 + 2c_2 \tau, \quad (2.26)$$

where  $c_2 = \theta L e^{\sqrt{44L}\theta/9} \sqrt{\frac{1}{1-2\sqrt{44L}\theta/9} + 1}$ .

Here, the first term on the right hand side of (2.26) comes from the linear approximation  $\Delta_{t+1} \approx (1 - \eta_t \Phi) \Delta_t \approx \prod_{s=0}^t (1 - \eta_s \Phi) \Delta_0$  in view of (2.24). The term  $2c_2 \|\Delta_0\|_2 + 2c_2 \tau$  is the sum of three errors. One is the accumulation of the perturbation error  $v_t$  from the stochastic gradients. Another is the accumulation of the approximation error  $\epsilon_t$  from the use of the linear approximation. The last one is the accumulation of the approximation error  $\eta_t (\Phi - H_t) \Delta_t$ .

From Theorem 2, we see that an early stopping time  $T$ , which is commonly used [SY19, AZLS19a], is needed to ensure the condition on the number of neurons per layer  $m$  is

satisfied. Intuitively, this dependency on  $T$  comes from two aspects. Firstly, to ensure the linear approximation holds, we crucially require  $\mathbf{W}(t)$  to be close to  $\mathbf{W}(0)$ , resulting in an upper bound on the number of SGD iterations  $T$ . Secondly, the accumulation of the approximation error  $\epsilon_t$ , albeit vanishing in  $m$ , grows in the number of iterations  $T$ . Thus to ensure the final approximation error is small, we need  $m$  to be sufficiently large compared with  $T$ .

Our result sheds light on the trade-off between the convergence rate and the accumulation of approximation errors. The trade-off is two-fold. One is between  $\prod_{s=0}^t (1 - \eta_s \lambda_\ell)$  and  $\mathcal{R}(\Delta_0, \ell)$  through  $\ell$ . Intuitively, on one hand, a larger  $\ell$  implies a larger principal space which yields a smaller  $\mathcal{R}(\Delta_0, \ell)$ . On the other hand, a larger  $\ell$  also implies a smaller  $\lambda_\ell$ . Thus, the contraction factor  $\prod_{s=0}^t (1 - \eta_s \lambda_\ell)$  is smaller, indicating slower convergence. The other trade-off is between the contraction factor  $\prod_{s=0}^t \left(1 - \frac{\theta \lambda_\ell}{s+1}\right)$  and the accumulation of approximation error and noise  $c_2$  through  $\theta$ . To make sure  $c_2$  is small, we need small  $\theta$ , thus yielding a small contraction factor. In return, we need more iterations to converge.

Now we present an application of Theorem 2 when  $f^*$  is a polynomial. Consider SGD under a symmetric initialization scheme of the last layer, i.e.,  $\mathbf{W}^{(L)}(0) = \begin{pmatrix} \mathbf{W} \\ \mathbf{W} \end{pmatrix}$  where  $\mathbf{W} \in \mathbb{R}^{\frac{m}{2} \times m}$  is a random matrix with i.i.d. standard normal entries and  $a = (b, -b)^\top$  where  $b \in \mathbb{R}^{m/2}$  has i.i.d. Rademacher entries.

**Corollary 1.** *Assume  $f^*$  is a degree  $\ell^*$  polynomial and the input data follows the uniform distribution over  $\mathbb{S}^{d-1}$ . Under the same condition as Theorem 2, we have with probability at least  $1 - \exp\left(-\Omega(C_4^{-L} m^{1/36})\right)$ ,*

$$\mathbb{E} [\|\Delta_{t+1}\|_2 | \mathbf{W}(0), a] \leq \prod_{s=0}^t (1 - \eta_s \lambda_{\ell^*+1}) \|f^*\|_2 + 2c_2 \|f^*\|_2 + 2c_2 \tau. \quad (2.27)$$

The proof is deferred to Section 3.4.

**Remark 2.4.2.** *From Corollary 1, for arbitrarily small constant  $\epsilon$ , by choosing small step sizes, a sufficiently long horizon and a sufficient wide neural network, we ensure that the average prediction error under SGD is smaller than  $\epsilon$ . To be more specific, for any  $0 <$*

$\epsilon < \|f^*\|_2 + \tau$ , by choosing  $T \geq \left(\frac{\epsilon}{6\|f^*\|_2}\right)^{-1/(\theta\lambda_{\ell^*+1})}$  and  $\theta \leq \frac{9\epsilon}{8\sqrt{44}(\|f^*\|_2 + \tau)L}$ , we ensure  $\mathbb{E}[\|\Delta_{t+1}\|_2 | \mathbf{W}(0), a] \leq \epsilon$ . To see why, note that

$$\prod_{s=0}^t (1 - \eta_s \lambda_{\ell^*+1}) \leq \exp(-\theta \lambda_{\ell^*+1} \log T) = T^{-\theta \lambda_{\ell^*+1}} \leq \frac{\epsilon}{6\|f^*\|_2}, \quad (2.28)$$

and

$$\begin{aligned} c_2(\|f^*\|_2 + \tau) &= \theta L e^{\sqrt{44}L\theta/9} \sqrt{\frac{1}{1 - 2\sqrt{44}L\theta/9} + 1} (\|f^*\|_2 + \tau) \\ &\stackrel{(a)}{\leq} \frac{9\epsilon}{8\sqrt{44}} e^{1/8} \sqrt{7/3} \leq \frac{5}{12}\epsilon, \end{aligned} \quad (2.29)$$

where (a) holds since  $\frac{2\sqrt{44}L\theta}{9} \leq \frac{\epsilon}{4(\|f^*\|_2 + \tau)} < \frac{1}{4}$ . The result follows by plugging (2.28) and (2.29) into (2.27).

## 2.5 Proof of Theorem 1

**Additional notation** Define  $\text{VC}(\mathcal{F})$  as the VC dimension of Boolean function class  $\mathcal{F}$ . For any matrix  $\mathbf{C} \in \mathbb{R}^{n \times m}$ , we define  $\|\mathbf{C}\|_\infty \triangleq \max_{1 \leq i \leq n, 1 \leq j \leq m} |\mathbf{C}_{ij}|$ . Throughout the remaining Chapter 2 and 3, we use  $C$  to denote absolute constant whose value may vary in lines.

We present several key lemmas that will be used in the proof of Theorem 1. First, we show that  $\langle o^{(\ell)}(x), o^{(\ell)}(x') \rangle$  concentrates on  $\mathbb{E}[\sigma(U^{(\ell)}(x)) \sigma(U^{(\ell)}(x'))]$  uniformly over all  $x, x' \in \mathbb{S}^{d-1}$  and all  $\ell \in [L]$ .

**Lemma 2.5.1.** *With probability at least  $1 - L \exp(O(d \log m) - \Omega(m^{1/3}))$ , for any  $1 \leq \ell \leq L$ ,*

$$\sup_{x, x'} \left| \langle o^{(\ell)}(x), o^{(\ell)}(x') \rangle - \mathbb{E}[\sigma(U^{(\ell)}(x)) \sigma(U^{(\ell)}(x'))] \right| = O\left(\frac{\ell C^{2\ell}}{m^{1/3}}\right), \quad (2.30)$$

where  $(U^{(\ell)}(x), U^{(\ell)}(x'))$  is defined in (2.15).

To prove Lemma 2.5.1, we follow the aforementioned heuristic in Section 2.4.1 to show that  $\langle o^{(\ell)}(x), o^{(\ell)}(x') \rangle$  concentrates on  $\mathbb{E} [\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))]$  for any fixed  $(x, x')$ . Then we establish that  $o^{(\ell)}(x)$  is Lipschitz in  $x$  with high probability. This enables us to apply an  $\epsilon$ -net argument to upgrade the pointwise concentration to the uniform one.

The next two lemmas together show that  $a^\top \mathbf{G}_L^{(\ell)}(x, x')a$  uniformly concentrates on  $q_L^{(\ell)}(x, x')$ .

**Lemma 2.5.2.** *With probability at least  $1 - \exp(O(dL \log m) - \Omega(m^{1/3}))$ , for  $\ell = 1, 2, \dots, L$ ,*

$$\sup_{x, x'} \left| a^\top \mathbf{G}_L^{(\ell)}(x, x')a - \text{Tr} \left( \mathbf{G}_L^{(\ell)}(x, x') \right) \right| = O \left( \frac{c_0^{2L-2\ell}}{m^{1/3}} \right). \quad (2.31)$$

The above lemma shows the uniform concentration of  $a^\top \mathbf{G}_L^{(\ell)}(x, x')a$  on  $\text{Tr}(\mathbf{G}_L^{(\ell)}(x, x'))$ . However, unlike the previous case, an  $\epsilon$ -net argument cannot be applied here, as  $x$  influences  $\mathbf{G}_L^{(\ell)}(x, x')$  through non-Lipschitz indicator functions  $\mathbf{1}_{\{\langle w^{(k+1)}, o^{(k)}(x) \rangle \geq 0\}}$  for  $k \geq \ell$ . As mentioned in Remark 2.4.1, the key to overcome this challenge lies on the following crucial observation. Although there are infinite number of different matrices  $\mathbf{G}_L^{(\ell)}(x, x')$  when varying  $x, x'$ , conditioning on  $\{\mathbf{W}^{(k)}\}_{k=1}^L$  the size of  $\mathcal{G}_L^{(\ell)} \triangleq \{\mathbf{G}_L^{(\ell)}(x, x') : x, x' \in \mathbb{S}^{d-1}\}$  depends only on the size of

$$\mathcal{D}_L \triangleq \left\{ \left( \mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(L)}(x) \right) : x \in \mathbb{S}^{d-1} \right\}.$$

Since  $\mathbf{D}^{(k)} \in \mathbb{R}^{m \times m}$  is diagonal with binary entries, one can directly bound  $|\mathcal{D}_L|$  by  $2^{mL}$ . Unfortunately, such naive bound is too loose to obtain a tight concentration. Instead, we show a much tighter bound  $|\mathcal{D}_L| \leq m^{dL}$  utilizing the recursive relation  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$  for all  $k$ . To obtain such recursive relation, a critical step is to decompose  $\mathbb{S}^{d-1}$  into disjoint regions  $\{V_j, j = 1, 2, \dots, |\mathcal{D}_{k-1}|\}$  so that for any  $x$  within the same  $V_j$ ,  $(\mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(k-1)}(x))$  is the same. With such decomposition, we can get

$$|\mathcal{D}_k| \leq \sum_{j=1}^{|\mathcal{D}_{k-1}|} \left| \left\{ \mathbf{D}^{(k)}(x) : x \in V_j \right\} \right|.$$

To further bound  $\left| \left\{ \mathbf{D}^{(k)}(x) : x \in V_j \right\} \right|$ , we crucially utilize the fact that for any fixed  $j$ ,

$o^{(k-1)}(x) = P_j x$  for all  $x \in V_j$  with some deterministic matrix  $P_j \in \mathbb{R}^{m \times d}$  independent of  $x$ . Hence,  $|\{\mathbf{D}^{(k)}(x) : x \in V_j\}| \leq m^d$  follows by applying [HR19, Proposition 7.1] and Sauer-Shelah Lemma (Lemma A.2.3). With this tighter bound in hand, we deduce the uniform concentration of  $a^\top \mathbf{G}_L^{(\ell)}(x, x') a$  on its mean  $\text{Tr}(\mathbf{G}_L^{(\ell)}(x, x'))$  by combining Hanson-Wright inequality with a union bound over  $\mathcal{G}_L^{(\ell)}$ .

It remains to show the uniform concentration of  $\text{Tr}(\mathbf{G}_L^{(\ell)}(x, x'))$  on  $q_L^{(\ell)}(x, x')$ .

**Lemma 2.5.3.** *With probability at least  $1 - \exp(-O(dL \log m)) - \Omega(m^{1/3})$ , for  $\ell = 1, 2, \dots, L$ ,*

$$\sup_{x, x'} \left| \text{Tr}(\mathbf{G}_L^{(\ell)}(x, x')) - q_L^{(\ell)}(x, x') \right| = O\left(\frac{\sqrt{L} C^L}{m^{1/6}} + \sqrt{\frac{d(1 + (L-1) \log m)}{m}}\right) \quad (2.32)$$

for some universal constant  $C$ .

To prove Lemma 2.5.3, we follow the heuristic argument in Section 2.4.1 to prove (2.17) and (2.18). The proof of (2.17) follows similarly as that of Lemma 2.5.2. To prove the first step of (2.18), we utilize the following observation. Conditioning on  $\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(L-1)}$ , the change of  $\sup_{x, x'} h^{(L)}(x, x')$  from the change of any single coordinate is bounded by  $\frac{1}{m}$ , where

$$\begin{aligned} & h^{(L)}(x, x') \\ & \triangleq \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x') \rangle \geq 0\}} - \mathbb{E}_w \left[ \mathbf{1}_{\{\langle w, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w, o^{(L-1)}(x') \rangle \geq 0\}} \right] \right|. \end{aligned}$$

This allows us to apply McDiarmid's inequality to show with high probability over the randomness of  $\{w_i^{(L)}\}_{i=1}^m$ ,  $\sup_{x, x'} h^{(L)}(x, x')$  concentrates on its mean. We then apply Lemma A.2.2 to bound  $\mathbb{E}[\sup_{x, x'} h^{(L)}(x, x')]$  by  $O\left(\sqrt{\frac{\text{VC}(\mathcal{H}^{(L)})}{m}}\right)$  where

$$\mathcal{H}^{(L)} \triangleq \left\{ f_{x, x'}(w) = \mathbf{1}_{\{\langle w, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w, o^{(L-1)}(x') \rangle \geq 0\}} : x, x' \in \mathbb{S}^{d-1} \right\}.$$

Afterwards, we apply Lemma A.2.1 to show  $\text{VC}(\mathcal{H}^{(L)}) = O(\text{VC}(\mathcal{F}^{(L)}))$  where

$$\mathcal{F}^{(L)} \triangleq \left\{ f_x(w) = \mathbf{1}_{\{\langle w, o^{(L-1)}(x) \rangle \geq 0\}} : x \in \mathbb{S}^{d-1} \right\}.$$

To bound  $\text{VC}(\mathcal{F}^{(L)})$ , we follow a similar decomposition strategy as Lemma 2.5.2 and show

$$\text{VC}(\mathcal{F}^{(L)}) = O(d(1 + (L - 1) \log m)).$$

To prove the second step of (2.18), we crucially establish that the arccos function is Hölder continuous of order 1/2 despite that it is non-Lipschitz.

With the above lemmas, we now present the proof of Theorem 1. The full proofs of Lemma 2.5.1– 2.5.3 are deferred to Section 3.1.

*Proof of Theorem 1.* Throughout the proof, we condition on the event such that (2.30), (2.31) and (2.32) hold simultaneously. By Lemma 2.5.1–Lemma 2.5.3, we get such event occurs with probability at least  $1 - \exp(-\Omega(m^{1/3}))$  for sufficiently large  $m$ .

For any  $1 \leq \ell \leq L$ , by the triangle inequality, we have

$$\begin{aligned} & \left\| H^{(\ell)} - \Phi^{(\ell)} \right\|_{\infty} \\ &= \sup_{x, x'} \left| \langle o^{(\ell-1)}(x), o^{(\ell-1)}(x') \rangle a^{\top} \mathbf{G}_L^{(\ell)}(x, x') a - \mathbb{E} \left[ \sigma \left( U^{(\ell-1)}(x) \right) \sigma \left( U^{(\ell-1)}(x') \right) \right] q_L^{(\ell)}(x, x') \right| \\ &\leq \sup_{x, x'} \left| \left( \langle o^{(\ell-1)}(x), o^{(\ell-1)}(x') \rangle - \mathbb{E} \left[ \sigma \left( U^{(\ell-1)}(x) \right) \sigma \left( U^{(\ell-1)}(x') \right) \right] \right) a^{\top} \mathbf{G}_L^{(\ell)}(x, x') a \right| \\ &+ \sup_{x, x'} \left| \mathbb{E} \left[ \sigma \left( U^{(\ell-1)}(x) \right) \sigma \left( U^{(\ell-1)}(x') \right) \right] \left( a^{\top} \mathbf{G}_L^{(\ell)}(x, x') a - q_L^{(\ell)}(x, x') \right) \right|. \end{aligned} \quad (2.33)$$

Here, we claim that

$$\sup_{x, x'} \left| a^{\top} \mathbf{G}_L^{(\ell)}(x, x') a \right| \leq 1, \quad (2.34)$$

and

$$\sup_{x, x'} \mathbb{E} \left[ \sigma(U^{(\ell)}(x)) \sigma(U^{(\ell)}(x')) \right] \leq \sup_x \sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x))]} \sup_{x'} \sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x'))]} = 2^{-\ell} \leq 1. \quad (2.35)$$

Plugging the above two claims into (2.33), we have

$$\begin{aligned}
& \left\| H^{(\ell)} - \Phi^{(\ell)} \right\|_{\infty} \\
& \leq \sup_{x, x'} \left| \langle o^{(\ell-1)}(x), o^{(\ell-1)}(x') \rangle - \mathbb{E} \left[ \sigma \left( U^{(\ell-1)}(x) \right) \sigma \left( U^{(\ell-1)}(x') \right) \right] \right| \\
& \quad + \sup_{x, x'} \left| a^{\top} \mathbf{G}_L^{(\ell)}(x, x') a - q_L^{(\ell)}(x, x') \right| \\
& \stackrel{(a)}{=} O \left( \frac{\ell C^{2\ell}}{m^{1/3}} \right) + O \left( \frac{C^L}{m^{1/6}} + \sqrt{\frac{d(1+(L-1)\log m)}{m}} \right) \\
& = O \left( \frac{C^L}{m^{1/6}} + \sqrt{\frac{d(1+(L-1)\log m)}{m}} \right),
\end{aligned}$$

where (a) holds by (2.30), (2.31), and (2.32); and the last equality holds since  $m = \Omega(\exp(L^2))$ .

It remains to prove (2.34) and (2.35). To prove (2.34), by definition (2.19), we have

$$0 \leq q_L^{(\ell)}(x, x') \leq 1/2. \quad (2.36)$$

Therefore, by the triangle inequality, we have

$$\sup_{x, x'} \left| a^{\top} \mathbf{G}_L^{(\ell)}(x, x') a \right| \leq \sup_{x, x'} \left| a^{\top} \mathbf{G}_L^{(\ell)}(x, x') a - q_L^{(\ell)}(x, x') \right| + \sup_{x, x'} \left| q_L^{(\ell)}(x, x') \right| \leq 1,$$

where the last inequality holds since  $O \left( \frac{\sqrt{\ell} C^L}{m^{1/6}} + \sqrt{\frac{d(1+(L-1)\log m)}{m}} \right) \leq \frac{1}{2}$  given  $m = \Omega(d^2 \exp(L^2))$ .

Now we prove (2.35). Since  $U^{(1)}(x) = \langle w, x \rangle \sim \mathcal{N}(0, 1)$  for any  $x$ , we have

$$\mathbb{E} \left[ \sigma^2(U^{(1)}(x)) \right] = \mathbb{E}_{Z \sim \mathcal{N}(0,1)} \left[ Z^2 \mathbf{1}_{\{Z \geq 0\}} \right] = \frac{1}{2}, \quad \forall x. \quad (2.37)$$

By the definition of  $\Sigma^{(\ell)}$ , it follows that

$$\begin{aligned}\mathbb{E} \left[ \sigma^2(U^{(\ell)}(x)) \right] &= \mathbb{E}_{U^{(\ell-1)}(x)} \left[ \mathbb{E}_{Z \sim \mathcal{N}(0, \sigma^2(U^{(\ell-1)}(x)))} \left[ Z^2 \mathbf{1}_{\{Z \geq 0\}} | U^{(\ell-1)}(x) \right] \right] \\ &= \frac{1}{2} \mathbb{E} \left[ \sigma^2(U^{(\ell-1)}(x)) \right], \quad \forall x.\end{aligned}$$

Recursively applying the above equality and noting (2.37), we get that

$$\mathbb{E} \left[ \sigma^2(U^{(\ell)}(x)) \right] = 2^{-\ell}, \quad \forall x. \quad (2.38)$$

By Cauchy-Schwartz inequality, we have

$$\sup_{x, x'} \mathbb{E} \left[ \sigma(U^{(\ell)}(x)) \sigma(U^{(\ell)}(x')) \right] \leq \sup_x \sqrt{\mathbb{E} \left[ \sigma^2(U^{(\ell)}(x)) \right]} \sup_{x'} \sqrt{\mathbb{E} \left[ \sigma^2(U^{(\ell)}(x')) \right]} = 2^{-\ell} \leq 1.$$

□

## 2.6 Bounding $\|H_t - H_0\|_\infty$

In this section, we prove that with high probability, for any sample path  $\{x_s, y_s\}_{s=0}^{T-1}$ ,  $\|H_t - H_0\|_\infty$  is small. As discussed in Section 2.4.2, this is crucial to the analysis of the average prediction error under SGD in the streaming data setup.

Recall from (2.1) that  $H_t = \sum_{\ell=1}^L H_t^{(\ell)}$  and

$$H_t^{(\ell)}(x, x') = \left\langle \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}, \frac{\partial f(x'; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} \right\rangle.$$

where

$$\frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} = \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(x) z_t^{(\ell)}(x) \left[ o_t^{(\ell-1)}(x) \right]^\top, \quad (2.39)$$

and  $z_t^{(\ell)}(x)$  measures the sensitivity of the output from the  $\ell$ -th hidden layer defined as

$$\left[ z_t^{(\ell)}(x) \right]^\top \triangleq \left[ \frac{\partial f(x; \mathbf{W}(t))}{\partial o^{(\ell)}(x)} \right]^\top = a^\top \frac{1}{\sqrt{m}} \mathbf{D}_t^{(L)}(x) \mathbf{W}^{(L)}(t) \cdots \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell+1)}(x) \mathbf{W}^{(\ell+1)}(t). \quad (2.40)$$

Throughout the section, we assume the width of each hidden layer  $m$  satisfies

$$m \geq d^9 \exp(\Omega(\theta L C^L \log T)) \quad (2.41)$$

for some absolute constant  $C$ . Also, recall from Section 2.3 that we assume

$\gamma \triangleq \max\{\|f^*\|_\infty, |u|\}$  is independent of  $m$  and choose step size  $\eta_t \leq \frac{\theta}{t+1}$ .

**Proposition 2.6.1.** *Assume (2.41) holds. With probability  $1 - \exp(-\Omega(C^{-L} m^{1/36}))$ , for any sample path  $\{x_s, y_s\}_{s=0}^{T-1}$ , all  $t \leq T$ , and all  $1 \leq \ell \leq L$ , we have*

$$\sup_x \left\| \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} - \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 = O\left(\frac{C^L}{m^{1/36}}\right),$$

and hence,

$$\left\| H_t^{(\ell)} - H_0^{(\ell)} \right\|_\infty = O\left(\frac{C^L}{m^{1/36}}\right). \quad (2.42)$$

To prove Proposition 2.6.1, in view of (2.39), the key is to control the deviations of  $\mathbf{D}_t^{(\ell)}(x)$ ,  $z_t^{(\ell)}(x)$  and  $o_t^{(\ell-1)}(x)$  uniformly, which will be done in the following Lemma 2.6.2–2.6.4. The detailed proof of Proposition 2.6.1 and Lemma 2.6.2–2.6.4 are deferred to Appendix 3.2.

We begin with bounding the deviation of  $o_t^{(\ell-1)}(x)$ . Define a sequence of real numbers:

$$\begin{aligned} R_0 &\triangleq m^{5/18}, \\ R_{t+1} &\triangleq R_0 + LC^{2L-2} \sum_{s=0}^t \eta_s (R_s + \gamma), t \geq 1. \end{aligned} \quad (2.43)$$

**Lemma 2.6.2.** *Assume (2.41) holds. Then we have  $R_t \leq m^{1/3}$  for all  $t \leq T$ . Moreover, with probability at least  $1 - \exp(-\Omega(C_1^{-L} m^{1/9}))$ , for any  $1 \leq \ell \leq L$ ,  $t \leq T$  and sample*

path  $\{x_s, y_s\}_{s=0}^{T-1}$ , the following holds:

$$\left\| \mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0) \right\|_2 \leq C_2^{L-1} \sum_{s=0}^{t-1} \eta_s (R_s + \gamma) \leq R_t \quad (2.44)$$

$$\sup_x \left\| o_t^{(\ell)}(x) - o_0^{(\ell)}(x) \right\|_2 \leq \frac{C_3^\ell}{m^{1/6}} \quad (2.45)$$

$$\sup_x |\Delta_t(x)| \leq R_t, \quad (2.46)$$

for some absolute constant  $C_1, C_2$  and  $C_3$ .

Lemma 2.6.2 is proved via induction over  $t$  in Section 3.2.2. A key underlying idea is as follows. While the weight vectors for some individual neurons may exhibit large deviations, collectively  $\mathbf{W}^{(\ell)}(t)$  is close to  $\mathbf{W}^{(\ell)}(0)$  in terms of the spectral norm, or equivalently the Frobenius norm as  $\mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0)$  is of rank no more than  $t$ , which is much smaller than  $m$  following (2.41). This allows us to further control the deviation of  $o_t^{(\ell)}(x)$  and  $|\Delta_t(x)|$  uniformly over all  $x$ , which in turn results in a small deviation of  $\mathbf{W}^{(\ell)}(t+1)$  in the next iteration. Departing from the previous work (e.g. [DLL<sup>+</sup>19, Lemma B.5]), here to control the deviation of  $\mathbf{W}^{(\ell)}(t+1)$ , it is crucial to bound  $|\Delta_t(x)|$  uniformly over all  $x$ . A critical intermediate step is to bound  $|f(x; \mathbf{W}(0))|$ . To this end, by observing that  $f^2(x; \mathbf{W}(0)) \leq a^\top \mathbf{Q}(x)a$  for some matrix  $\mathbf{Q}(x)$  independent of  $a$ , we apply the Hanson-Wright inequality for a fixed  $\mathbf{Q}(x)$  and then apply a union bound over  $\{\mathbf{Q}(x) : x \in \mathbb{S}^{d-1}\}$ , analogous to the proof of Lemma 2.5.2.

Next, we show  $\mathbf{D}_t^{(\ell)}(x)$  is close to  $\mathbf{D}_0^{(\ell)}(x)$  for any  $x$ . As such, define

$$S_t^{(\ell)}(x) \triangleq \|\mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\|_F.$$

Equivalently,  $S_t^{(\ell)}(x)$  measures the number of sign flips of the neurons at the  $\ell$ -th layer.

**Lemma 2.6.3.** *Assume (2.41) holds. Then with probability  $1 - \exp\left(-\Omega\left(C_1^{-L}m^{1/9}\right)\right)$  for any  $1 \leq \ell \leq L$ ,  $t \leq T$  and sample path  $\{x_s, y_s\}_{s=0}^{t-1}$ ,*

$$\sup_x S_t^{(\ell)}(x) \leq C_2^\ell m^{8/9}, \quad (2.47)$$

for some absolute constant  $C_1$  and  $C_2$ .

Note that the previous work [DZPS19] has obtained bounds to the number of sign flips in the batch setting with one hidden layer. They crucially require every individual weight vector  $w_i^{(1)}$  not to change much and hence only the neurons with small  $|\langle w_i^{(1)}(0), x \rangle|$  can have sign flips. However, in our setting, we need to further bound the number of neurons with relatively large deviations based on our bound of  $\|\mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0)\|_2$ .

Finally, we bound the deviation of the sensitivity  $z_t^{(\ell)}(x)$ .

**Lemma 2.6.4.** *Assume (2.41) holds. With probability at least  $1 - \exp(-\Omega(C_3^{-L+\ell}m^{1/36}))$ , for layer  $\ell$  and  $t \leq T$  and any sample path  $\{x_s, y_s\}_{s=0}^{T-1}$ , we have*

$$\sup_x \left\| z_0^{(\ell)}(x) \right\|_\infty \leq m^{1/36}, \quad (2.48)$$

and

$$\sup_x \left\| z_t^{(\ell)}(x) - z_0^{(\ell)}(x) \right\|_2 = O\left(C_4^{2L-\ell}m^{17/36}\right), \quad (2.49)$$

for some absolute constant  $C_3$  and  $C_4$ .

Lemma 2.6.4 is proved via a backward induction over  $\ell$  in Section 3.2.3. In particular, we crucially utilize the following layer-wise recursive relation

$$z_t^{(\ell)}(x) = \frac{1}{\sqrt{m}} \left[ \mathbf{W}^{(\ell+1)}(t) \right]^\top \mathbf{D}_t^{(\ell+1)}(x) z_t^{(\ell+1)}(x) \quad (2.50)$$

and apply the aforementioned deviation bounds of  $\mathbf{W}^{(\ell+1)}(t)$  and  $\mathbf{D}_t^{(\ell+1)}(x)$ . Note that even if there is only a single sign flip at some  $r$ -th neuron, an enormous value of the  $r$ -th coordinate of  $z_0^{(\ell+1)}(x)$  can possibly induce a large change of  $z_t^{(\ell)}(x)$ . To circumvent this issue, we derive a uniform bound to  $\|z_0^{(\ell)}(x)\|_\infty$ . Specifically, we observe that the  $r$ -th coordinate of  $z_0^{(\ell)}(x)$  equals  $\langle a, v_r^{(\ell)}(x) \rangle$  where  $v_r^{(\ell)}(x)$  is the  $r$ -th column of matrix  $\frac{1}{\sqrt{m}} \mathbf{D}^{(L)}(x) \mathbf{W}^{(L)}(0) \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell+1)}(x) \mathbf{W}^{(\ell+1)}(0)$  in view of (2.40). Analogous to the proof of Lemma 2.5.2, by conditioning on  $\{\mathbf{W}^{(k)}\}_{k=1}^L$ , we first show the concentration of  $\langle a, v_r^{(\ell)}(x) \rangle$  for a fixed  $v_r^{(\ell)}(x)$  and then apply a union bound by counting the number of  $v_r^{(\ell)}(x)$ .

## 2.7 Proof of Theorem 2

Recall from Section 2.4.2 that the recursion (2.24) plays a key role in showing Theorem 2. In the following lemma, we prove the recursion (2.24). Denote operators as

$$\mathbf{K}_t = \mathbf{I} - \eta_t \Phi, \quad \mathbf{Q}_t = \mathbf{I} - \eta_t \mathbf{H}_t, \quad \mathbf{D}_t = \mathbf{Q}_t - \mathbf{K}_t, \quad (2.51)$$

where  $\Phi$  is the integral operator associated with  $\Phi$ ,  $\mathbf{H}_t$  is the integral operator associated with  $H_t$  defined in (2.1), and  $\Phi \triangleq \sum_{\ell=1}^L \Phi^{(\ell)}$  with  $\Phi^{(\ell)}$  defined in (2.21).

**Lemma 2.7.1.** *For any  $t$ , we have*

$$\Delta_{t+1} = \mathbf{K}_t \circ \Delta_t + \mathbf{D}_t \circ \Delta_t + v_t + \epsilon_t, \quad (2.52)$$

and hence,

$$\begin{aligned} & \mathbb{E} [\|\Delta_{t+1}\|_2 | \mathbf{W}(0), a] \\ & \leq \left\| \prod_{s=0}^t \mathbf{K}_s \circ \Delta_0 \right\|_2 + \sum_{r=0}^t \mathbb{E} \left[ \left\| \prod_{i=r+1}^t \mathbf{Q}_i \mathbf{D}_r \prod_{j=0}^{r-1} \mathbf{K}_j \circ \Delta_0 \right\|_2 \middle| \mathbf{W}(0), a \right] \\ & \quad + \sum_{s=0}^t \mathbb{E} \left[ \left\| \prod_{r=s+1}^t \mathbf{Q}_r \circ \epsilon_s \right\|_2 \middle| \mathbf{W}(0), a \right] + \mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s \right\|_2 \middle| \mathbf{W}(0), a \right]. \end{aligned} \quad (2.53)$$

where  $\epsilon_t \equiv \epsilon_t(x, X_t; \mathbf{W}(t), \mathbf{W}(t+1))$  with

$$\epsilon_t(x, x'; \mathbf{W}(t), \mathbf{W}(t+1)) \triangleq f(x; \mathbf{W}(t)) - f(x; \mathbf{W}(t+1)) + \eta_t H_t(x, x') (f^*(x') + u_t - f(x'; \mathbf{W}(t)))$$

and

$$v_t \equiv v_t(x, X_t) = -\eta_t [(\Delta_t(X_t) + u_t) H_t(x, X_t) - \mathbb{E}_{X_t} [\Delta_t(X_t) H_t(x, X_t) | \mathbf{W}(0), a]]. \quad (2.54)$$

*Proof of Lemma 2.7.1.* By the definition of  $\epsilon_t$ , we have

$$\begin{aligned}\Delta_{t+1}(x) &= \Delta_t(x) - \eta_t H_t(x, X_t) (\Delta_t(X_t) + u_t) + \epsilon_t(x, X_t) \\ &= \Delta_t - \eta_t \mathbb{E}_{X_t} [H_t(x, X_t) \Delta_t(X_t) | \mathbf{W}(0), a] + \epsilon_t(x, X_t) + v_t(x, X_t).\end{aligned}$$

Using the notation in (2.51), we get the first equality of the lemma

$$\begin{aligned}\Delta_t &= \mathbf{Q}_t \circ \Delta_t + v_t + \epsilon_t \\ &= \mathbf{K}_t \circ \Delta_t + \mathbf{D}_t \circ \Delta_t + \epsilon_t + v_t,\end{aligned}$$

where the last equality holds since  $\mathbf{Q}_t = \mathbf{D}_t + \mathbf{K}_t$ .

Unrolling the above equality, we have

$$\Delta_{t+1} \leq \prod_{s=0}^t \mathbf{K}_s \circ \Delta_0 + \sum_{r=0}^t \prod_{i=r+1}^t \mathbf{Q}_i \mathbf{D}_r \prod_{j=0}^{r-1} \mathbf{K}_j \circ \Delta_0 + \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ \epsilon_s + \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s.$$

Taking  $L_2$  norm and conditional expectation, following the triangle inequality, we obtain the second inequality of the lemma.  $\square$

To bound the average prediction error  $\mathbb{E} [\|\Delta_{t+1}\|_2 | \mathbf{W}(0), a]$ , it suffices to bound the right hand side of (2.53). The first term can be bounded using the eigen-decomposition of  $\mathbf{K}_t$ . As an intermediate step, we prove both  $\Phi$  and  $\mathbf{H}_t$  are positive semi-definite with a bounded spectral norm in Lemma 2.7.2 below. This will be useful in bounding the spectrum of  $\mathbf{K}_t$  and  $\mathbf{Q}_t$ .

**Lemma 2.7.2.**  $\Phi$  is positive semi-definite with  $\|\Phi\|_2 \leq \|\Phi\|_\infty \leq \frac{L}{2}$ . Hence, for  $\eta_t \leq \frac{2}{L}$ , we have

$$0 \leq \lambda_i(\mathbf{K}_t) \leq 1,$$

for all  $i$  where  $\lambda_i(\mathbf{K}_t)$  is the  $i$ -th largest eigen-value of  $\mathbf{K}_t$ .

Assume (2.41) holds. With probability at least  $1 - \exp(-\Omega(C^{-L} m^{1/36}))$ ,  $\mathbf{H}_t$  are positive

semi-definite for all  $t \leq T$  with  $\|\mathbf{H}_t\|_2 \leq \frac{2L}{3}$ , and hence for  $\eta_t \leq \frac{3}{2L}$ ,

$$0 \leq \lambda_i(\mathbf{Q}_t) \leq 1,$$

where  $\lambda_i(\mathbf{Q}_t)$  is the  $i$ -th largest eigen-value of  $\mathbf{Q}_t$ .

The second term of (2.53) is the approximation error of using  $\Phi$  instead of  $\mathbf{H}_t$ . To bound the second term, we apply Lemma 2.7.2 which bounds  $\|\mathbf{Q}_t\|_2$  and  $\|\mathbf{K}_t\|_2$  for all  $t$ . Then we apply Proposition 2.6.1 as well as Theorem 1 to bound  $\|\mathbf{D}_t\|_2$ .

It remains to bound the last two terms. Intuitively, the third term is the accumulation of  $\epsilon_t$  and the last term is the accumulation of the noise from the stochastic gradients  $v_t$ .

The following lemma bounds the approximation error.

**Lemma 2.7.3.** *Assume (2.41) holds. With probability at least  $1 - \exp(-\Omega(C^{-L+1}m^{1/36}))$ , we have*

$$\mathbb{E} [\|\epsilon_t\|_2 | \mathbf{W}(0), a] = O\left(\frac{\eta_t C^L \sigma_t}{m^{1/36}}\right), \quad (2.55)$$

where

$$\sigma_t^2 = \mathbb{E} \left[ \|\Delta_t\|_2^2 | \mathbf{W}(0), a \right] + \tau^2. \quad (2.56)$$

and  $\tau$  is the variance of the noise  $u$ .

Next, we bound the noise from the stochastic gradients in expectation.

**Lemma 2.7.4.** *Assume (2.41) holds. With probability at least  $1 - \exp(-\Omega(C^{-L}m^{1/36}))$ , we have*

$$\mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s \right\|_2 \middle| \mathbf{W}(0), a \right] \leq c_2 \sigma_0, \quad (2.57)$$

where  $c_2 = L\theta e^{\sqrt{44}L\theta/9} \sqrt{\frac{1}{1-2\sqrt{44}L\theta/9} + 1}$ .

To prove Lemma 2.7.4, we first utilize  $\|\mathbf{Q}_t\|_2 \leq 1$  and the observation

$$\mathbb{E} \left[ v_s | \{X_r, y_r\}_{r=0}^{s-1}, \mathbf{W}(0), a \right] = 0$$

to show

$$\mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s \right\|_2^2 \middle| \mathbf{W}(0), a \right] \leq \sum_{s=0}^t \mathbb{E} \left[ \|v_s\|_2^2 \middle| \mathbf{W}(0), a \right].$$

Then following the definition of  $v_t$  and the upper bound of  $\|H_t\|_\infty$ , we have

$$\mathbb{E} \left[ \|v_s\|_2^2 \middle| \mathbf{W}(0), a \right] \leq \frac{4L\eta_s^2}{9} \sigma_s^2$$

where  $\sigma_s^2$  is defined in (2.56).

Finally we bound  $\sum_{s=0}^T \eta_s^2 \sigma_s^2$ . Note that  $\eta_s \leq \frac{\theta}{s+1}$ . Therefore, by showing  $\sigma_t$  does not grow too fast in  $t$ , i.e.,  $\sigma_{t+1} \leq \left(1 + \frac{\sqrt{44L\eta_t}}{9}\right) \sigma_t$ , we guarantee  $\sum_{s=0}^\infty \eta_s^2 \sigma_s^2$  converges and hence obtain the upper bound of  $\sum_{s=0}^T \eta_s^2 \sigma_s^2$ .

*Proof of Theorem 2.* Throughout the proof, we condition on  $\mathbf{W}(0)$  and  $a$  such that (2.22), (2.42), (2.55), (2.57) hold. This can be guaranteed with probability  $1 - \exp(-\Omega(C^{-L}m^{1/36}))$  following Theorem 1, Proposition 2.6.1, Lemma 2.7.3 and Lemma 2.7.4. For simplicity, we abbreviate the conditional expectation  $\mathbb{E}[\cdot | \mathbf{W}(0), a]$  as  $\mathbb{E}[\cdot]$ . We now prove the theorem by induction.

When  $t = 0$ , clearly  $\|\Delta_0\|_2 \leq \|\Delta_0\|_2 + 2c_2 \|\Delta_0\|_2 + 2c_2\tau$ .

Suppose (2.26) holds at all time  $s \leq t$ , now we show it also holds at time  $t+1$ . Following (2.53) in Lemma 2.7.1, we have

$$\begin{aligned} & \mathbb{E} [\|\Delta_{t+1}\|_2] \\ & \leq \left\| \prod_{s=0}^t \mathbf{K}_s \circ \Delta_0 \right\|_2 + \sum_{r=0}^t \mathbb{E} \left[ \left\| \prod_{i=r+1}^t \mathbf{Q}_i \mathbf{D}_r \prod_{j=0}^{r-1} \mathbf{K}_j \circ \Delta_0 \right\|_2 \right] \\ & \quad + \sum_{s=0}^t \mathbb{E} \left[ \left\| \prod_{r=s+1}^t \mathbf{Q}_r \circ \epsilon_s \right\|_2 \right] + \mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s \right\|_2 \right] \\ & \leq \left\| \prod_{s=0}^t \mathbf{K}_s \circ \Delta_0 \right\|_2 + \sum_{r=0}^t \mathbb{E} [\|\mathbf{D}_r\|_2] \|\Delta_0\|_2 + \sum_{s=0}^t \mathbb{E} [\|\epsilon_s\|_2] + \mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s \right\|_2 \right], \end{aligned} \tag{2.58}$$

where the last inequality holds by Lemma 2.7.2 that gives  $\|\mathbf{Q}_t\|_2 \leq 1$  and  $\|\mathbf{K}_t\|_2 \leq 1$  for all  $t \leq T$ .

Now we bound each term on the right hand side above.

Denote  $\rho_i(t) \triangleq \prod_{s=0}^t (1 - \eta_s \lambda_i)$  where  $\{\lambda_i\}_{i=1}^\infty$  are eigenvalues of  $\Phi$ . Since  $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$  and  $\sup_i (1 - \eta_s \lambda_i) \leq 1$  for all  $s$ , we know  $\rho_i(t)$  is bounded above by 1 and is increasing in  $i$ . To bound the first term on the right hand side of (2.58), here we use an induction to prove that

$$\prod_{s=0}^t \mathbf{K}_s \circ g = \sum_{i=1}^{\infty} \rho_i(t) \langle g, \phi_i \rangle \phi_i, \quad (2.59)$$

where  $\phi_i$  is the eigenfunction of  $\Phi$  associated with eigenvalue  $\lambda_i$ .

When  $t = 0$ , since  $\mathbf{K}_0$  is positive semi-definite, by Lemma A.3.3, we have

$$\mathbf{K}_0 \circ g = \sum_{i=1}^{\infty} \rho_i(0) \langle g, \phi_i \rangle \phi_i,$$

where  $\phi_i$  is the eigenfunction of  $\Phi$  associated with eigenvalue  $\lambda_i$ .

Suppose (2.59) holds for some time  $t$ . Since  $\mathbf{K}_{t+1}$  is PSD, by Lemma A.3.3, we have

$$\begin{aligned} \mathbf{K}_{t+1} \circ \left( \prod_{s=0}^t \mathbf{K}_s \circ g \right) &= \sum_{i=1}^{\infty} (1 - \eta_{t+1} \lambda_i) \left\langle \sum_{j=1}^{\infty} \rho_j(t) \langle g, \phi_j \rangle \phi_j, \phi_i \right\rangle \phi_i \\ &\stackrel{(a)}{=} \sum_{i=1}^{\infty} (1 - \eta_{t+1} \lambda_i) \langle \rho_i(t) \langle g, \phi_i \rangle \phi_i, \phi_i \rangle \phi_i \\ &\stackrel{(b)}{=} \sum_{i=1}^{\infty} \rho_i(t+1) \langle g, \phi_i \rangle \phi_i, \end{aligned}$$

where (a) holds by orthogonality of  $\{\phi_i\}$  and (b) holds by the definition of  $\rho_i$  and the normality of  $\phi_i$ . Therefore, taking  $L_2$  norm square on both hand sides of (2.59), for any

$r \in \mathbb{N}$ , we get

$$\begin{aligned}
\left\| \prod_{s=0}^t \mathsf{K}_s \circ \Delta_0 \right\|_2^2 &= \sum_{i=0}^{\infty} \rho_i^2(t) \langle \Delta_0, \phi_i \rangle^2 \\
&\stackrel{(a)}{\leq} \sum_{i=0}^r \rho_i^2(t) \langle \Delta_0, \phi_i \rangle^2 + \sum_{i=r+1}^{\infty} \langle \Delta_0, \phi_i \rangle^2 \\
&\stackrel{(b)}{\leq} \rho_r^2(t) \sum_{i=0}^r \langle \Delta_0, \phi_i \rangle^2 + \mathcal{R}^2(\Delta_0, r) \\
&\leq \rho_r^2(t) \|\Delta_0\|_2^2 + \mathcal{R}^2(\Delta_0, r),
\end{aligned}$$

where the residual projection error  $\mathcal{R}$  is defined in (2.25); (a) holds since  $\rho_i(t) \leq 1$  for all  $i$  and  $t$ ; (b) holds since  $\rho_i(t)$  is monotonic increasing in  $i$ . Hence, we have for any  $r \in \mathbb{N}$ ,

$$\left\| \prod_{s=0}^t \mathsf{K}_s \circ \Delta_0 \right\|_2 \leq \prod_{s=0}^t (1 - \eta_s \lambda_r) \|\Delta_0\|_2 + \mathcal{R}(\Delta_0, r). \quad (2.60)$$

To bound  $\sum_{r=0}^t \mathbb{E} [\|\mathsf{D}_r\|_2]$ , note that

$$\|\mathsf{D}_r\|_2 \leq \eta_r \|H_t - \Phi\|_{\infty} = O\left(\frac{\eta_r C^L}{m^{1/36}}\right).$$

Thus, we get

$$\sum_{r=0}^t \mathbb{E} [\|\mathsf{D}_r\|_2] = O\left(\frac{C^L \sum_{r=0}^t \eta_r}{m^{1/36}}\right) \leq C_1 \frac{\theta C^L \log T}{m^{1/36}}, \quad (2.61)$$

for some absolute constant  $C_1$  where the last inequality holds by plugging in  $\eta_r \leq \frac{\theta}{r+1}$ .

By (2.55), we have

$$\sum_{s=0}^t \mathbb{E} [\|\epsilon_s\|_2] = O\left(\frac{C_2^L}{m^{1/36}} \sum_{s=0}^t \eta_s \sigma_s\right).$$

By definition of  $\sigma_s$ , we have

$$\sigma_s = \sqrt{\mathbb{E} [\|\Delta_s\|_2^2] + \tau^2} \leq \mathbb{E} [\|\Delta_s\|_2] + \tau.$$

Now we prove that

$$\mathbb{E} [\|\Delta_s\|_2] \leq (1 + 2c_2) \|\Delta_0\|_2 + 2c_2\tau.$$

and hence

$$\sigma_s \leq (1 + 2c_2) \|\Delta_0\|_2 + (1 + 2c_2)\tau.$$

To see this, note that for any  $\epsilon > 0$ ,  $\mathcal{R}(\Delta_0, \ell) < \epsilon$  for sufficiently large  $\ell$ . Therefore, since (2.26) holds for all  $s \leq t$ , we have

$$\mathbb{E} [\|\Delta_s\|_2] \leq \prod_{r=0}^s (1 - \eta_r \lambda_r) \|\Delta_0\|_2 + \epsilon + 2c_2 \|\Delta_0\|_2 + 2c_2\tau \leq (1 + 2c_2) \|\Delta_0\|_2 + \epsilon + 2c_2\tau.$$

Since  $\epsilon$  can be arbitrary, we have  $\mathbb{E} [\|\Delta_s\|_2] \leq (1 + 2c_2) \|\Delta_0\|_2 + 2c_2\tau$ . Plugging the bound to  $\sigma_s$  and  $\eta_s \leq \frac{\theta}{s+1}$ , we get

$$\sum_{s=0}^t \mathbb{E} [\|\epsilon_s\|_2] \leq \frac{C_3 \theta C_2^L \log T}{m^{1/36}} (1 + 2c_2) (\|\Delta_0\|_2 + \tau). \quad (2.62)$$

for some absolute constant  $C_3$  where we use the fact that  $\sum_{t=0}^T \eta_t \leq \sum_{t=0}^T \frac{\theta}{t+1} \leq C' \theta \log T$ .

Lastly, from (2.57), we have

$$\mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t Q_r \circ v_s \right\|_2 \right] \leq c_2 (\|\Delta_0\|_2 + \tau).$$

Plugging the above bound as well as (2.60), (2.61) and (2.62) into (2.58), we have

$$\begin{aligned} & \mathbb{E} [\|\Delta_{t+1}\|_2] \\ & \leq \prod_{s=0}^t (1 - \eta_s \lambda_r) \|\Delta_0\|_2 + \mathcal{R}(\Delta_0, r) + C_1 \frac{\theta C^L \log T}{m^{1/36}} \|\Delta_0\|_2 \\ & \quad + \frac{C_3 \theta C_2^L \log T}{m^{1/36}} (1 + 2c_2) (\|\Delta_0\|_2 + \tau) + c_2 (\|\Delta_0\|_2 + \tau) \\ & = \left\{ \prod_{s=0}^t (1 - \eta_s \lambda_r) \right\} \|\Delta_0\|_2 + \mathcal{R}(\Delta_0, r) \\ & \quad + \left[ (C_1 + C_3(1 + 2c_2)) \frac{\theta C_4^L \log T}{m^{1/36}} + c_2 \right] \|\Delta_0\|_2 + \left( \frac{(1 + 2c_2) C_3 \theta C_2^L \log T}{m^{1/36}} + c_2 \right) \tau, \end{aligned} \quad (2.63)$$

where  $C_4 = \max \{C, C_2\}$ .

When  $m = \Omega \left( (C_1 + C_3(1 + c_2))^{36} \theta^{36} c_2^{36} C_4^{36L} \log^{36} T \right)$ , we have

$$(C_1 + C_3(1 + 2c_2)) \frac{\theta C_4^L \log T}{m^{1/36}} \leq c_2,$$

and

$$\frac{(1 + 2c_2) C_3 \theta C_2^L \log T}{m^{1/36}} \leq c_2.$$

As a result, we have

$$\mathbb{E} [\|\Delta_{t+1}\|_2] \leq \prod_{s=0}^t (1 - \eta_s \lambda_r) \|\Delta_0\|_2 + \mathcal{R}(\Delta_0, r) + 2c_2 \|\Delta_0\|_2 + 2c_2 \tau,$$

which completes the induction. □

## 2.8 Numerical Study

In this section, we provide some numerical studies.

### 2.8.1 Synthetic data

We consider the following different choices of  $f^*$ :

- Linear:  $f^*(x) = \langle b, x \rangle$  with parameter  $b \in \mathbb{R}^d$ ;
- Quadratic:  $f^*(x) = x^\top A x + \langle b, x \rangle$ , where  $A \in \mathbb{R}^{d \times d}$  and  $b \in \mathbb{R}^d$ ;
- Teacher neural network:  $f^*(x) = \sum_{i=1}^3 b_i \psi(\langle v_i, x \rangle)$ , where  $\psi(z) = \frac{1}{1+e^{-z}}$  is the sigmoid function,  $b_i \in \{-1, 1\}$ , and  $v_i \in \mathbb{R}^d$ ;
- Random label:  $f^*(x)$  is an i.i.d. Bernoulli random variable with parameter 1/2.

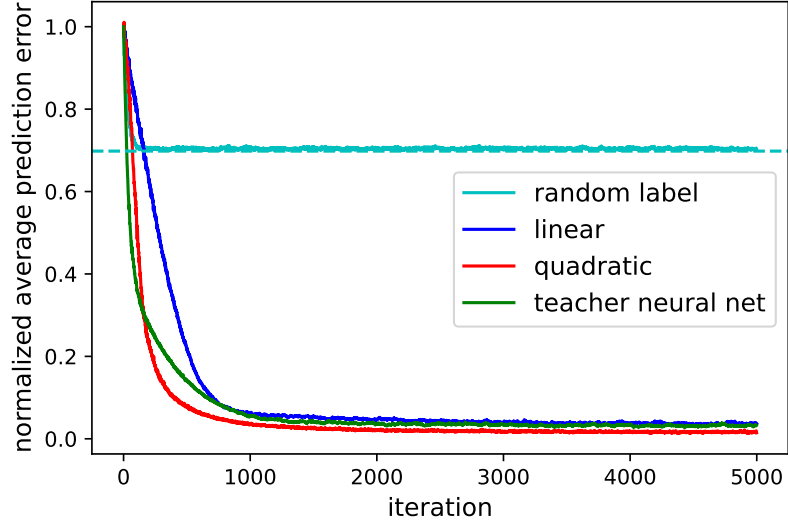
We use the symmetric initialization introduced in Section 2.4 as Corollary 1 suggests a zero residual projection error  $\mathcal{R}(\Delta_0, \ell^* + 1)$  for a degree  $\ell^*$  polynomial. We run the

stochastic gradient descent algorithm (2.10) on the streaming data with constant step size  $\eta = 0.3$  to train a four-layer neural network. At each iteration, we randomly draw data  $X$  uniformly from  $\mathbb{S}^{d-1}$  and  $u$  from  $\mathcal{N}(0, \tau^2)$  to obtain  $(X, y)$  where  $y = f^*(X) + u$ . The average prediction error is estimated using freshly drawn 200 data points, and the resulting error is further averaged over 20 independent runs.

In Section 2.4, we prove that the average prediction error converges under SGD in the streaming setting. Here we show the numerical performance of SGD. We study the normalized average prediction error  $\mathbb{E}[\|\Delta_t\|_2] / \mathbb{E}[\|\Delta_0\|_2]$  for different  $f^*$  with  $d = 5$ ,  $m = 1000$ , and  $\tau = 0.1$ . For linear, quadratic and teacher neural network  $f^*$ , the best achievable value of the normalized error equals 0. For random label  $f^*$ , since  $f^*(x)$  is an i.i.d. Bernoulli random variable with parameter 1/2 for any  $x$ , we get

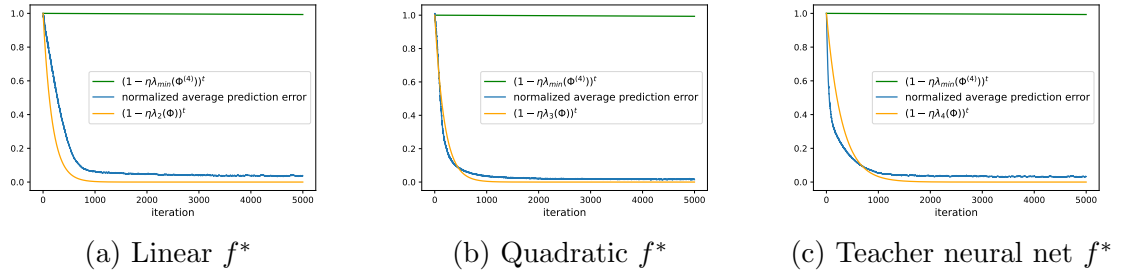
$$\|\Delta_t\|_2^2 = \mathbb{E}_X \left[ (f^*(X) - f(X; \mathbf{W}(t)))^2 \right] = \frac{1}{2} \left[ (f(X; \mathbf{W}(t)) - 1)^2 + f^2(X; \mathbf{W}(t)) \right] \geq \frac{1}{4}, \forall t.$$

Hence, the best achievable value of the normalized average prediction error equals  $\frac{1/2}{\mathbb{E}[\|\Delta_0\|_2]}$ , which is represented by the horizontal dashed line in Figure 2.3. From Figure 2.3, we clearly see that SGD learns  $f^*$  efficiently for all four choices: the normalized average prediction error converges to the best achievable values.



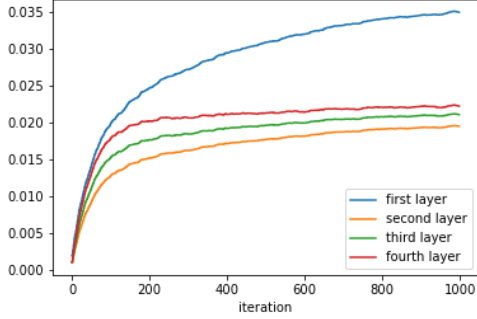
**Figure 2.3:** Normalized prediction error for different  $f^*$

As discussed in Section 2.4, our result which captures the contribution of NTK from all hidden layers, characterizes the average prediction error better than existing works du2019gradient . Here, we provide numerical studies to verify this statement. Figure 2.4 plots the evolution of the average prediction error normalized by the error at initialization and the characterizations utilizing the spectrum of  $\Phi$  and  $\Phi^{(4)}$ . It can be seen that our characterization based on  $\Phi$  is close to the actual SGD dynamic when  $f^*$  is linear, quadratic or teacher neural network function. Note that under the symmetric initialization,  $\Delta_0 = f^*$ . According to Corollary 1, we choose  $\lambda_2(\Phi)$  for linear and  $\lambda_3(\Phi)$  for quadratic  $f^*$  since the residual projection error equals 0. For teacher neural network  $f^*$  which is not polynomial, we cannot find some  $\ell^*$  such that the residual projection error  $\mathcal{R}(f^*, \ell^*) = 0$ . Instead, we choose  $\lambda_4(\Phi)$  as it provides the best fit among all  $\lambda_i(\Phi), i \geq 1$ .

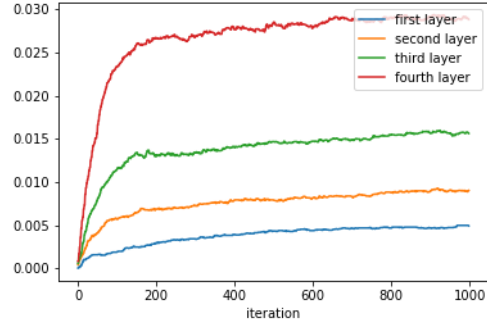


**Figure 2.4:** Average prediction error and the characterization based on  $\Phi$

Lastly, we provide some numerical study to verify the phenomenon that  $\mathbf{W}^{(\ell)}(t)$  stays relatively close to  $\mathbf{W}^{(\ell)}(0)$  in Lemma 2.6.2 and the number of sign changes for each hidden layer is relatively small in Lemma 2.6.3. Figure 2.5 studies teacher neural network  $f^*$ . Figure 2.5a shows that the relative deviation of weight matrix  $\|\mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0)\|_2 / \|\mathbf{W}^{(\ell)}(0)\|_2$  for each  $\ell$ -th hidden layer is small. This is consistent with the trend implied by Lemma 2.6.2 which shows that with high probability, the numerator  $\|\mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0)\|_2$  is small compared to the denominator  $\|\mathbf{W}^{(\ell)}(0)\|_2$ . In Figure 2.5b, we observe only a small fraction of sign changes in each hidden layer throughout the training. Furthermore, we see the proportion of sign changes increase when the layer index  $\ell$  increases. Both are consistent with the trend indicated by Lemma 2.6.3.



(a) Relative deviation of weight matrices

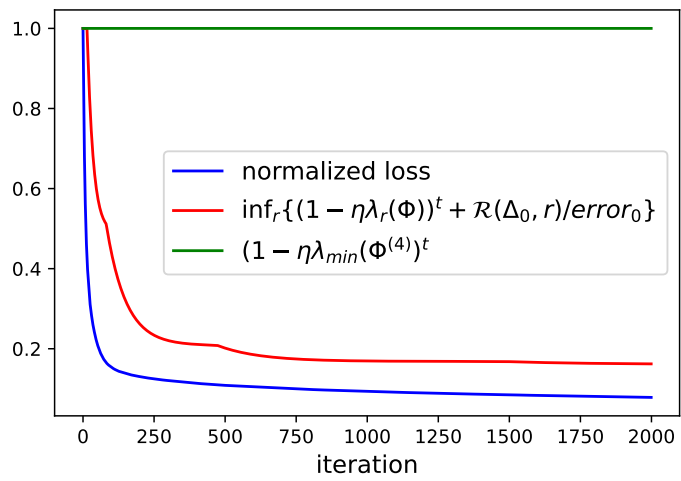


(b) Proportion of sign flips

**Figure 2.5:** Evolution of weight and sign flips for a 4-layer teacher neural network  $f^*$

## 2.8.2 Real data experiment

To illustrate the characterization from our theoretical result on real data, we run a numerical experiment on MNIST dataset. For simplicity, we only use the data corresponding to digit 0 and digit 1. We randomly draw 1500 images with  $28 \times 28$  pixels from each digit and treat the empirical distribution of these 3000 images as the underlying true data distribution. We reshape the data to have  $x_i \in \mathbb{R}^{784}$ . For each  $x_i \in \mathbb{R}^{784}$  in the dataset, we assign  $y_i = 1$  if the corresponding image is digit 1 and  $y_i = -1$  if the image is digit 0. We then normalize  $x_i$  to have  $\|x_i\|_2 = 1$ . We run the mini-batch SGD with mini-batch size 100 on streaming data with step size  $\eta = 0.7$  using a 4-layer neural network with 10000 neurons in each hidden layer. The reason for the use of mini-batch is to limit the noise from the stochastic gradient. Figure 2.6 shows the training loss normalized by the loss at initialization and two characterizations from the spectrum of NTK. We use  $\text{error}_0$  to denote the average prediction error at initialization. It can be seen that the characterization from our result provides a much tighter bound than the characterization from existing works du2019gradient using only the spectrum of the NTK from the last hidden layer. In addition, we clearly see two elbow points on our characterization. The first 100 iterations correspond to  $(1 - \eta\lambda_1(\Phi))^t + \mathcal{R}(\Delta_0, 1)/\text{error}_0$  while the next 400 iterations correspond to  $(1 - \eta\lambda_2(\Phi))^t + \mathcal{R}(\Delta_0, 2)/\text{error}_0$ .



**Figure 2.6:** Normalized training loss for the first 2000 iterations

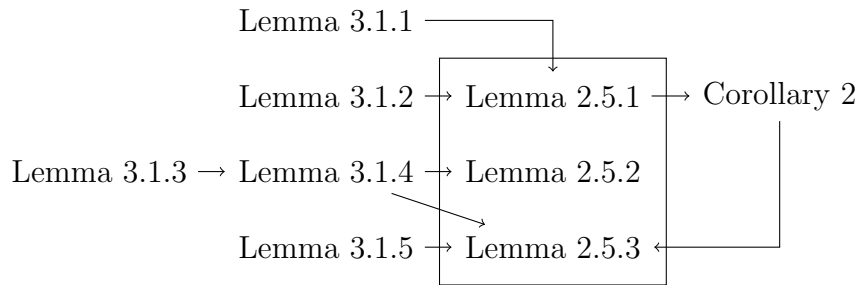
# Chapter 3

## Proofs of Chapter 2 Intermediate Results

In this chapter, we provide the detailed proof of technical lemmas in Chapter 2.

### 3.1 Proofs in Section 2.5

Recall from Section 2.5 that the proof of Theorem 1 consists of Lemma 2.5.1–2.5.3. Here, we present the full proofs of these lemmas. Since the proofs involve several key intermediate results, to ease the reading, we present the following diagram to illustrate the proof structure.



**Figure 3.1:** Diagram of the proof structure in Section 3.1

#### 3.1.1 Proof of Lemma 2.5.1

As mentioned in Section 2.5, to prove Lemma 2.5.1, we need to establish that  $o^{(\ell)}(x)$  is Lipschitz in  $x$ . This is done in the following lemma.

Define

$$\mathcal{E}_0^{(k)} = \left\{ \left\| \mathbf{W}^{(k)} \right\|_2 \leq c_0 \sqrt{m} \right\}. \quad (3.1)$$

where  $\mathbf{W}^{(k)}$  is the weight matrix of the  $k$ -th hidden layer.

**Lemma 3.1.1.** *For any  $0 \leq \ell \leq L$ , under event  $\cap_{k=1}^{\ell} \mathcal{E}_0^{(k)}$ ,*

- $\sup_x \|o^{(\ell)}(x)\|_2 \leq c_0^\ell$ ;
- $\|o^{(\ell)}(x) - o^{(\ell)}(z)\|_2 \leq c_0^\ell \|x - z\|_2$ .

*Proof of Lemma 3.1.1.* Recall that  $o^{(0)}(x) = x$  and

$$o^{(\ell)}(x) = \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x) \mathbf{W}^{(\ell)} \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(1)}(x) \mathbf{W}^{(1)} x, \quad \forall \ell \geq 1$$

where  $\mathbf{D}^{(\ell)}(x) = \text{diag} \left\{ \mathbf{1}_{\{\langle w_i^{(\ell)}, o^{(\ell-1)}(x) \rangle \geq 0\}} \right\}$  and  $w_i^{(\ell)}$  is the  $i$ -th row of  $\mathbf{W}^{(\ell)}$ .

When  $\ell = 0$ , since  $o^{(0)}(x) = x$ , both inequalities of Lemma 3.1.1 hold directly.

Now consider the case for  $\ell \geq 1$ . Under  $\cap_{k=1}^{\ell} \mathcal{E}_0^{(k)}$ , we know  $\|\mathbf{W}^{(k)}\|_2 \leq c_0 \sqrt{m}$  for all  $k = 1, 2, \dots, \ell$ . Therefore, for any  $x$ ,  $\left\| \frac{1}{\sqrt{m}} \mathbf{D}^{(k)}(x) \mathbf{W}^{(k)} \right\|_2 \leq c_0$  for all  $k = 1, 2, \dots, \ell$ . Thus, we have

$$\begin{aligned} \sup_x \|o^{(\ell)}(x)\|_2 &\leq \sup_x \left\| \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x) \mathbf{W}^{(\ell)} \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(1)}(x) \mathbf{W}^{(1)} \right\|_2 \\ &\leq \sup_x \left\| \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x) \mathbf{W}^{(\ell)} \right\|_2 \dots \sup_x \left\| \frac{1}{\sqrt{m}} \mathbf{D}^{(1)}(x) \mathbf{W}^{(1)} \right\|_2 \\ &\leq c_0^\ell. \end{aligned}$$

This completes the proof of the first inequality of Lemma 3.1.1.

Now we prove the second inequality of Lemma 3.1.1. By the definition of  $o^{(\ell)}(x)$ , we know

$$\begin{aligned} [o^{(\ell)}(x)]_i &= \left[ \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x) \mathbf{W}^{(\ell)} o^{(\ell-1)}(x) \right]_i = \frac{1}{\sqrt{m}} \mathbf{1}_{\{\langle w_i^{(\ell)}, o^{(\ell-1)}(x) \rangle \geq 0\}} \langle w_i^{(\ell)}, o^{(\ell-1)}(x) \rangle \\ &= \frac{1}{\sqrt{m}} \sigma \left( \langle w_i^{(\ell)}, o^{(\ell-1)}(x) \rangle \right), \end{aligned}$$

where  $[o^{(\ell)}(x)]_i$  is the  $i$ -th coordinate of  $o^{(\ell)}(x)$ .

As a result, for any  $x$  and  $z$ , we have

$$\begin{aligned}
\left\| o^{(\ell)}(x) - o^{(\ell)}(z) \right\|_2^2 &= \frac{1}{m} \sum_{i=1}^m \left( \sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x) \rangle) - \sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(z) \rangle) \right)^2 \\
&\stackrel{(i)}{\leq} \frac{1}{m} \sum_{i=1}^m \left( \langle w_i^{(\ell)}, o^{(\ell-1)}(x) \rangle - \langle w_i^{(\ell)}, o^{(\ell-1)}(z) \rangle \right)^2 \\
&= \frac{1}{m} \left\| \mathbf{W}^{(\ell)} \left( o^{(\ell-1)}(x) - o^{(\ell-1)}(z) \right) \right\|_2^2 \\
&\leq c_0^2 \left\| o^{(\ell-1)}(x) - o^{(\ell-1)}(z) \right\|_2^2.
\end{aligned}$$

where (i) holds since ReLU function is 1-Lipchitz and the last inequality holds under  $\cap_{k=1}^{\ell} \mathcal{E}_0^{(k)}$ .

Recursively applying the above displayed equation, we obtain the second inequality of Lemma 3.1.1.  $\square$

The following lemma from [DLL<sup>+</sup>19, Lemma G.4] shows that if the covariance matrices of two pairs of bivariate normal random variables are close entrywise, then the expectation of some function  $F$  on these two pairs are also close.

**Lemma 3.1.2.** *Let*

$$\mathbf{A} = \begin{pmatrix} a_1^2 & \rho_1 a_1 b_1 \\ \rho_1 a_1 b_1 & b_1^2 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} a_2^2 & \rho_2 a_2 b_2 \\ \rho_2 a_2 b_2 & b_2^2 \end{pmatrix}.$$

*Suppose there exists some constant  $C > 0$  such that*

$$\frac{1}{C} \leq \min(a_1, b_1, a_2, b_2) \leq \max(a_1, b_1, a_2, b_2) \leq C.$$

*Define  $F(\mathbf{X}) = \mathbb{E}_{(U,V) \sim \mathcal{N}(0,\mathbf{X})} [\sigma(U)\sigma(V)]$  for any positive definite matrix  $X$ . Then we have*

$$|F(\mathbf{A}) - F(\mathbf{B})| = O(\|\mathbf{A} - \mathbf{B}\|_{\infty}).$$

**Proof of Lemma 2.5.1** Denote  $V_0$  as a  $\frac{1}{m^2}$ -net of  $\mathbb{S}^{d-1}$ . By [Ver19, Corollary 4.2.13], we have  $V_0$  is of size  $O(m^{2d})$ . Define event  $\mathcal{E}_1^{(k)}$  such that the following holds for any

$x_0, x'_0 \in V_0$ :

$$\left| \frac{1}{m} \sum_{i=1}^m \sigma(\langle w_i^{(k)}, o^{(k-1)}(x_0) \rangle) \sigma(\langle w_i^{(k)}, o^{(k-1)}(x'_0) \rangle) - \mathbb{E}_w \left[ \sigma(\langle w, o^{(k-1)}(x_0) \rangle) \sigma(\langle w, o^{(k-1)}(x'_0) \rangle) \right] \right| \leq C \frac{c_0^{2(k-1)}}{m^{1/3}}. \quad (3.2)$$

Denote  $\mathcal{E}_1 = \cap_{k=1}^L \mathcal{E}_1^{(k)}$  and  $\mathcal{E}_0 = \cap_{k=1}^L \mathcal{E}_0^{(k)}$ .

To prove the lemma, we first bound  $\mathbb{P}[\mathcal{E}_0 \cap \mathcal{E}_1]$  and then show (2.30) holds under  $\mathcal{E}_0 \cap \mathcal{E}_1$ .

By Lemma A.4.1, we have

$$\mathbb{P}[\mathcal{E}_0 \cap \mathcal{E}_1] \geq 1 - \sum_{\ell=1}^L \mathbb{P} \left[ \left( \mathcal{E}_1^{(\ell)} \right)^c \mid \left( \cap_{k=1}^{\ell-1} \mathcal{E}_1^{(k)} \right) \cap \left( \cap_{k=1}^{\ell-1} \mathcal{E}_0^{(k)} \right) \right] - \sum_{\ell=1}^L \mathbb{P} \left[ \left( \mathcal{E}_0^{(\ell)} \right)^c \right]. \quad (3.3)$$

For  $1 \leq \ell \leq L$ , since  $\mathbf{W}^{(\ell)}$  has i.i.d. standard Gaussian entries, by [Ver19, Theorem 4.4.5],

$$\mathbb{P} \left[ \left( \mathcal{E}_0^{(\ell)} \right)^c \right] \leq \exp(-\Omega(m)). \quad (3.4)$$

Next we condition on  $\{\mathbf{W}^{(k)}\}_{k=1}^{\ell-1}$  such that  $\left( \cap_{k=1}^{\ell-1} \mathcal{E}_1^{(k)} \right) \cap \left( \cap_{k=1}^{\ell-1} \mathcal{E}_0^{(k)} \right)$  holds. Since  $w_i^{(\ell)}$ 's are independent of  $\{\mathbf{W}^{(k)}\}_{k=1}^{\ell-1}$ ,  $\langle w_i^{(\ell)}, o^{(\ell-1)}(x) \rangle \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \|o^{(\ell-1)}(x)\|_2^2)$ . Therefore,

$$\begin{aligned} & \|\sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x_0) \rangle) \sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x'_0) \rangle)\|_{\psi_1} \\ & \leq \|\sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x_0) \rangle)\|_{\psi_2} \|\sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x'_0) \rangle)\|_{\psi_2} \leq c_0^{2\ell-2}, \end{aligned}$$

where  $\|X\|_{\psi_2} \triangleq \inf \{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$ , and

$\|X\|_{\psi_1} \triangleq \inf \{t > 0 : \mathbb{E}[\exp(|X|/t)] \leq 2\}$  for any random variable  $X$ ; the first inequality holds by [Ver19, Lemma 2.7.7] and the second inequality holds by Lemma 3.1.1 under  $\cap_{k=1}^{\ell-1} \mathcal{E}_0^{(k)}$ .

It follows from the sub-exponential concentration inequality (Lemma A.1.2) that for any fixed  $(x_0, x'_0) \in V_0$ , (3.2) holds with probability at least  $1 - \exp(-\Omega(m^{1/3}))$ . Further

taking union bounds over  $V_0$ , we have

$$\mathbb{P} \left[ \mathcal{E}_1^{(\ell)} \mid \left( \bigcap_{k=1}^{\ell-1} \mathcal{E}_1^{(k)} \right) \cap \left( \bigcap_{k=1}^{\ell-1} \mathcal{E}_0^{(k)} \right) \right] \geq 1 - \exp \left( O(d \log m) - \Omega(m^{1/3}) \right). \quad (3.5)$$

Plugging (3.5) and (3.4) into (3.3), we have

$$\begin{aligned} \mathbb{P} [\mathcal{E}_0 \cap \mathcal{E}_1] &\geq 1 - L \exp \left( O(d \log m) - \Omega(m^{1/3}) \right) - L \exp (-\Omega(m)) \\ &\geq 1 - L \exp \left( O(d \log m) - \Omega(m^{1/3}) \right). \end{aligned} \quad (3.6)$$

It remains to show (2.30) under  $\mathcal{E}_0 \cap \mathcal{E}_1$ . Fix any  $(x, x')$  and denote  $(x_0, x'_0) \in V_0 \times V_0$  such that  $\|x - x_0\|_2 \leq \frac{1}{m^2}$  and  $\|x' - x'_0\|_2 \leq \frac{1}{m^2}$ . For any  $0 \leq \ell \leq L - 1$ , by the triangle inequality,

$$\begin{aligned} &\left| \langle o^{(\ell+1)}(x), o^{(\ell+1)}(x') \rangle - \mathbb{E} \left[ \sigma(U^{(\ell+1)}(x)) \sigma(U^{(\ell+1)}(x')) \right] \right| \\ &\leq \underbrace{\left| \langle o^{(\ell+1)}(x), o^{(\ell+1)}(x') \rangle - \langle o^{(\ell+1)}(x_0), o^{(\ell+1)}(x'_0) \rangle \right|}_{\text{(I)}} \\ &\quad + \underbrace{\left| \langle o^{(\ell+1)}(x_0), o^{(\ell+1)}(x'_0) \rangle - \mathbb{E}_w \left[ \sigma(\langle w, o^{(\ell)}(x_0) \rangle) \sigma(\langle w, o^{(\ell)}(x'_0) \rangle) \right] \right|}_{\text{(II)}} \\ &\quad + \underbrace{\left| \mathbb{E}_w \left[ \sigma(\langle w, o^{(\ell)}(x_0) \rangle) \sigma(\langle w, o^{(\ell)}(x'_0) \rangle) \right] - \mathbb{E} \left[ \sigma(U^{(\ell+1)}(x)) \sigma(U^{(\ell+1)}(x')) \right] \right|}_{\text{(III)}}, \end{aligned} \quad (3.7)$$

where

$$\begin{aligned} &(U^{(\ell+1)}(x), U^{(\ell+1)}(x')) \sim \mathcal{N} \left( 0, \Sigma^{(\ell)}(x, x') \right) \\ &\Sigma^{(\ell)}(x, x') \triangleq \begin{pmatrix} \mathbb{E} [\sigma^2(U^{(\ell)}(x))] & \mathbb{E} [\sigma(U^{(\ell)}(x)) \sigma(U^{(\ell)}(x'))] \\ \mathbb{E} [\sigma(U^{(\ell)}(x)) \sigma(U^{(\ell)}(x'))] & \mathbb{E} [\sigma^2(U^{(\ell)}(x'))] \end{pmatrix} \end{aligned} \quad (3.8)$$

$$\text{with } \Sigma^{(0)}(x, x') = \begin{pmatrix} 1 & \langle x, x' \rangle \\ \langle x, x' \rangle & 1 \end{pmatrix}.$$

To bound (I), note that for any  $y, z, y', z'$ , by the triangle inequality and Cauchy-

Schwartz inequality, we have

$$|\langle y, y' \rangle - \langle z, z' \rangle| \leq \|y - z\| \|y'\| + \|y' - z'\| \|z\|. \quad (3.9)$$

Thus, we get

$$\begin{aligned} \text{(I)} &\leq \sup_{x, x'} \left( \left\| o^{(\ell+1)}(x) - o^{(\ell+1)}(x_0) \right\|_2 \left\| o^{(\ell+1)}(x') \right\|_2 + \left\| o^{(\ell+1)}(x') - o^{(\ell+1)}(x'_0) \right\|_2 \left\| o^{(\ell+1)}(x_0) \right\|_2 \right) \\ &\leq \frac{2c_0^{2\ell+2}}{m^2}. \end{aligned} \quad (3.10)$$

where the last inequality holds under  $\mathcal{E}_0$  by Lemma 3.1.1.

For term (II), recall by (2.13) that

$$\langle o^{(\ell+1)}(x_0), o^{(\ell+1)}(x'_0) \rangle = \frac{1}{m} \sum_{i=1}^m \sigma(\langle w_i^{(\ell+1)}, o^{(\ell)}(x_0) \rangle) \sigma(\langle w_i^{(\ell+1)}, o^{(\ell)}(x'_0) \rangle).$$

Thus, under  $\mathcal{E}_1$ ,

$$\text{(II)} \leq C \frac{c_0^{2\ell}}{m^{1/3}}. \quad (3.11)$$

To bound (III), note that conditioning on  $o^{(\ell)}$ ,  $(\langle w, o^{(\ell)}(x_0) \rangle, \langle w, o^{(\ell)}(x'_0) \rangle)$  is a bivariate normal random vector with mean 0 and covariance

$$\mathbf{A}^{(\ell)}(x_0, x'_0) = \begin{pmatrix} \left\| o^{(\ell)}(x_0) \right\|_2^2 & \langle o^{(\ell)}(x_0), o^{(\ell)}(x'_0) \rangle \\ \langle o^{(\ell)}(x_0), o^{(\ell)}(x'_0) \rangle & \left\| o^{(\ell)}(x'_0) \right\|_2^2 \end{pmatrix}.$$

Thus, by Lemma 3.1.2, we have

$$\begin{aligned} \text{(III)} &= O \left( \left\| \mathbf{A}^{(\ell)}(x_0, x'_0) - \Sigma^{(\ell)}(x, x') \right\|_{\infty} \right) \\ &\leq O \left( \left\| \mathbf{A}^{(\ell)}(x_0, x'_0) - \mathbf{A}^{(\ell)}(x, x') \right\|_{\infty} \right) + O \left( \left\| \mathbf{A}^{(\ell)}(x, x') - \Sigma^{(\ell)}(x, x') \right\|_{\infty} \right) \\ &= O \left( \frac{c_0^{2\ell}}{m^2} \right) + O \left( \left\| \mathbf{A}^{(\ell)}(x, x') - \Sigma^{(\ell)}(x, x') \right\|_{\infty} \right) \end{aligned} \quad (3.12)$$

where the last equality holds by (3.9).

Plugging (3.10), (3.11) and (3.12) into (3.7) and taking supremum over  $(x, x')$ , we get

$$\begin{aligned} & \sup_{x, x'} \left| \langle o^{(\ell+1)}(x), o^{(\ell+1)}(x') \rangle - \mathbb{E} \left[ \sigma(U^{(\ell+1)}(x)) \sigma(U^{(\ell+1)}(x')) \right] \right| \\ & \leq O \left( \frac{c_0^{2\ell}}{m^{1/3}} \right) + O \left( \sup_{x, x'} \left\| \mathbf{A}^{(\ell)}(x, x') - \Sigma^{(\ell)}(x, x') \right\|_{\infty} \right). \end{aligned} \quad (3.13)$$

By definition of  $\mathbf{A}^{(\ell)}$  and  $\Sigma^{(\ell)}$ , for  $\ell \geq 1$ ,

$$\begin{aligned} & \sup_{x, x'} \left\| \mathbf{A}^{(\ell)}(x, x') - \Sigma^{(\ell)}(x, x') \right\|_{\infty} \\ & \leq O \left( \sup_{x, x'} \left| \langle o^{(\ell)}(x), o^{(\ell)}(x') \rangle - \mathbb{E} \left[ \sigma(U^{(\ell)}(x)) \sigma(U^{(\ell)}(x')) \right] \right| \right). \end{aligned} \quad (3.14)$$

Recursively applying (3.13) and (3.14), and noting that

$$\sup_{x, x'} \left\| \mathbf{A}^{(0)}(x, x') - \Sigma^{(0)}(x, x') \right\|_{\infty} = 0,$$

we complete the proof of (2.30).

### 3.1.2 Proof of Lemma 2.5.2

Recall from the discussion in Section 2.5 that a crucial step in the proof of Lemma 2.5.2 is to bound the number of different matrices  $\mathbf{G}_k^{(\ell)}(x, x')$  by varying  $x, x'$  through bounding the cardinality of  $\mathcal{D}_k$  where  $\mathbf{G}_k^{(\ell)}(x, x') = \left[ \mathbf{V}_k^{(\ell)}(x) \right]^{\top} \mathbf{V}_k^{(\ell)}(x')$  from (2.12),

$$\left[ \mathbf{V}_k^{(\ell)}(x) \right]^{\top} \triangleq \frac{1}{\sqrt{m}} \mathbf{D}^{(k)}(x) \mathbf{W}^{(k)} \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell+1)}(x) \mathbf{W}^{(\ell+1)} \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x)$$

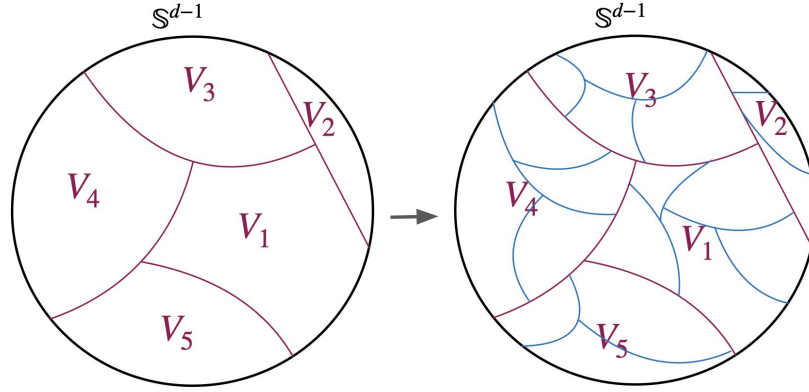
from (2.8) and  $\mathcal{D}_k = \{(\mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(k)}(x)) : x \in \mathbb{S}^{d-1}\}$ .

**Lemma 3.1.3.** *Fix any  $k > 0$  and  $\ell \leq k$ . For any fixed  $\{\mathbf{W}^{(r)}\}_{r=1}^k$ , we have  $|\mathcal{D}_k| \leq m^{dk}$ , and hence  $|\mathcal{G}_k^{(\ell)}| \leq m^{2dk}$  where  $\mathcal{G}_k^{(\ell)}(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k)}) \triangleq \left\{ \mathbf{G}_k^{(\ell)}(x, x') : x, x' \in \mathbb{S}^{d-1} \right\}$ .*

Intuitively, Lemma 3.1.3 implies that while there are infinite many different choices of  $x, x'$  on the unit sphere  $\mathbb{S}^{d-1}$ , the number of different matrices  $\mathbf{G}_k^{(\ell)}(x, x')$  is finite for any

fixed  $\{\mathbf{W}^{(r)}\}_{r=1}^k$ .

Before presenting the proof of Lemma 3.1.3, we provide a proof sketch for the ease of reading. To prove Lemma 3.1.3, note that given fixed  $\{\mathbf{W}^{(r)}\}_{r=1}^k$ , by the definition of  $\mathcal{G}_k^{(\ell)}$ , we have  $|\mathcal{G}_k^{(\ell)}| \leq |\mathcal{D}_k|^2$ . The proof is then completed by showing  $|\mathcal{D}_k| \leq m^{dk}$ . To obtain this, we show  $|\mathcal{D}_1| \leq m^d$  and  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$  for all  $k$ . The key of proving  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$  lies on a refinement idea illustrated in Figure 3.2. In particular, we partition  $\mathbb{S}^{d-1}$  into disjoint  $V_j$  for  $j = 1, 2, \dots, |\mathcal{D}_{k-1}|$  such that  $\cup_j V_j = \mathbb{S}^{d-1}$  and  $V_j \cap V_{j'} = \emptyset$  for all  $j \neq j'$ , and that for any  $x$  within the same  $V_j$ ,  $(\mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(k-1)}(x))$  is the same. We then refine  $V_j$  so that within each subregion after refinement,  $\mathbf{D}^{(k)}(x)$  is the same. Here, we crucially show  $|\{\mathbf{D}^{(k)}(x) : x \in V_j\}| \leq m^d$  for all  $j$ , i.e., the refinement within each  $V_j$  cannot exceed  $m^d$  and hence conclude  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$ .



**Figure 3.2:** Illustration of the key idea in showing  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$ . Left hand side shows the partition of  $\mathbb{S}^{d-1}$  into disjoint  $\{V_j\}_{j=1}^{|\mathcal{D}_{k-1}|}$  so that for any  $x$  within  $V_j$ ,  $(\mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(k-1)}(x))$  is the same. We then refine each  $V_j$  to obtain the right hand side so that within each refined sub-region in  $V_j$ ,  $\mathbf{D}^{(k)}(x)$  is also the same. We then show the number of such sub-region cannot exceed  $m^d$  for any  $V_j$ , which leads to  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$ .

*Proof of Lemma 3.1.3.* Throughout the proof, we fix  $\{\mathbf{W}^{(r)}\}_{r=1}^k$ . Since  $\{\mathbf{W}^{(r)}\}_{r=1}^k$  is fixed,

we have

$$\mathcal{G}_k^{(\ell)} \subset \left\{ \mathbf{V}^\top \tilde{\mathbf{V}} : \mathbf{V} = \mathbf{D}^{(k)} \mathbf{W}^{(k)} \dots \mathbf{D}^{(\ell+1)} \mathbf{W}^{(\ell+1)} \mathbf{D}^{(\ell)}, \tilde{\mathbf{V}} = \tilde{\mathbf{D}}^{(k)} \mathbf{W}^{(k)} \dots \tilde{\mathbf{D}}^{(\ell+1)} \mathbf{W}^{(\ell+1)} \tilde{\mathbf{D}}^{(\ell)}, \right. \\ \left. \left( \mathbf{D}^{(1)}, \dots, \mathbf{D}^{(k)} \right) \in \mathcal{D}_k, \left( \tilde{\mathbf{D}}^{(1)}, \dots, \tilde{\mathbf{D}}^{(k)} \right) \in \mathcal{D}_k \right\}.$$

Thus  $|\mathcal{G}_k^{(\ell)}| \leq |\mathcal{D}_k|^2$ , and the proof is completed by the following claim:

$$|\mathcal{D}_k| \leq m^{dk}. \quad (3.15)$$

To prove this claim, we first show  $|\mathcal{D}_1| \leq m^d$  and then show the recursion  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$  for all  $k \geq 2$ .

**Step 1 bounding  $|\mathcal{D}_1|$ :** Note that  $\mathbf{D}^{(1)}(x)$  is diagonal whose  $i$ -th diagonal element equals  $f_x(w_i^{(1)})$ , where  $f_x(w) = \mathbf{1}_{\{(w,x) \geq 0\}}$ . Thus, letting

$$\mathcal{F}^{(1)} = \left\{ f_x(w) = \mathbf{1}_{\{(w,x) \geq 0\}} : x \in \mathbb{S}^{d-1} \right\},$$

we have

$$\left| \left\{ \mathbf{D}^{(1)}(x) : x \in \mathbb{S}^{d-1} \right\} \right| = \left| \left\{ \left( f(w_1^{(1)}), \dots, f(w_m^{(1)}) \right) : f \in \mathcal{F}^{(1)} \right\} \right|.$$

It follows from Lemma A.2.3 that  $\left| \left\{ \mathbf{D}^{(1)}(x) : x \in \mathbb{S}^{d-1} \right\} \right| \leq m^{\text{VC}(\mathcal{F}^{(1)})}$ . By [HR19, Proposition 7.1],

$$\text{VC}(\mathcal{F}^{(1)}) = d. \quad (3.16)$$

As a result, we get  $|\mathcal{D}_1| = \left| \left\{ \mathbf{D}^{(1)}(x) : x \in \mathbb{S}^{d-1} \right\} \right| \leq m^d$ .

**Step 2 showing  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$  for any  $k \geq 2$ :** Partition  $\mathbb{S}^{d-1}$  into disjoint  $V_j$  for  $j = 1, 2, \dots, |\mathcal{D}_{k-1}|$  such that for any  $x$  and  $x'$  within the same  $V_j$ ,

$$\left( \mathbf{D}^{(1)}(x), \mathbf{D}^{(2)}(x), \dots, \mathbf{D}^{(k-1)}(x) \right) = \left( \mathbf{D}^{(1)}(x'), \mathbf{D}^{(2)}(x'), \dots, \mathbf{D}^{(k-1)}(x') \right).$$

Note that  $\mathcal{D}_k = \cup_{j=1}^{|\mathcal{D}_{k-1}|} \{(\mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(k)}(x)) : x \in V_j\}$ . Thus,

$$|\mathcal{D}_k| \leq \sum_{j=1}^{|\mathcal{D}_{k-1}|} \left| \left\{ \left( \mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(k)}(x) \right) : x \in V_j \right\} \right| = \sum_{j=1}^{|\mathcal{D}_{k-1}|} \left| \left\{ \mathbf{D}^{(k)}(x) : x \in V_j \right\} \right|, \quad (3.17)$$

where the last equality holds because  $(\mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(k-1)}(x))$  is the same for all  $x \in V_j$ .

It remains to bound  $|\{\mathbf{D}^{(k)}(x) : x \in V_j\}|$ . The  $i$ -th diagonal element of  $\mathbf{D}^{(k)}(x)$  equals  $f_x(w_i^{(k)})$ , where  $f_x(w) = \mathbf{1}_{\langle w, o^{(k-1)}(x) \rangle \geq 0}$ . Therefore, by letting

$$\mathcal{F}_j^{(k)}(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k-1)}) \triangleq \left\{ f_x(w) = \mathbf{1}_{\langle w, o^{(k-1)}(x) \rangle \geq 0} : x \in V_j \right\}, \quad (3.18)$$

we have

$$\left| \left\{ \mathbf{D}^{(k)}(x) : x \in V_j \right\} \right| = \left| \left\{ \left( f(w_1^{(k)}), \dots, f(w_m^{(k)}) \right) : f \in \mathcal{F}_j^{(k)} \right\} \right| \leq m^{\text{VC}(\mathcal{F}_j^{(k)})}, \quad (3.19)$$

where the last inequality holds by Lemma A.2.3.

Now we bound  $\text{VC}(\mathcal{F}_j^{(k)})$ . Since  $(\mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(k-1)}(x))$  is the same across all  $x \in V_j$  and  $\{\mathbf{W}^{(r)}\}_{r=1}^{k-1}$  are fixed, by definition of  $o^{(k-1)}$ , we have  $o^{(k-1)}(x) = P_j x$ , for all  $x \in V_j$ , where  $P_j = \frac{1}{\sqrt{m}} \mathbf{D}^{(k-1)}(x) \mathbf{W}^{(k-1)} \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(1)}(x) \mathbf{W}^{(1)} \in \mathbb{R}^{m \times d}$  is some matrix independent of  $x$ . Therefore,  $o^{(k-1)}(x)$  lies on the same  $d$ -dimensional subspace of  $\mathbb{R}^m$  for all  $x \in V_j$ . By [HR19, Proposition 7.1],

$$\text{VC}(\mathcal{F}_j^{(k)}) = d. \quad (3.20)$$

It then follows from (3.19) that  $|\{\mathbf{D}^{(k)}(x) : x \in V_j\}| \leq m^d$  for all  $j = 1, 2, \dots, |\mathcal{D}_{k-1}|$ . Further plugging this bound into (3.17) yields that  $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$ .  $\square$

With Lemma 3.1.3, we prove the following key intermediate result by applying Hanson-Wright inequality with a union bound on  $\mathbf{G}_k^{(\ell)}(x, x')$ .

**Lemma 3.1.4.** *Let  $Y = (Y_1, Y_2, \dots, Y_m) \in \mathbb{R}^m$  be a random vector with mean zero, independent, sub-Gaussian coordinates with  $\|Y_i\|_{\psi_2} \leq C$ . Assume  $Y$  is independent of  $\{\mathbf{W}^{(r)}\}_{r=1}^k$ .*

For any  $\ell = 1, 2, \dots, k$ , we have,

$$\begin{aligned} & \mathbb{P} \left[ \sup_{\ell \in [k]} \sup_{x, x'} \left| Y^\top \mathbf{G}_k^{(\ell)}(x, x') Y - \text{Tr} \left( \mathbf{G}_k^{(\ell)}(x, x') \right) \right| \geq \frac{c_0^{2k-2}}{m^{1/3}} \left| \cap_{r=0}^k \mathcal{E}_0^{(r)} \right| \right] \\ & \leq k \exp \left( O(dk \log m) - \Omega(m^{1/3}) \right). \end{aligned} \quad (3.21)$$

In the above lemma, by taking  $Y = a$  and  $k = L$ , we obtain the uniform concentration of  $a^\top \mathbf{G}_L^{(\ell)}(x, x') a$  on  $\text{Tr} \left( \mathbf{G}_L^{(\ell)}(x, x') \right)$  conditional on  $\cap_{r=1}^L \mathcal{E}_0^{(r)}$ , which readily implies Lemma 2.5.2. Furthermore, by taking  $Y = w_i^{(k+1)}$ , we obtain the uniform concentration of  $\left[ w_i^{(k+1)} \right]^\top \mathbf{G}_k^{(\ell)}(x, x') w_i^{(k+1)}$  on  $\text{Tr} \left( \mathbf{G}_k^{(\ell)}(x, x') \right)$  for any  $i \in [m]$ , where  $w_i^{(k+1)}$  is the  $i$ -th row of  $\mathbf{W}^{(k+1)}$ . That turns out to be instrumental in the proof of Lemma 2.5.3.

*Proof of Lemma 3.1.4.* Fix arbitrary  $\ell \leq k$ . We condition on  $\{\mathbf{W}^{(r)}\}_{r=1}^k$  such that  $\cap_{r=1}^k \mathcal{E}_0^{(r)}$  holds. Under  $\cap_{r=1}^k \mathcal{E}_0^{(r)}$ , for any  $x \in \mathbb{S}^{d-1}$ , we have

$$\begin{aligned} \left\| \mathbf{V}_k^{(\ell)}(x) \right\|_2 &= \left\| \left[ \frac{1}{\sqrt{m}} \mathbf{D}^{(k)}(x) \mathbf{W}^{(k)} \dots \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell+1)}(x) \mathbf{W}^{(\ell+1)} \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x) \right]^\top \right\|_2 \\ &\leq \frac{c_0^{k-\ell}}{\sqrt{m}}. \end{aligned}$$

By definition of  $\mathbf{G}_k^{(\ell)}$ , we get

$$\left\| \mathbf{G}_k^{(\ell)}(x, x') \right\|_2 = \left\| \left[ \mathbf{V}_k^{(\ell)}(x) \right]^\top \mathbf{V}_k^{(\ell)}(x') \right\|_2 \leq \left\| \mathbf{V}_k^{(\ell)}(x) \right\|_2 \left\| \mathbf{V}_k^{(\ell)}(x') \right\|_2 \leq \frac{c_0^{2k-2\ell}}{m}. \quad (3.22)$$

and

$$\left\| \mathbf{G}_k^{(\ell)}(x, x') \right\|_F \leq \sqrt{m} \left\| \mathbf{G}_k^{(\ell)}(x, x') \right\|_2 \leq \frac{c_0^{2k-2\ell}}{\sqrt{m}}. \quad (3.23)$$

Since  $Y$  is mean zero and is independent of  $\{\mathbf{W}^{(r)}\}_{r=1}^k$ , we have

$$\mathbb{E} \left[ Y^\top \mathbf{G}_k^{(\ell)}(x, x') Y \left| \left\{ \mathbf{W}^{(r)} \right\}_{r=1}^k \right. \right] = \text{Tr} \left( \mathbf{G}_k^{(\ell)}(x, x') \right).$$

Thus, under event  $\cap_{r=1}^k \mathcal{E}_0^{(r)}$ , by Hanson-Wright inequality, we have for any fixed  $x, x'$ ,

$$\begin{aligned} & \mathbb{P} \left[ \left| Y^\top \mathbf{G}_k^{(\ell)}(x, x') Y - \text{Tr} \left( \mathbf{G}_k^{(\ell)}(x, x') \right) \right| > \frac{c_0^{2k-2\ell}}{m^{1/3}} \left| \left\{ \mathbf{W}^{(r)} \right\}_{r=1}^k \right. \right] \\ & \leq 2 \exp \left( -C \min \left( \frac{c_0^{4k-4\ell} m^{-2/3}}{\|\mathbf{G}_k^{(\ell)}(x, x')\|_{\mathbb{F}}^2}, \frac{c_0^{2k-2\ell} m^{-1/3}}{\|\mathbf{G}_k^{(\ell)}(x, x')\|_2} \right) \right) = \exp \left( -\Omega(m^{1/3}) \right). \end{aligned}$$

By Lemma 3.1.3, we have  $|\mathcal{G}_k^{(\ell)}(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(k)})| \leq m^{2dk}$ . Taking union bounds over all possible  $\mathbf{G}_k^{(\ell)}$ , we have under event  $\cap_{r=1}^k \mathcal{E}_0^{(r)}$ ,

$$\begin{aligned} & \mathbb{P} \left[ \sup_{x, x'} \left| Y^\top \mathbf{G}_k^{(\ell)}(x, x') Y - \text{Tr} \left( \mathbf{G}_k^{(\ell)}(x, x') \right) \right| > \frac{c_0^{2k-2\ell}}{m^{1/3}} \left| \left\{ \mathbf{W}^{(r)} \right\}_{r=1}^k \right. \right] \\ & = \mathbb{P} \left[ \sup_{\mathbf{G} \in \mathcal{G}_k^{(\ell)}} \left| Y^\top \mathbf{G} Y - \text{Tr}(\mathbf{G}) \right| > \frac{c_0^{2k-2\ell}}{m^{1/3}} \left| \left\{ \mathbf{W}^{(r)} \right\}_{r=1}^k \right. \right] \\ & = m^{2dk} \exp \left( -\Omega(m^{1/3}) \right) = \exp \left( O(dk \log m) - \Omega(m^{1/3}) \right). \end{aligned}$$

Further take union bounds over  $\ell$ , we obtain that

$$\begin{aligned} & \mathbb{P} \left[ \sup_{\ell \in [k]} \sup_{x, x'} \left| Y^\top \mathbf{G}_k^{(\ell)}(x, x') Y - \text{Tr} \left( \mathbf{G}_k^{(\ell)}(x, x') \right) \right| > \frac{c_0^{2k-2\ell}}{m^{1/3}} \left| \left\{ \mathbf{W}^{(r)} \right\}_{r=1}^k \right. \right] \\ & = k \exp \left( O(dk \log m) - \Omega(m^{1/3}) \right). \end{aligned}$$

Taking the average of  $\left\{ \mathbf{W}^{(r)} \right\}_{r=1}^k$  on the event  $\cap_{r=1}^k \mathcal{E}_0^{(r)}$ , we get the desired conclusion.  $\square$

**Proof of Lemma 2.5.2:** Denote

$$\mathcal{E}_2 = \left\{ \sup_{\ell \in [L]} \sup_{x, x'} \left| a^\top \mathbf{G}_L^{(\ell)}(x, x') a - \text{Tr} \left( \mathbf{G}_L^{(\ell)}(x, x') \right) \right| \leq C \frac{c_0^{2L-2}}{m^{1/3}} \right\}.$$

Note that  $a$  is mean zero, sub-Gaussian and is independent of  $\left\{ \mathbf{W}^{(k)} \right\}_{k=1}^L$ .

Thus, by Lemma 3.1.4, we have

$$\mathbb{P} \left[ \mathcal{E}_2 | \cap_{k=1}^L \mathcal{E}_0^{(k)} \right] \geq 1 - L \exp \left( O(dL \log m) - \Omega(m^{1/3}) \right)$$

From (3.4), we have

$$\mathbb{P} \left[ \cap_{k=1}^L \mathcal{E}_0^{(k)} \right] \geq 1 - L \exp(-\Omega(m)). \quad (3.24)$$

Therefore,

$$\begin{aligned} \mathbb{P} [\mathcal{E}_2] &\geq \mathbb{P} \left[ \mathcal{E}_2 | \cap_{k=1}^L \mathcal{E}_0^{(k)} \right] \mathbb{P} \left[ \cap_{k=1}^L \mathcal{E}_0^{(k)} \right] \\ &\geq \left( 1 - L \exp \left( O(dL \log m) - \Omega(m^{1/3}) \right) \right) (1 - L \exp(-\Omega(m))) \\ &\geq 1 - \exp \left( O(dL \log m) - \Omega(m^{1/3}) \right). \end{aligned}$$

### 3.1.3 Proof of Lemma 2.5.3

Recall from (2.19) that

$$\begin{aligned} q_L^{(\ell)}(x, x') &= \frac{\pi - \arccos \rho^{(L-1)}(x, x')}{2\pi} q_{L-1}^{(\ell)}(x, x'), \quad \forall \ell \leq L, \\ q_L^{(L)}(x, x') &= \frac{\pi - \arccos \rho^{(L-1)}(x, x')}{2\pi}, \end{aligned}$$

where

$$\rho^{(L-1)}(x, x') = \frac{\mathbb{E} [\sigma(U^{(L-1)}(x)) \sigma(U^{(L-1)}(x'))]}{\sqrt{\mathbb{E} [\sigma^2(U^{(L-1)}(x))]} \sqrt{\mathbb{E} [\sigma^2(U^{(L-1)}(x'))]}}, \quad (3.25)$$

and  $U^{(L-1)}$  is defined in (3.8).

To prove Lemma 2.5.3, we crucially show the concentration of  $\text{Tr} \left( \mathbf{G}_L^{(\ell)}(x, x') \right)$  on  $q_{L-1}^{(L-1)}(x, x') \text{Tr} \left( \mathbf{G}_{L-1}^{(\ell)}(x, x') \right)$ . Then, by repeatedly applying this recursive relation of  $\text{Tr} \left( \mathbf{G}_{L-1}^{(\ell)}(x, x') \right)$ , we obtain the concentration of  $\text{Tr} \left( \mathbf{G}_L^{(\ell)}(x, x') \right)$  on  $q_L^{(\ell)}(x, x')$ .

Proving the concentration of  $\text{Tr} \left( \mathbf{G}_L^{(\ell)}(x, x') \right)$  on  $q_{L-1}^{(L-1)}(x, x') \text{Tr} \left( \mathbf{G}_{L-1}^{(\ell)}(x, x') \right)$  consists of the following three steps.

**Step 1:** As the first step, we show the concentration of  $\text{Tr} \left( \mathbf{G}_L^{(\ell)}(x, x') \right)$  on

$$\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x') \rangle \geq 0\}} \text{Tr} \left( \mathbf{G}_{L-1}^{(\ell)}(x, x') \right).$$

This is achieved by applying Lemma 3.1.4.

In Step 2 and 3, we show  $\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x') \rangle \geq 0\}}$  concentrates on  $q_{L-1}^{(\ell-1)}(x, x')$ .

**Step 2:** Here, we show that  $\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(L)}, o^{(L-1)}(x') \rangle \geq 0\}}$  concentrates on  $\mathbb{E}_{w \sim \mathcal{N}(0, \mathbf{I})} \left[ \mathbf{1}_{\{\langle w, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w, o^{(L-1)}(x') \rangle \geq 0\}} \right]$ .

**Lemma 3.1.5.** Let  $\{w_i\}_{i=1}^m \in \mathbb{R}^d$  be i.i.d. Gaussian random vectors with standard normal entries. Define for  $0 \leq \ell \leq L-1$ ,

$$\begin{aligned} & h_{x, x'}^{(\ell+1)}(z_1, \dots, z_m) \\ & \triangleq \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle z_i, o^{(\ell)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle z_i, o^{(\ell)}(x') \rangle \geq 0\}} - \mathbb{E}_{w \sim \mathcal{N}(0, \mathbf{I})} \left[ \mathbf{1}_{\{\langle w, o^{(\ell)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w, o^{(\ell)}(x') \rangle \geq 0\}} \right] \right|. \end{aligned}$$

Conditioning on  $\{\mathbf{W}^{(k)}\}_{k=1}^{\ell}$ , with probability at least  $1 - \exp(-2m^{1/3})$ ,

$$\sup_{x, x'} h_{x, x'}^{(\ell+1)}(w_1, \dots, w_m) \leq C \sqrt{\frac{d(1 + \ell \log m)}{m}} + \frac{1}{m^{1/3}}.$$

*Proof of Lemma 3.1.5.* Throughout the proof, we condition on  $\{\mathbf{W}^{(k)}\}_{k=1}^{\ell}$ . We first show that

$$\begin{aligned} & \mathbb{P} \left[ \sup_{x, x'} h_{x, x'}^{(\ell+1)}(w_1, \dots, w_m) \leq \mathbb{E} \left[ \sup_{x, x'} h_{x, x'}^{(\ell+1)}(w_1, \dots, w_m) \right] + \frac{1}{m^{1/3}} \right] \\ & \geq 1 - \exp \left( -2m^{1/3} \right). \end{aligned} \tag{3.26}$$

To prove this, note that by the triangle inequality, for arbitrary  $i$ ,  $\exists(x_0, x'_0) \in \mathbb{S}^{d-1}$  such

that

$$\begin{aligned}
& \left| \sup_{x,x'} h^{(\ell+1)}(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_m) - \sup_{x,x'} h^{(\ell+1)}(z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_m) \right| \\
& \leq \frac{1}{m} \left| \mathbf{1}_{\{z_i, o^{(\ell)}(x_0) \geq 0\}} \mathbf{1}_{\{z_i, o^{(\ell)}(x'_0) \geq 0\}} - \mathbf{1}_{\{z'_i, o^{(\ell)}(x_0) \geq 0\}} \mathbf{1}_{\{z'_i, o^{(\ell)}(x'_0) \geq 0\}} \right| \\
& \leq \frac{1}{m}.
\end{aligned} \tag{3.27}$$

Therefore, (3.26) follows by applying McDiarmid's inequality (Lemma A.1.1).

To complete the proof of Lemma 3.1.5, it remains to show

$$\mathbb{E} \left[ \sup_{x,x'} h_{x,x'}^{(\ell+1)}(w_1, \dots, w_m) \right] = O \left( \sqrt{\frac{d(1 + \ell \log m)}{m}} \right). \tag{3.28}$$

Since  $\{w_i\}_{i=1}^m$  are i.i.d. conditional on  $\{\mathbf{W}^{(r)}\}_{r=1}^\ell$ , by Lemma A.2.2,

$$\mathbb{E} \left[ \sup_{x,x'} h_{x,x'}^{(\ell+1)}(w_1, \dots, w_m) \right] \leq C \sqrt{\frac{\text{VC}(\mathcal{H}^{(\ell+1)})}{m}}, \tag{3.29}$$

where

$$\mathcal{H}^{(\ell+1)}(\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(\ell)}) = \left\{ \alpha_{x,x'}^{(\ell)}(w) : x, x' \in \mathbb{S}^{d-1} \right\},$$

with  $\alpha_{x,x'}^{(\ell)}(w) = \mathbf{1}_{\{w, o^{(\ell)}(x) \geq 0\}} \mathbf{1}_{\{w, o^{(\ell)}(x') \geq 0\}}$ . Now we bound  $\text{VC}(\mathcal{H}^{(\ell+1)})$ . Let  $\mathcal{F}^{(\ell+1)} = \left\{ f_x(w) = \mathbf{1}_{\{w, o^{(\ell)}(x) \geq 0\}} : x \in \mathbb{S}^{d-1} \right\}$ . Then for any  $\alpha(w) \in \mathcal{H}^{(\ell+1)}$ , we can always find  $f(w)$  and  $g(w)$  in  $\mathcal{F}^{(\ell+1)}$  such that  $\alpha = f \times g$ . Thus, by Lemma A.2.1,  $\text{VC}(\mathcal{H}^{(\ell+1)}) \leq C \text{VC}(\mathcal{F}^{(\ell+1)})$  for some universal constant  $C$ . We claim that  $\text{VC}(\mathcal{F}^{(\ell+1)}) = O(d(1 + \ell \log m))$ . Plugging this bound into the above displayed equation and combining it with (3.29), we obtain (3.28).

Finally, we prove the claim. When  $\ell = 0$ , by (3.16), we have  $\text{VC}(\mathcal{F}^{(1)}) = d$ .

Now consider the case when  $\ell \geq 1$ . Similar to the proof of Lemma 3.1.3, we decompose  $\mathbb{S}^{d-1}$  into  $\mathcal{V} = \{V_j, j = 1, 2, \dots, |\mathcal{D}_\ell|\}$  where  $\cup V_j = \mathbb{S}^{d-1}$  and  $V_j \cap V_{j'} = \emptyset$  whenever  $j \neq j'$  such that for any  $x, x'$  within the same  $V_j$ ,

$$\left( \mathbf{D}^{(1)}(x), \dots, \mathbf{D}^{(\ell)}(x) \right) = \left( \mathbf{D}^{(1)}(x'), \dots, \mathbf{D}^{(\ell)}(x') \right).$$

Recall from (3.18) that  $\mathcal{F}_j^{(\ell+1)} \triangleq \left\{ f_x(w) = \mathbf{1}_{\{\langle w, o^{(\ell)}(x) \rangle \geq 0\}} : x \in V_j \right\}$ . Since  $\cup V_j = \mathbb{S}^{d-1}$ , we have  $\mathcal{F}^{(\ell+1)} = \cup_{j=1}^{|\mathcal{D}_\ell|} \mathcal{F}_j^{(\ell+1)}$ . From (3.20), we have  $\text{VC}(\mathcal{F}_j^{(\ell+1)}) \leq d$  for all  $j$ . Thus, by Lemma A.2.5, we have

$$\text{VC}(\mathcal{F}^{(\ell+1)}) = O(\max\{d \log d, \log |\mathcal{D}_\ell|\}) = O(\max\{d \log d, d\ell \log m\}) = O(d\ell \log m),$$

where the second equality holds by (3.15) which gives  $|\mathcal{D}_\ell| \leq m^{d\ell}$  and the last equality holds since  $\log d \leq \ell \log m$  as  $m \geq d$ .  $\square$

**Step 3:** Note that  $\mathbb{E}_{w \sim \mathcal{N}(0, \mathbf{I})} \left[ \mathbf{1}_{\{\langle w, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w, o^{(L-1)}(x') \rangle \geq 0\}} \right] = \frac{\pi - \arccos \hat{\rho}^{(L-1)}(x, x')}{2\pi}$

where

$$\hat{\rho}^{(L-1)}(x, x') \triangleq \left\langle \frac{o^{(L-1)}(x)}{\|o^{(L-1)}(x)\|_2}, \frac{o^{(L-1)}(x')}{\|o^{(L-1)}(x')\|_2} \right\rangle. \quad (3.30)$$

To show the concentration of  $\mathbb{E}_{w \sim \mathcal{N}(0, \mathbf{I})} \left[ \mathbf{1}_{\{\langle w, o^{(L-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w, o^{(L-1)}(x') \rangle \geq 0\}} \right]$  on  $q_{L-1}^{(L-1)}(x, x')$ , we show the concentration of  $\arccos \hat{\rho}^{(L-1)}(x, x')$  on  $\arccos \rho^{(L-1)}(x, x')$  through the following corollary.

**Corollary 2.** Fix any  $\ell \leq L$ . Under  $\left( \cap_{k=1}^\ell \mathcal{E}_0^{(k)} \right) \cap \left( \cap_{k=1}^\ell \mathcal{E}_1^{(k)} \right)$  where  $\mathcal{E}_0^{(k)}$  is defined in (3.1) and  $\mathcal{E}_1^{(k)}$  is defined in (3.2),

$$\sup_{x, x'} \left| \hat{\rho}^{(\ell)}(x, x') - \rho^{(\ell)}(x, x') \right| = O\left( \frac{\ell C^\ell}{m^{1/3}} \right).$$

and hence

$$\sup_{x, x'} \left| \arccos \rho^{(\ell)}(x, x') - \arccos \hat{\rho}^{(\ell)}(x, x') \right| = O\left( \frac{\sqrt{\ell} C^\ell}{m^{1/6}} \right).$$

To see why Corollary 2 holds, note that Lemma 2.5.1 implies both the numerator and denominator of  $\hat{\rho}^{(\ell)}$  are close to those of  $\rho^{(\ell)}$ . To obtain the second bound of Corollary 2, we prove that the arccos function is Hölder continuous of order 1/2, that is,

$$\arccos z - \arccos y \leq \arccos(1 - (y - z)) \leq 3\sqrt{y - z}, \quad \forall 0 \leq z \leq y \leq 1. \quad (3.31)$$

Combining the above with the first bound of Corollary 2 finishes the proof.

*Proof of Corollary 2.* We first prove  $\widehat{\rho}^{(\ell)}(x, x')$  is close to  $\rho^{(\ell)}(x, x')$ . Note that for any  $y, y', z, z'$ , by the triangle inequality we have

$$\left| \frac{y}{z} - \frac{y'}{z'} \right| \leq \left| \frac{y - y'}{z} \right| + \left| \frac{y'}{z} - \frac{y'}{z'} \right| = \left| \frac{y - y'}{z} \right| + \left| \frac{y'(z' - z)}{zz'} \right| \leq \left| \frac{y - y'}{z} \right| + \left| \frac{z' - z}{z} \right|,$$

where the last inequality holds under the assumption that  $|y'/z'| \leq 1$ . Taking  $y = \mathbb{E} [\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))]$ ,  $y' = \langle o^{(\ell)}(x), o^{(\ell)}(x') \rangle$ ,  $z = \sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x))]} \sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x'))]}$ , and  $z' = \|o^{(\ell)}(x)\|_2 \|o^{(\ell)}(x')\|_2$ , by definition (3.25) and (3.30), we have

$$\widehat{\rho}^{(\ell)}(x, x') = \frac{y}{z}, \quad \rho^{(\ell)}(x, x') = \frac{y'}{z'},$$

By Cauchy Schwartz inequality,  $|\rho^{(\ell)}(x, x')| \leq 1$ . As a result, we have

$$\begin{aligned} \left| \widehat{\rho}^{(\ell)}(x, x') - \rho^{(\ell)}(x, x') \right| &\leq \underbrace{\left| \frac{\langle o^{(\ell)}(x), o^{(\ell)}(x') \rangle - \mathbb{E} [\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))]}{\sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x))]} \sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x'))]}} \right|}_{\text{(I)}} \\ &+ \underbrace{\left| \frac{\|o^{(\ell)}(x)\|_2 \|o^{(\ell)}(x')\|_2 - \sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x))]} \sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x'))]}}{\sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x))]} \sqrt{\mathbb{E} [\sigma^2(U^{(\ell)}(x'))]}} \right|}_{\text{(II)}}. \end{aligned}$$

Note that

$$\text{(I)} = 2^\ell \left| \langle o^{(\ell)}(x), o^{(\ell)}(x') \rangle - \mathbb{E} [\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))] \right| = O\left(\frac{\ell C^\ell}{m^{1/3}}\right)$$

where the first equality holds by (2.38) which gives

$$\mathbb{E} [\sigma^2(U^{(\ell)}(x))] = \mathbb{E} [\sigma^2(U^{(\ell)}(x'))] = 2^{-\ell}, \quad \forall x, x',$$

and the last equality holds by Lemma 2.5.1.

To bound (II), by (2.38), we have

$$(II) = 2^\ell \left| \left\| o^{(\ell)}(x) \right\|_2 \left\| o^{(\ell)}(x') \right\|_2 - \sqrt{\mathbb{E} [\sigma^2 (U^{(\ell)}(x))]} \sqrt{\mathbb{E} [\sigma^2 (U^{(\ell)}(x'))]} \right|. \quad (3.32)$$

Note that for any  $y, \tilde{y}, z, \tilde{z} \geq 0$ ,

$$\begin{aligned} |y\tilde{y} - z\tilde{z}| &\leq \tilde{y}|y - z| + z|\tilde{y} - \tilde{z}| \leq \tilde{y} \frac{|y^2 - z^2|}{y + z} + z \frac{|\tilde{y}^2 - \tilde{z}^2|}{\tilde{y} + \tilde{z}} \\ &\leq \tilde{y} \frac{|y^2 - z^2|}{z} + z \frac{|\tilde{y}^2 - \tilde{z}^2|}{\tilde{z}}. \end{aligned}$$

Taking  $y = \|o^{(\ell)}(x)\|_2$ ,  $\tilde{y} = \|o^{(\ell)}(x')\|_2$ ,  $z = \sqrt{\mathbb{E} [\sigma^2 (U^{(\ell)}(x))]}$ , and  $z' = \sqrt{\mathbb{E} [\sigma^2 (U^{(\ell)}(x'))]}$ , we have  $\tilde{y} \leq c_0^\ell$  by Lemma 3.1.1 under event  $\cap_{k=1}^\ell \mathcal{E}_0^{(k)}$ ,  $z, \tilde{z} = 2^{-\ell/2}$  by (2.38), and  $|y^2 - z^2|$  and  $|\tilde{y}^2 - \tilde{z}^2|$  are upper bounded by  $\ell C^\ell / m^{1/3}$  by Lemma 2.5.1. Therefore,

$$\begin{aligned} &\left| \left\| o^{(\ell)}(x) \right\|_2 \left\| o^{(\ell)}(x') \right\|_2 - \sqrt{\mathbb{E} [\sigma^2 (U^{(\ell)}(x))]} \sqrt{\mathbb{E} [\sigma^2 (U^{(\ell)}(x'))]} \right| \\ &= O \left( c_0^\ell \frac{\ell C^\ell / m^{1/3}}{2^{-\ell/2}} \right) + O \left( \frac{\ell C^\ell}{m^{1/3}} \right) = O \left( \frac{\ell C^\ell}{m^{1/3}} \right). \end{aligned}$$

Plugging the above bound into (3.32), we have (II) =  $O \left( \frac{\ell C^\ell}{m^{1/3}} \right)$ .

Combining the bound of (I) and (II), we have for any  $x$  and  $x'$ ,

$$|\hat{\rho}^{(\ell)}(x, x') - \rho^{(\ell)}(x, x')| = O \left( \frac{\ell C^\ell}{m^{1/3}} \right). \quad (3.33)$$

Next we prove  $\arccos \hat{\rho}^{(\ell)}(x, x')$  is close to  $\arccos \rho^{(\ell)}(x, x')$  for any  $x$  and  $x'$  on  $\mathbb{S}^{d-1}$ . For notation simplicity, in the remaining part of the proof, we denote  $\rho$  as  $\rho^{(\ell)}$  and  $\hat{\rho}$  as  $\hat{\rho}^{(\ell)}$ . Here, we claim for any  $y$  and  $z$ ,  $|\arccos y - \arccos z| \leq 3\sqrt{|y - z|}$ . Given the claim, taking  $y = \hat{\rho}$  and  $z = \rho$ , we complete the proof since  $|\arccos \rho - \arccos \hat{\rho}| \leq 3\sqrt{|\rho - \hat{\rho}|} = O \left( \sqrt{\ell} C^\ell m^{-1/6} \right)$ .

Now we prove the claim. WLOG, assume  $\rho \leq \hat{\rho} \leq 1$ , so  $\arccos \rho \geq \arccos \hat{\rho}$ . By [ASR88, 4.4.33], we have  $\arccos \rho - \arccos \hat{\rho} = \arccos \left( \rho \hat{\rho} + \sqrt{1 - \rho^2} \sqrt{1 - \hat{\rho}^2} \right) \triangleq \arccos \xi$ . Define  $\delta \triangleq \hat{\rho} - \rho$ . Note that  $\xi = \rho \hat{\rho} + \sqrt{1 - \rho^2} \sqrt{1 - \hat{\rho}^2} \geq \hat{\rho}^2 - \delta \hat{\rho} + 1 - \hat{\rho}^2 \geq 1 - \delta$ , where the second

inequality holds by  $1 - \rho^2 \geq 1 - \widehat{\rho}^2$  and the last inequality holds by  $\widehat{\rho} \leq 1$ .

Since arccos function is monotonic decreasing, it remains to show  $\arccos(1 - \delta) \leq 3\sqrt{\delta}$ . Denote  $h(x) = 3\sqrt{x} - \arccos(1 - x)$ ,  $x \in (0, 1]$ . Since  $\frac{dh}{dx} = \frac{1}{\sqrt{x}} \left(3 - \frac{2}{\sqrt{2-x}}\right) > 0$  for any  $x \in (0, 1]$  and  $h(0) = 0$ , we have  $\arccos(1 - x) \leq 3\sqrt{x}$  for any  $x \in [0, 1]$ .  $\square$

**Proof of Lemma 2.5.3:** Denote for any  $k$  and all  $\ell \leq k$ ,

$$\mathcal{E}_{d,k}^{(\ell)} = \left\{ \sup_{x,x'} \left| \text{Tr} \left( \mathbf{G}_k^{(\ell)}(x, x') \right) - q_k^{(\ell)}(x, x') \right| = O \left( \frac{\sqrt{k} C^k}{m^{1/6}} + \sqrt{\frac{d \log^{k-1} m}{m}} \right) \right\},$$

and  $\mathcal{E}_{d,k} = \cap_{\ell=1}^k \mathcal{E}_{d,k}^{(\ell)}$ .

Note that under  $\mathcal{E}_{d,L}$ , (2.32) holds directly. Thus, it suffices to prove

$$\mathbb{P}[\mathcal{E}_{d,L} \cap \mathcal{E}_0 \cap \mathcal{E}_1] = 1 - \exp \left( O(dL \log m) - \Omega(m^{1/3}) \right).$$

where  $\mathcal{E}_0 = \cap_{k=1}^L \mathcal{E}_0^{(k)}$  and  $\mathcal{E}_1 = \cap_{k=1}^L \mathcal{E}_1^{(k)}$  with  $\mathcal{E}_0^{(k)}$  defined in (3.1) and  $\mathcal{E}_1^{(k)}$  defined in (3.2). For notation simplicity, denote  $\mathcal{V}^{(k)} = \cap_{r=1}^k \left( \mathcal{E}_0^{(r)} \cap \mathcal{E}_1^{(r)} \right)$ . By the second inequality of Lemma A.4.1, we have

$$\mathbb{P}[\mathcal{E}_{d,L} \cap \mathcal{E}_0 \cap \mathcal{E}_1] \geq 1 - \sum_{k=1}^L \mathbb{P} \left[ \mathcal{E}_{d,k}^c | \mathcal{E}_{d,k-1} \cap \mathcal{V}^{(k-1)} \right] - \sum_{k=1}^L \mathbb{P} \left[ \left( \mathcal{V}^{(k)} \right)^c \right]. \quad (3.34)$$

By (3.6), we have

$$\mathbb{P} \left[ \mathcal{V}^{(k)} \right] \geq 1 - k \exp \left( O(d \log m) - \Omega(m^{1/3}) \right). \quad (3.35)$$

Next we bound  $\mathbb{P} \left[ \mathcal{E}_{d,k}^c | \mathcal{E}_{d,k-1} \cap \mathcal{V}^{(k-1)} \right]$ . By definition of  $\mathcal{E}_{d,k}$ ,

$$\mathbb{P} \left[ \mathcal{E}_{d,k}^c | \mathcal{E}_{d,k-1} \cap \mathcal{V}^{(k-1)} \right] \leq \sum_{\ell=1}^k \mathbb{P} \left[ \left( \mathcal{E}_{d,k}^{(\ell)} \right)^c | \mathcal{E}_{d,k-1} \cap \mathcal{V}^{(k-1)} \right]. \quad (3.36)$$

In the remaining proof, we condition on  $\{\mathbf{W}^{(r)}\}_{r=1}^{k-1}$  such that  $\mathcal{E}_{d,k-1} \cap \mathcal{V}^{(k-1)}$  holds.

**Case 1  $\ell = k$ :** By definition,

$$\mathrm{Tr} \left( \mathbf{G}_k^{(k)}(x, x') \right) = \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}}, \quad (3.37)$$

and

$$q_k^{(k)}(x, x') = \frac{\pi - \arccos \rho^{(k-1)}(x, x')}{2\pi},$$

where  $\rho^{(k-1)}$  is defined in (3.25).

By the triangle inequality, we have

$$\begin{aligned} & \left| \mathrm{Tr} \left( \mathbf{G}_k^{(k)}(x, x') \right) - q_k^{(k)}(x, x') \right| \\ &= \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}} - \frac{\pi - \arccos \rho^{(k-1)}(x, x')}{2\pi} \right| \\ &\leq \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}} - \frac{\pi - \arccos \widehat{\rho}^{(k-1)}(x, x')}{2\pi} \right| \\ &+ \left| \frac{1}{2\pi} \left( \arccos \widehat{\rho}^{(k-1)}(x, x') - \arccos \rho^{(k-1)}(x, x') \right) \right|, \end{aligned} \quad (3.38)$$

where  $\widehat{\rho}^{(k-1)}$  is defined in (3.30).

Under  $\mathcal{V}^{(k-1)}$ , by Corollary 2, we have

$$\sup_{x, x'} \left| \frac{1}{2\pi} \left( \arccos \widehat{\rho}^{(k-1)}(x, x') - \arccos \rho^{(k-1)}(x, x') \right) \right| = O \left( \frac{\sqrt{k-1} C^{k-1}}{m^{1/6}} \right). \quad (3.39)$$

Now we bound the first term on the RHS of (3.38). Note that  $\{w_i^{(k)}\}_{i=1}^m$  is independent of  $\{\mathbf{W}^{(\ell)}\}_{\ell=1}^{k-1}$  and

$$\mathbb{E}_{w_i^{(k)}} \left[ \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}} \right] = \frac{\pi - \arccos \widehat{\rho}^{(k-1)}(x, x')}{2\pi}.$$

Thus, by Lemma 3.1.5, with probability at least  $1 - \exp(-2m^{1/3})$ ,

$$\begin{aligned} & \sup_{x, x'} \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}} - \frac{\pi - \arccos \widehat{\rho}^{(k-1)}(x, x')}{2\pi} \right| \\ &= O \left( \sqrt{\frac{d(1 + (k-1) \log m)}{m}} + \frac{1}{m^{1/3}} \right), \end{aligned}$$

Combining the above displayed equation with (3.39), we have with probability at least  $1 - \exp(-2m^{1/3})$ ,

$$\begin{aligned} \sup_{x, x'} \left| \text{Tr} \left( \mathbf{G}_k^{(k)}(x, x') \right) - q_k^{(k)}(x, x') \right| &= O \left( \frac{\sqrt{k-1} C^{k-1}}{m^{1/6}} + \sqrt{\frac{d(1 + (k-1) \log m)}{m}} + \frac{1}{m^{1/3}} \right) \\ &= O \left( \frac{\sqrt{k-1} C^{k-1}}{m^{1/6}} + \sqrt{\frac{d(1 + (k-1) \log m)}{m}} \right). \end{aligned}$$

Thus,

$$\mathbb{P} \left[ \mathcal{E}_{d,k}^{(k)} | \mathcal{E}_{d,k-1} \cap \mathcal{V}^{(k-1)} \right] \geq 1 - \exp(-2m^{1/3}). \quad (3.40)$$

**Case 2**  $\ell < k$ : By the definition of  $\mathbf{G}_k^{(\ell)}$ , we have

$$\begin{aligned} & \text{Tr} \left( \mathbf{G}_k^{(\ell)}(x, x') \right) \\ &= \text{Tr} \left( \mathbf{D}^{(k)}(x) \mathbf{W}^{(k)} \mathbf{G}_{k-1}^{(\ell)}(x, x') \left[ \mathbf{W}^{(k)} \right]^\top \mathbf{D}^{(k)}(x') \right) \\ &= \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}} \left[ w_i^{(k)} \right]^\top \mathbf{G}_{k-1}^{(\ell)}(x, x') w_i^{(k)}. \end{aligned}$$

Thus, by the triangle inequality, we have

$$\begin{aligned}
& \sup_{x,x'} \left| \text{Tr} \left( \mathbf{G}_k^{(\ell)}(x, x') \right) - q_k^{(\ell)}(x, x') \right| \\
& \leq \sup_{x,x'} \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}} \left[ w_i^{(k)} \right]^\top \mathbf{G}_{k-1}^{(\ell)}(x, x') w_i^{(k)} \right. \\
& \quad \left. - \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}} q_{k-1}^{(\ell)}(x, x') \right| \\
& + \sup_{x,x'} \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}} q_{k-1}^{(\ell)}(x, x') - q_k^{(\ell)}(x, x') \right| \\
& \leq \underbrace{\sup_{x,x',i} \left| \left[ w_i^{(k)} \right]^\top \mathbf{G}_{k-1}^{(\ell)}(x, x') w_i^{(k)} - q_{k-1}^{(\ell)}(x, x') \right|}_{\text{(I)}} \\
& + \underbrace{\sup_{x,x'} \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x) \rangle \geq 0\}} \mathbf{1}_{\{\langle w_i^{(k)}, o^{(k-1)}(x') \rangle \geq 0\}} q_{k-1}^{(\ell)}(x, x') - q_k^{(\ell)}(x, x') \right|}_{\text{(II)}}.
\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}
\text{(I)} & \leq \sup_{x,x', 1 \leq i \leq m} \left| \left[ w_i^{(k)} \right]^\top \mathbf{G}_{k-1}^{(\ell)}(x, x') w_i^{(k)} - \text{Tr} \left( \mathbf{G}_{k-1}^{(\ell)}(x, x') \right) \right| \\
& + \sup_{x,x'} \left| \text{Tr} \left( \mathbf{G}_{k-1}^{(\ell)}(x, x') \right) - q_{k-1}^{(\ell)}(x, x') \right| \\
& \leq \sup_{x,x', 1 \leq i \leq m} \left| \left[ w_i^{(k)} \right]^\top \mathbf{G}_{k-1}^{(\ell)}(x, x') w_i^{(L)} - \text{Tr} \left( \mathbf{G}_{k-1}^{(\ell)}(x, x') \right) \right| \\
& + O \left( \frac{\sqrt{k-1} C^{k-1}}{m^{1/6}} + \sqrt{\frac{d(1+(k-2)\log m)}{m}} \right), \tag{3.41}
\end{aligned}$$

where the last equality holds under  $\mathcal{E}_{d,k-1}^{(\ell)}$ .

We next bound the first term in the RHS of (3.41). Since  $\{w_i^{(k)}\}_{i=1}^m$  are i.i.d.  $\mathcal{N}(0, \mathbf{I}_m)$  and are independent of  $\{\mathbf{W}^{(r)}\}_{r=1}^{k-1}$ , by Lemma 3.1.4, for any  $i \in [m]$ ,

$$\begin{aligned}
& \mathbb{P}_{w_i^{(k)}} \left[ \sup_{x,x'} \left| \left[ w_i^{(k)} \right]^\top \mathbf{G}_{k-1}^{(\ell)}(x, x') w_i^{(k)} - \text{Tr} \left( \mathbf{G}_{k-1}^{(\ell)}(x, x') \right) \right| > \frac{c_0^{2k-2-2\ell}}{m^{1/3}} \middle| \mathcal{E}_{d,k-1} \cap \mathcal{V}^{(k-1)} \right] \\
& = \exp \left( O(d(k-1)\log m) - \Omega(m^{1/3}) \right).
\end{aligned}$$

Further taking union bounds over  $i$ , we have

$$\begin{aligned} & \mathbb{P} \left[ \sup_{x, x', i \in [m]} \left| \left[ w_i^{(k)} \right]^\top \mathbf{G}_{k-1}^{(\ell)}(x, x') w_i^{(k)} - \text{Tr}(\mathbf{G}_{k-1}^{(\ell)}(x, x')) \right| > \frac{c_0^{2k-2-2\ell}}{m^{1/3}} \middle| \mathcal{E}_{d, k-1} \cap \mathcal{V}^{(k-1)} \right] \\ & \leq m \exp \left[ O(d(k-1) \log m) - \Omega(m^{1/3}) \right] = \exp \left[ O(d(k-1) \log m) - \Omega(m^{1/3}) \right]. \end{aligned} \quad (3.42)$$

Plugging (3.42) into (3.41), we get

$$\begin{aligned} & \mathbb{P} \left[ \text{(I)} = O \left( \frac{\sqrt{k-1} C^{k-1}}{m^{1/6}} + \sqrt{\frac{d(1+(k-2) \log m)}{m}} \right) + \frac{c_0^{2k-2-2\ell}}{m^{1/3}} \middle| \mathcal{E}_{d, k-1} \cap \mathcal{V}^{(k-1)} \right] \\ & = \mathbb{P} \left[ \text{(I)} = O \left( \frac{\sqrt{k-1} C^{k-1}}{m^{1/6}} + \sqrt{\frac{d(1+(k-2) \log m)}{m}} \right) \middle| \mathcal{E}_{d, k-1} \cap \mathcal{V}^{(k-1)} \right] \\ & \geq 1 - \exp \left( O(d(k-1) \log m) - \Omega(m^{1/3}) \right). \end{aligned}$$

To bound (II), we have with probability at least  $1 - \exp(-2m^{1/3})$ ,

$$\begin{aligned} \text{(II)} & \stackrel{(a)}{=} \sup_{x, x'} \left| q_{k-1}^{(\ell)}(x, x') \left( \text{Tr} \left( \mathbf{G}_k^{(k)} \right) - q_k^{(k)}(x, x') \right) \right| \\ & \stackrel{(b)}{\leq} \sup_{x, x'} \left| \text{Tr} \left( \mathbf{G}_k^{(k)}(x, x') \right) - q_k^{(k)}(x, x') \right| \\ & = O \left( \frac{\sqrt{k} C^k}{m^{1/6}} + \sqrt{\frac{d(1+(k-1) \log m)}{m}} \right). \end{aligned}$$

where (a) holds by (2.19), that is  $q_k^{(\ell)}(x, x') = q_{k-1}^{(\ell)}(x, x') q_k^{(k)}(x, x')$ , and (3.37); (b) holds as  $\sup_{x, x'} \left| q_{k-1}^{(\ell)}(x, x') \right| \leq 1$ ; and the last equality holds from (3.40).

Combining the above bounds on (I) and (II), we have for any  $\ell < k$ ,

$$\mathbb{P} \left[ \left[ \mathcal{E}_{d, k}^{(\ell)} \right]^c \middle| \mathcal{E}_{d, k-1} \cap \mathcal{V}_0^{(k-1)} \cap \mathcal{V}_1^{(k-1)} \right] \leq \exp \left[ O(d(k-1) \log m) - \Omega(m^{1/3}) \right] + \exp \left( -2m^{1/3} \right).$$

Combining the last displayed equation with (3.36) and (3.40) yields that

$$\begin{aligned} & \mathbb{P} \left[ \mathcal{E}_{d, k}^c \middle| \mathcal{E}_{d, k-1} \cap \mathcal{V}_0^{(k-1)} \cap \mathcal{V}_1^{(k-1)} \right] \\ & \leq (k-1) \exp \left[ O(d(k-1) \log m) - \Omega(m^{1/3}) \right] + k \exp \left( -2m^{1/3} \right). \end{aligned} \quad (3.43)$$

Plugging (3.43) and (3.35) into (3.34), we get

$$\begin{aligned}
& \mathbb{P}[\mathcal{E}_{d,k} \cap \mathcal{E}_0 \cap \mathcal{E}_1] \\
& \geq 1 - \sum_{k=1}^L \left[ (k-1) \exp \left[ O(d(k-1) \log m) - \Omega(m^{1/3}) \right] + k \exp \left( -2m^{1/3} \right) \right] \\
& \quad - \sum_{k=1}^L \exp \left( O(d \log m) - \Omega(m^{1/3}) \right) \\
& = 1 - \exp \left[ O(dL \log m) - \Omega(m^{1/3}) \right].
\end{aligned}$$

## 3.2 Proofs in Section 2.6

Recall  $\mathcal{E}_0 = \cap_{\ell=1}^L \mathcal{E}_0^{(\ell)}$  where  $\mathcal{E}_0^{(\ell)}$  is defined in (3.1). By (3.24), we have

$$\mathbb{P}[\mathcal{E}_0] \geq 1 - L \exp(-\Omega(m)). \quad (3.44)$$

### 3.2.1 Proof of Proposition 2.6.1

Throughout the proof, we assume  $\mathcal{E}_0$ , all three conclusions of Lemma 2.6.2, conclusion of Lemma 2.6.3, and conclusions of Lemma 2.6.4 hold. These altogether can be guaranteed with probability at least  $1 - \exp(-\Omega(C^{-L}m^{1/36}))$  by union bounds following (3.44) and Lemma 2.6.2–2.6.4.

Recall from (2.39) that

$$\frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} = \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(x) z_t^{(\ell)}(x) \left[ o_t^{(\ell-1)}(x) \right]^\top,$$

where

$$\left[ z_t^{(\ell)}(x) \right]^\top = a^\top \frac{1}{\sqrt{m}} \mathbf{D}_t^{(L)}(x) \mathbf{W}^{(L)}(t) \cdots \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell+1)}(x) \mathbf{W}^{(\ell+1)}(t). \quad (3.45)$$

Therefore, by the triangle inequality, we have

$$\begin{aligned}
& \left\| \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} - \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 \\
&= \frac{1}{\sqrt{m}} \left\| \mathbf{D}_t^{(\ell)}(x) z_t^{(\ell)}(x) \left[ o_t^{(\ell-1)}(x) \right]^\top - \mathbf{D}_0^{(\ell)}(x) z_0^{(\ell)}(x) \left[ o_0^{(\ell-1)}(x) \right]^\top \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \left\| \mathbf{D}_t^{(\ell)}(x) \left( z_t^{(\ell)}(x) - z_0^{(\ell)}(x) \right) \left[ o_t^{(\ell-1)}(x) \right]^\top \right\|_2 \\
&+ \frac{1}{\sqrt{m}} \left\| \left( \mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_0^{(\ell)}(x) \left[ o_t^{(\ell-1)}(x) \right]^\top \right\|_2 \\
&+ \frac{1}{\sqrt{m}} \left\| \mathbf{D}_0^{(\ell)}(x) z_0^{(\ell)}(x) \left( o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right)^\top \right\|_2. \tag{3.46}
\end{aligned}$$

Now we bound the first term on the right hand side of (3.46). Note that

$$\begin{aligned}
\sup_x \left\| o_t^{(\ell-1)}(x) \right\|_2 &\leq \sup_x \left\| o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right\|_2 + \sup_x \left\| o_0^{(\ell-1)}(x) \right\|_2 \\
&\leq \frac{C^{\ell-1}}{m^{1/6}} + c_0^{\ell-1} \leq C^{\ell-1}, \tag{3.47}
\end{aligned}$$

where the second inequality holds by (2.45) in Lemma 2.6.2, and Lemma 3.1.1 under  $\mathcal{E}_0$ .

Since  $\left\| \mathbf{D}_t^{(\ell)}(x) \right\|_2 \leq 1$  for all  $x$  and  $t$ , we have

$$\begin{aligned}
& \frac{1}{\sqrt{m}} \left\| \mathbf{D}_t^{(\ell)}(x) \left( z_t^{(\ell)}(x) - z_0^{(\ell)}(x) \right) \left[ o_t^{(\ell-1)}(x) \right]^\top \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \left\| z_t^{(\ell)}(x) - z_0^{(\ell)}(x) \right\|_2 \left\| o_t^{(\ell-1)}(x) \right\|_2 \\
&= O\left( C^{2L-1} m^{-1/36} \right) \tag{3.48}
\end{aligned}$$

where the last inequality holds by (2.49) from Lemma 2.6.4.

To bound the second term of (3.46), note that by the definition of  $\mathbf{D}_t^{(\ell)}$ , we have

$$\begin{aligned}
& \left\| \left( \mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_0^{(\ell)}(x) \right\|_2^2 \\
&= \sum_{i=1}^m \left( \mathbf{1}_{\{ \langle w_i^{(\ell)}(t), o_t^{(\ell-1)}(x) \rangle \geq 0 \}} - \mathbf{1}_{\{ \langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x) \rangle \geq 0 \}} \right)^2 \left[ z_0^{(\ell)}(x) \right]_i^2 \\
&\leq \left\| z_0^{(\ell)}(x) \right\|_\infty^2 \sum_{i=1}^m \left| \mathbf{1}_{\{ \langle w_i^{(\ell)}(t), o_t^{(\ell-1)}(x) \rangle \geq 0 \}} - \mathbf{1}_{\{ \langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x) \rangle \geq 0 \}} \right| \\
&\leq 4m^{1/18} S_t^{(\ell)}(x) = O\left( C^\ell m^{17/18} \right)
\end{aligned}$$

where the last inequality holds by  $\sup_x \left\| z_0^{(\ell)}(x) \right\|_\infty \leq m^{1/36}$  from (2.48) in Lemma 2.6.4, and the definition of  $S_t(x)$  and the last equality holds by (2.47) from Lemma 2.6.3, i.e.,  $\sup_x S_t^{(\ell)}(x) \leq C_2^\ell m^{8/9}$ . Thus, by (3.47), we have

$$\begin{aligned}
& \frac{1}{\sqrt{m}} \left\| \left( \mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_0^{(\ell)}(x) \left[ o_t^{(\ell-1)}(x) \right]^\top \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \left\| \left( \mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_0^{(\ell)}(x) \right\|_2 \left\| o_t^{(\ell-1)}(x) \right\|_2 \\
&= \frac{1}{\sqrt{m}} O\left( C^{2\ell-1} m^{17/36} \right) = O\left( C^{2\ell-1} m^{-1/36} \right). \tag{3.49}
\end{aligned}$$

Now we bound the last term on the right hand side of (3.46). By the definition of  $z_t^{(\ell)}(x)$  in (3.45), since  $\left\| \mathbf{D}_0^{(\ell)}(x) \right\|_2 \leq 1$  for all  $x$ , under  $\mathcal{E}_0$ , we have

$$\left\| z_0^{(\ell)}(x) \right\|_2 \leq \prod_{k=\ell+1}^L \left\| \frac{1}{\sqrt{m}} \mathbf{D}_0^{(k)}(x) \mathbf{W}^{(k)}(0) \right\|_2 \|a\|_2 \leq c_0^{L-\ell} \sqrt{m}. \tag{3.50}$$

Thus, we have

$$\begin{aligned}
& \frac{1}{\sqrt{m}} \left\| \mathbf{D}_0^{(\ell)}(x) z_0^{(\ell)}(x) \left( o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right)^\top \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \left\| z_0^{(\ell)}(x) \right\|_2 \left\| o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right\|_2 \leq \frac{C^L}{m^{1/6}}, \tag{3.51}
\end{aligned}$$

where the last inequality holds by (3.50) and (2.45) of Lemma 2.6.2.

Plugging (3.48), (3.49) and (3.51) back into (3.46), we get

$$\begin{aligned} \left\| \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} - \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 &= O \left( C^{2L-1} m^{-1/36} + C^{2\ell-1} m^{-1/36} + \frac{C^L}{m^{1/6}} \right) \\ &= O \left( C^{2L} m^{-1/36} \right). \end{aligned} \quad (3.52)$$

Next, we prove  $\|H_t - H_0\|_\infty = O(C^{2L} m^{-1/36})$ . By (3.9), we have

$$\begin{aligned} & \left| H_t^{(\ell)}(x, x') - H^{(\ell)}(x, x') \right| \\ &= \left| \left\langle \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}, \frac{\partial f(x'; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} \right\rangle - \left\langle \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}, \frac{\partial f(x'; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\rangle \right| \\ &\leq \left\| \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} - \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 \left\| \frac{\partial f(x'; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 \\ &\quad + \left\| \frac{\partial f(x'; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} - \frac{\partial f(x'; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 \left\| \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 \\ &= O \left( C^{2L} m^{-1/36} \right) \left( \left\| \frac{\partial f(x'; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 + \left\| \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 \right), \end{aligned} \quad (3.53)$$

where the last equality holds by (3.52).

From (3.50) and Lemma 3.1.1, we get under  $\mathcal{E}_0$ ,

$$\left\| \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 \leq \frac{1}{\sqrt{m}} \left\| z_0^{(\ell)}(x) \right\|_2 \left\| o_0^{(\ell-1)}(x) \right\|_2 = O(c_0^L). \quad (3.54)$$

By the triangle inequality, we further have

$$\left\| \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 \leq \left\| \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 + \left\| \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} - \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 = O(C^{2L})$$

where the last equality holds by plugging in (3.54) and (3.52).

Plugging the above bound and (3.54) into (3.53), we complete the proof.

### 3.2.2 Proof of Lemma 2.6.2

**Step 1, showing  $R_t \leq m^{1/3}$ :** Recall that  $R_0 = m^{5/18}$  and

$$R_{t+1} = R_0 + LC^{2L-2} \sum_{s=0}^t \eta_s (R_s + \gamma).$$

Therefore  $R_{t+1} + \gamma - (R_t + \gamma) = LC^{2L-2} \eta_t (R_t + \gamma)$ , which is equivalent as  $R_{t+1} + \gamma = (1 + LC^{2L-2} \eta_t) (R_t + \gamma)$ . Thus,

$$\begin{aligned} R_t + \gamma &= \prod_{s=0}^{t-1} (1 + LC^{2L-2} \eta_s) (R_0 + \gamma) \\ &\leq \exp \left( LC^{2L-2} \sum_{s=0}^{t-1} \eta_s \right) (R_0 + \gamma) \leq 2 \exp (LC^{2L-2} \theta \log T) m^{5/18} \end{aligned}$$

where the second inequality holds since  $1 + z \leq e^z$  for any  $z$  and the last equality holds by plugging in  $\eta_s \leq \frac{\theta}{s+1}$  and the fact that  $R_0 + \gamma \leq 2m^{5/18}$ .

As a result, when  $m = \exp(\Omega(LC^{2L-2} \theta \log T))$ , we have  $R_t \leq R_t + \gamma \leq m^{1/3}$  for all  $t \leq T$ .

In the following Step 2, we show  $\mathcal{E}_0 \cap \mathcal{E}_3$  occurs with high probability where

$$\mathcal{E}_3 \triangleq \left\{ \sup_x |\Delta_0(x)| \leq m^{5/18}, \forall t \leq T \right\},$$

with  $\Delta_0(x) = f^*(x) - f(x; \mathbf{W}(t))$ .

Then in Step 3, we use an inductive argument to show under  $\mathcal{E}_0 \cap \mathcal{E}_3$ , (2.44)–(2.46) in Lemma 2.6.2 hold for all  $t \leq T$ .

**Step 2, bounding  $\mathbb{P}[\mathcal{E}_0 \cap \mathcal{E}_3]$ :** Note that it suffices to show

$$\mathbb{P}[\mathcal{E}_3 | \mathcal{E}_0] \geq 1 - \exp \left( -\Omega(C^{-L} m^{1/9}) \right). \quad (3.55)$$

With (3.55), by (3.44), we then have

$$\begin{aligned}\mathbb{P}[\mathcal{E}_0 \cap \mathcal{E}_3] &\geq \left(1 - \exp\left(-\Omega(C^{-L}m^{1/9})\right)\right) (1 - L \exp(-\Omega(m))) \\ &= 1 - \exp\left(-\Omega(C_1^{-L}m^{1/9})\right),\end{aligned}\tag{3.56}$$

for some constant  $C$  and  $C_1$ .

Now we prove (3.55). By the triangle inequality, we have  $\sup_x |\Delta_0(x)| \leq \sup_x |f^*(x)| + \sup_x |f(x; \mathbf{W}(0))|$ . Since  $\sup_x |f^*(x)| \leq m^{5/18}$  by assumption, we have

$$\mathbb{P}[\mathcal{E}_3 | \mathcal{E}_0] \geq \mathbb{P}\left[\sup_x f^2(x; \mathbf{W}(0)) \leq m^{5/9} \mid \mathcal{E}_0\right].$$

Thus, it remains to show

$$\mathbb{P}\left[\sup_x f^2(x; \mathbf{W}(0)) \leq m^{5/9} \mid \mathcal{E}_0\right] = 1 - \exp\left(-\Omega(C^{-L}m^{1/9})\right).$$

Throughout the remaining proof of Step 2, we condition on  $\{\mathbf{W}^{(k)}(0)\}_{k=1}^L$  such that  $\mathcal{E}_0$  holds. Following the definition of  $f$  in (2.4),

$$\begin{aligned}f^2(x; \mathbf{W}(0)) &= \left[a^\top \left(\frac{1}{\sqrt{m}}\mathbf{D}_0^{(L)}(x)\mathbf{W}^{(L)}(0) \cdots \frac{1}{\sqrt{m}}\mathbf{D}_0^{(1)}(x)\mathbf{W}^{(1)}(0)\right) x\right]^2 \\ &\leq \left\|a^\top \left(\frac{1}{\sqrt{m}}\mathbf{D}_0^{(L)}(x)\mathbf{W}^{(L)}(0) \cdots \frac{1}{\sqrt{m}}\mathbf{D}_0^{(1)}(x)\mathbf{W}^{(1)}(0)\right)\right\|_2^2 \\ &= a^\top \mathbf{Q}(x)a,\end{aligned}$$

where  $\mathbf{Q}(x) = \mathbf{V}(x)(\mathbf{V}(x))^\top$  with

$$\mathbf{V}(x) \triangleq \left(\frac{1}{\sqrt{m}}\mathbf{D}_0^{(L)}(x)\mathbf{W}^{(L)}(0) \cdots \frac{1}{\sqrt{m}}\mathbf{D}_0^{(1)}(x)\mathbf{W}^{(1)}(0)\right).$$

Under  $\mathcal{E}_0$ , we have  $\|\mathbf{Q}(x)\|_2 \leq c_0^{2L}$  and hence,  $\|\mathbf{Q}(x)\|_F \leq \sqrt{m} \|\mathbf{Q}(x)\|_2 \leq c_0^{2L} \sqrt{m}$ . Since

$a$  is independent with  $\{\mathbf{W}^{(k)}(0)\}_{k=1}^L$ , by Hanson-Wright inequality, for any fixed  $x$ ,

$$\begin{aligned} & \mathbb{P} \left[ \left| a^\top \mathbf{Q}(x) a \right| \geq m^{5/9} \left| \left\{ \mathbf{W}^{(k)}(0) \right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds} \right] \\ & \leq 2 \exp \left( -C \min \left( \frac{m^{10/9}}{c_0^{4L} m}, \frac{m^{5/9}}{c_0^{2L}} \right) \right) = \exp \left( -\Omega(C^{-L} m^{1/9}) \right). \end{aligned}$$

Denote  $\mathcal{Q}(\mathbf{W}^{(1)}(0), \dots, \mathbf{W}^{(L)}(0)) = \{\mathbf{Q}(x, x') : x, x' \in \mathbb{S}^{d-1}\}$ . Note that for any given  $\mathbf{W}^{(1)}(0), \dots, \mathbf{W}^{(L)}(0)$ , by the definition of  $\mathcal{D}_k$  in Section 3.1.2, we have

$$\mathcal{Q} \subset \left\{ \mathbf{V} \mathbf{V}^\top : \mathbf{V} = \mathbf{D}_L \mathbf{W}^{(L)}(0) \cdots \mathbf{D}_1 \mathbf{W}^{(1)}(0), (\mathbf{D}_1, \dots, \mathbf{D}_L) \in \mathcal{D}_L \right\}.$$

Thus by (3.15), we have  $|\mathcal{Q}| \leq |\mathcal{D}_L| \leq m^{dL}$ . Taking union bounds over  $\mathcal{Q}$ , for sufficiently large  $m$ , we have

$$\begin{aligned} & \mathbb{P} \left[ \sup_x \left| a^\top \mathbf{Q}(x) a \right| \geq m^{5/9} \left| \left\{ \mathbf{W}^{(k)}(0) \right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds} \right] \\ & = \mathbb{P} \left[ \sup_{\mathbf{Q} \in \mathcal{Q}} \left| a^\top \mathbf{Q} a \right| \geq m^{5/9} \left| \left\{ \mathbf{W}^{(k)}(0) \right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds} \right] \\ & \leq m^{dL} \exp \left( -\Omega(C^{-L} m^{1/9}) \right) = \exp \left( -\Omega(C^{-L} m^{1/9}) \right). \end{aligned}$$

Averaging over  $\{\mathbf{W}^{(k)}(0)\}_{k=1}^L$ , we get

$$\mathbb{P} \left[ \sup_x f^2(x; \mathbf{W}(0)) \leq m^{5/9} \mid \mathcal{E}_0 \right] = 1 - \exp \left( -\Omega(C^{-L} m^{1/9}) \right).$$

**Step 3, inductive argument to show small deviations of  $\mathbf{W}^{(\ell)}$ ,  $o^{(\ell)}$  and bounded  $\Delta_t$ :** Throughout step 3, we assume  $\mathcal{E}_0 \cap \mathcal{E}_3$  holds. Here, we use an inductive argument to show that under  $\mathcal{E}_0 \cap \mathcal{E}_3$ , (2.44)–(2.46) hold for all  $t \leq T$ .

When  $t = 0$ , (2.44) and (2.45) hold by definition. By the definition of  $\Delta_0$ , we know (2.46) holds under  $\mathcal{E}_3$ .

Now suppose (2.44)–(2.46) hold for some  $t$ . We first show (2.44) holds at  $t + 1$  by

showing

$$\left\| \mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t) \right\|_2 \leq C^{L-1} \eta_t (R_s + \gamma) \quad (3.57)$$

for all  $\ell$ .

By (2.10), we have

$$\begin{aligned} & \left\| \mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t) \right\|_2 \\ &= \eta_t |\Delta_t(X_t) + u_t| \left\| \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(X_t) \left( \prod_{k=\ell+1}^L \left[ \frac{1}{\sqrt{m}} \mathbf{W}^{(k)}(t) \right]^\top \mathbf{D}_t^{(k)}(X_t) \right) a \left[ o_t^{(\ell-1)}(X_t) \right]^\top \right\|_2 \\ &\leq \eta_t (R_t + \gamma) \left\| \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(X_t) \left( \prod_{k=\ell+1}^L \left[ \frac{1}{\sqrt{m}} \mathbf{W}^{(k)}(t) \right]^\top \mathbf{D}_t^{(k)}(X_t) \right) a \left[ o_t^{(\ell-1)}(X_t) \right]^\top \right\|_2, \end{aligned} \quad (3.58)$$

where the last inequality holds since  $|\Delta_t(X_t) + u_t| \leq \sup_x |\Delta_t(x)| + \sup |u_t| \leq R_t + \gamma$ .

Now we show

$$\left\| \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(X_t) \left( \prod_{k=\ell+1}^L \left[ \frac{1}{\sqrt{m}} \mathbf{W}^{(k)}(t) \right]^\top \mathbf{D}_t^{(k)}(X_t) \right) a \left[ o_t^{(\ell-1)}(X_t) \right]^\top \right\|_2 \leq C^{L-1}.$$

Plugging the above inequality into (3.58), we obtain (3.57).

By the triangle inequality, under  $\mathcal{E}_0 \cap \mathcal{E}_3$ , for any  $k$ , we have

$$\left\| \mathbf{W}^{(k)}(t) \right\|_2 \leq \left\| \mathbf{W}^{(k)}(t) - \mathbf{W}^{(k)}(0) \right\|_2 + \left\| \mathbf{W}^{(k)}(0) \right\|_2 = O(\sqrt{m}), \quad (3.59)$$

where the last equality holds since under  $\mathcal{E}_0 \cap \mathcal{E}_3$ ,  $\left\| \mathbf{W}^{(k)}(0) \right\|_2 = O(\sqrt{m})$  and

$$\left\| \mathbf{W}^{(k)}(t) - \mathbf{W}^{(k)}(0) \right\|_2 \leq C^{L-1} \sum_{s=0}^{t-1} \eta_s (R_s + \gamma) \leq R_t \leq m^{1/3}, \forall 0 \leq t \leq T. \quad (3.60)$$

Similarly, by the triangle inequality, we have

$$\sup_x \left\| o_t^{(\ell-1)}(x) \right\|_2 \leq \sup_x \left\| o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right\|_2 + \sup_x \left\| o_0^{(\ell-1)}(x) \right\|_2 \leq \frac{\ell c_0^\ell}{m^{1/6}} + c_0^{\ell-1} \leq C^{\ell-1}, \quad (3.61)$$

where the last inequality holds by Lemma 3.1.1 under  $\mathcal{E}_0$ .

As a result, under  $\mathcal{E}_0 \cap \mathcal{E}_3$ , we have

$$\begin{aligned}
& \left\| \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(X_t) \left( \prod_{k=\ell+1}^L \left[ \frac{1}{\sqrt{m}} \mathbf{W}^{(k)}(t) \right]^\top \mathbf{D}_t^{(k)}(X_t) \right) a \left[ o_t^{(\ell-1)}(X_t) \right]^\top \right\|_2 \\
& \leq \left\| \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(X_t) \right\|_2 \left( \prod_{k=\ell+1}^L \left\| \left[ \frac{1}{\sqrt{m}} \mathbf{W}^{(k)}(t) \right]^\top \mathbf{D}_t^{(k)}(X_t) \right\|_2 \right) \|a\|_2 \left\| o_t^{(\ell-1)}(X_t) \right\|_2 \\
& \leq C^{L-1},
\end{aligned}$$

where the last inequality holds by (3.59), (3.61) and the fact that  $\left\| \mathbf{D}_t^{(k)}(x) \right\|_2 \leq 1$  for any  $k$  and  $x$ .

Next, we show (2.45) holds at  $t+1$ . When  $\ell = 0$ , since  $o_{t+1}^{(0)}(x) = x$  for all  $t$ , we get

$$\sup_x \left\| o_{t+1}^{(0)}(x) - o_0^{(0)}(x) \right\|_2 = 0. \tag{3.62}$$

Fix arbitrary  $\ell$ . By the definition of  $o^{(\ell)}$ , under  $\mathcal{E}_0 \cap \mathcal{E}_3$ , for any  $x \in \mathbb{S}^{d-1}$ , we have

$$\begin{aligned}
& \left\| o_{t+1}^{(\ell+1)}(x) - o_0^{(\ell+1)}(x) \right\|_2 \\
& = \frac{1}{\sqrt{m}} \left\| \sigma(\mathbf{W}^{(\ell+1)}(t+1)) o_{t+1}^{(\ell)}(x) - \sigma(\mathbf{W}^{(\ell+1)}(0)) o_0^{(\ell)}(x) \right\|_2 \\
& \leq \frac{1}{\sqrt{m}} \left\| \left( \mathbf{W}^{(\ell+1)}(t+1) - \mathbf{W}^{(\ell+1)}(0) \right) o_{t+1}^{(\ell)}(x) \right\|_2 + \frac{1}{\sqrt{m}} \left\| \mathbf{W}^{(\ell+1)}(0) \left( o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x) \right) \right\|_2 \\
& \stackrel{(a)}{\leq} m^{-1/6} \left\| o_{t+1}^{(\ell)}(x) \right\|_2 + c_0 \left\| o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x) \right\|_2 \\
& \leq m^{-1/6} \left( \left\| o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x) \right\|_2 + \left\| o_0^{(\ell)}(x) \right\|_2 \right) + c_0 \left\| o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x) \right\|_2 \\
& \stackrel{(b)}{\leq} \left( c_0 + m^{-1/6} \right) \left\| o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x) \right\|_2 + c_0^\ell m^{-1/6},
\end{aligned}$$

where the first inequality holds by the triangle inequality, (a) holds by (3.60) and the definition of  $\mathcal{E}_0$  which gives  $\left\| \mathbf{W}^{(\ell+1)}(0) \right\|_2 \leq c_0 \sqrt{m}$ , and (b) holds by Lemma 3.1.1 under  $\mathcal{E}_0$ .

Recursively applying the above inequality and being aware of (3.62), we get for any  $x \in \mathbb{S}^{d-1}$ ,  $\left\| o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x) \right\|_2 \leq c_0^\ell m^{-1/6} \sum_{k=0}^{\ell-1} (c_0 + m^{-1/6})^k = O(C^\ell m^{-1/6})$ .

Finally, we show (2.46) holds at  $t+1$ . For notation simplicity, define  $\mathbf{E}^{(k)}(t) \triangleq \mathbf{W}^{(k)}(t) -$

$\mathbf{W}^{(k)}(0)$ . By the triangle inequality, we have for any  $x \in \mathbb{S}^{d-1}$ ,

$$\begin{aligned}
|\Delta_{t+1}(x)| &= |f^*(x) - f(x; \mathbf{W}(t+1))| \\
&\leq |f^*(x) - f(x; \mathbf{W}(0))| + |f(x; \mathbf{W}(0)) - f(x; \mathbf{W}(t+1))| \\
&= |\Delta_0(x)| + \left| \frac{1}{\sqrt{m}} a^\top \left( \sigma(\mathbf{W}^{(L)}(t+1) o_{t+1}^{(L-1)}(x)) - \sigma(\mathbf{W}^{(L)}(0) o_0^{(L-1)}(x)) \right) \right| \\
&\stackrel{(a)}{\leq} R_0 + \left\| \mathbf{W}^{(L)}(t+1) o_{t+1}^{(L-1)}(x) - \mathbf{W}^{(L)}(0) o_0^{(L-1)}(x) \right\|_2 \left\| \frac{1}{\sqrt{m}} a \right\|_2 \\
&\stackrel{(b)}{\leq} R_0 + \left\| \mathbf{E}^{(L)}(t+1) o_{t+1}^{(L-1)}(x) \right\|_2 + \left\| \mathbf{W}^{(L)}(0) \left( o_{t+1}^{(L-1)}(x) - o_0^{(L-1)}(x) \right) \right\|_2 \\
&\stackrel{(c)}{\leq} R_0 + \sum_{\ell=0}^{L-1} \left\| \mathbf{E}^{(L-\ell)}(t+1) o_t^{(L-\ell-1)}(x) \right\|_2
\end{aligned}$$

where (a) holds by Cauchy-Schwartz inequality under  $\mathcal{E}_3$  and the fact that ReLU is 1-Lipschitz, (b) holds by the triangle inequality and (c) holds by recursively decomposing  $o_{t+1}^{(L-1)}(x) - o_0^{(L-1)}(x)$ .

Plugging (3.61) and the assumption that

$$\left\| \mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(0) \right\|_2 \leq C^{L-1} \sum_{s=0}^t \eta_s (R_s + \gamma)$$

for any  $\ell$  in the above displayed equation, we have for any  $x$ ,

$$\begin{aligned}
|\Delta_{t+1}(x)| &\leq R_0 + \sum_{\ell=1}^L C^{L-1} \left\| \mathbf{E}^{(\ell)}(t+1) \right\|_2 \\
&\leq R_0 + LC^{2L-2} \sum_{s=0}^t \eta_s (R_s + \gamma) \\
&= R_{t+1},
\end{aligned}$$

where the last equality holds by the definition of  $R_{t+1}$ .

### 3.2.3 Proof of Lemma 2.6.3

Denote  $O_t^{(\ell)}(x) \triangleq \left\{ i : \mathbf{1}_{\{\langle w_i^{(\ell)}(t), o_t^{(\ell)}(x) \rangle \geq 0\}} - \mathbf{1}_{\{\langle w_i^{(\ell)}(0), o_0^{(\ell)}(x) \rangle \geq 0\}} \neq 0 \right\}$ . Therefore, we have  $S_t^{(\ell)}(x) = |O_t^{(\ell)}(x)|$ .

Note that if any neuron at layer  $\ell$  has a sign flip, then it has either a small output value at initialization or a larger deviation than the initial output value. As such, we define

$$B^{(\ell)}(x) \triangleq \left\{ i : |\langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-1/3} m^{-1/9} \right\}, \quad (3.63)$$

as the set of neurons with small output values at initialization. Then

$$\sup_x S_t^{(\ell)}(x) \leq \sup_x |B^{(\ell)}(x)| + \sup_x \left| O_t^{(\ell)}(x) \cap [B^{(\ell)}(x)]^c \right|. \quad (3.64)$$

It remains to bound both  $\sup_x |B^{(\ell)}(x)|$  and  $\sup_x \left| O_t^{(\ell)}(x) \cap [B^{(\ell)}(x)]^c \right|$ .

Define  $\mathcal{E}_4 \triangleq \{ \sup_x |B^{(\ell)}(x)| = O(C^\ell m^{8/9}), \forall 1 \leq \ell \leq L \}$ . It can be shown that  $\mathcal{E}_4$  occurs with high probability. The proof is deferred to the end.

**Step 1, bounding  $\sup_x S_t^{(\ell)}(x)$ :** Throughout Step 1, we assume  $\mathcal{E}_0 \cap \mathcal{E}_3 \cap \mathcal{E}_4$  and all conclusions of Lemma 2.6.2 hold.

Fix arbitrary  $\ell$ . We first bound the deviation of the output value:

$$\sup_x \left\| \mathbf{W}^{(\ell)}(t) o_t^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(0) o_0^{(\ell-1)}(x) \right\|_2.$$

From (3.47), under  $\mathcal{E}_0$ , we have  $\sup_x \left\| o_t^{(\ell)}(x) \right\|_2 \leq C^\ell$ . Thus, by the triangle inequality,

we have

$$\begin{aligned}
& \sup_x \left\| \mathbf{W}^{(\ell)}(t) o_t^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(0) o_0^{(\ell-1)}(x) \right\|_2 \\
& \leq \sup_x \left\| \left( \mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0) \right) o_t^{(\ell-1)}(x) \right\|_2 + \sup_x \left\| \mathbf{W}^{(\ell)}(0) \left( o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right) \right\|_2 \\
& \leq C^\ell m^{1/3} + c_0 \sqrt{m} \frac{C_1^\ell}{m^{1/6}} \\
& \leq C_2^\ell m^{1/3},
\end{aligned}$$

for some constant  $C_1$  and  $C_2$  where the second inequality holds by (2.45) from Lemma 2.6.2 under  $\mathcal{E}_0$  and (3.60) under  $\mathcal{E}_0 \cap \mathcal{E}_3$ .

For neuron  $i$  in  $[B^{(\ell)}(x)]^c$ , we know  $\left| \langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x) \rangle \right| > \ell^{-1/3} C^{-1/3} m^{-1/9}$ . It follows that

$$\begin{aligned}
\sup_x \left| O_t^{(\ell)}(x) \cap [B^{(\ell)}(x)]^c \right| & \leq \frac{\sup_x \sum_{i=1}^m \left( \langle w_i^{(\ell)}(t), o_t^{(\ell-1)}(x) \rangle - \langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x) \rangle \right)^2}{\ell^{-2/3} C^{-2/3} m^{-2/9}} \\
& = \frac{\sup_x \left\| \mathbf{W}^{(\ell)}(t) o_t^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(0) o_0^{(\ell-1)}(x) \right\|_2^2}{\ell^{-2/3} C^{-2/3} m^{-2/9}} \\
& = O\left(C^\ell m^{8/9}\right).
\end{aligned}$$

Plugging the above displayed equation into (3.64), under  $\mathcal{E}_0 \cap \mathcal{E}_3 \cap \mathcal{E}_4$ , we have

$$\sup_x S_t^{(\ell)}(x) = O\left(C^\ell m^{8/9}\right).$$

**Step 2,  $\mathcal{E}_4$  occurs with high probability:** Here, we prove

$$\mathbb{P}[\mathcal{E}_4] = 1 - L \exp\left(O(d \log m) - \Omega(m^{1/3})\right). \tag{3.65}$$

Define the deviation for any  $1 \leq \ell \leq L$

$$\begin{aligned} \phi_x^{(\ell-1)}(z_1, \dots, z_m) &\equiv \phi_x^{(\ell-1)}\left(z_1, \dots, z_m; \mathbf{W}^{(1)}(0), \dots, \mathbf{W}^{(\ell-1)}(0)\right) \\ &= \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle z_i, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} - \mathbb{E}_w \left[ \mathbf{1}_{\{|\langle w, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} \right] \right| \end{aligned}$$

where  $\mathbb{E}_w[\cdot]$  is the expectation over  $w$ .

We first show  $\phi^{(\ell-1)}(w_1^{(\ell)}, \dots, w_m^{(\ell)})$  concentrates on its mean for any  $x$ . By the triangle inequality, we have

$$\begin{aligned} &\left| \sup_x \phi_x^{(\ell-1)}(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_m) - \sup_x \phi_x^{(\ell-1)}(z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_m) \right| \\ &\leq \frac{1}{m} \left| \mathbf{1}_{\{|\langle z_i, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} - \mathbf{1}_{\{|\langle z'_i, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} \right| \leq \frac{1}{m}. \end{aligned}$$

Thus, by McDiarmid's inequality (Lemma A.1.1), we get

$$\begin{aligned} \mathbb{P} \left[ \sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \dots, w_m^{(\ell)}(0)) \leq \mathbb{E} \left[ \sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \dots, w_m^{(\ell)}(0)) \right] + m^{-1/3} \right] \\ = 1 - \exp\left(-m^{1/3}\right). \end{aligned} \quad (3.66)$$

Now we bound  $\mathbb{E} \left[ \sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \dots, w_m^{(\ell)}(0)) \right]$ . By Lemma A.2.2, we have

$$\mathbb{E} \left[ \sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \dots, w_m^{(\ell)}(0)) \right] \leq C \sqrt{\frac{\text{VC}(\mathcal{Y})}{m}}, \quad (3.67)$$

where  $\mathcal{Y} = \left\{ f_x(w) = \mathbf{1}_{\{|\langle w, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} : x \in \mathbb{S}^{d-1} \right\}$ .

Note that for any  $f \in \mathcal{Y}(\mathbf{W}^{(1)}(0), \dots, \mathbf{W}^{(\ell-1)}(0))$ , we can always find  $g \in \mathcal{W}$  and  $h \in \mathcal{W}'$  such that  $f(w) = g(w)h(w)$  where

$$\mathcal{W}(\mathbf{W}^{(1)}(0), \dots, \mathbf{W}^{(\ell-1)}(0)) = \left\{ g_x(w) = \mathbf{1}_{\{|\langle w, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} : x \in \mathbb{S}^{d-1} \right\},$$

and

$$\mathcal{W}'(\mathbf{W}^{(1)}(0), \dots, \mathbf{W}^{(\ell-1)}(0)) = \left\{ h_x(w) = \mathbf{1}_{\{\langle w, o_0^{(\ell-1)}(x) \rangle \geq -\ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} : x \in \mathbb{S}^{d-1} \right\}.$$

Therefore, by Lemma A.2.1, we have

$$\text{VC}(\mathcal{Y}) = O(\text{VC}(\mathcal{W}) + \text{VC}(\mathcal{W}')).$$

Following the same procedure as bounding  $\text{VC}(\mathcal{F}^{(\ell+1)})$  in the proof of Lemma 3.1.5 from Section 3.1.3, we can get

$$\text{VC}(\mathcal{W}) = \text{VC}(\mathcal{W}') = O(d\ell \log m).$$

As a result,  $\text{VC}(\mathcal{Y}) = O(d\ell \log m)$ . Plugging the bound of  $\text{VC}(\mathcal{Y})$  into (3.67), we get

$$\mathbb{E} \left[ \sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \dots, w_m^{(\ell)}(0)) \right] \leq C \sqrt{\frac{d\ell \log m}{m}} \leq m^{-1/3} \quad (3.68)$$

when  $m$  satisfies (2.41).

Plugging (3.68) into (3.66) and taking union bounds over  $\ell$ , we get with probability at least  $1 - L \exp(-m^{1/3})$ , for all  $x$  and  $\ell$ ,

$$\begin{aligned} & \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle w_i^{(\ell)}, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} \\ & \leq \mathbb{E}_w \left[ \mathbf{1}_{\{|\langle w, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} \right] + 2m^{-1/3}. \end{aligned} \quad (3.69)$$

Next, we bound  $\sup_{x, \ell} \mathbb{E}_w \left[ \mathbf{1}_{\{|\langle w, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} \right]$ . Note that conditioning on

$\{\mathbf{W}^{(k)}(0)\}_{k=1}^L$ ,  $\langle w, o_0^{(\ell-1)}(x) \rangle \sim \mathcal{N}\left(0, \left\|o_0^{(\ell-1)}(x)\right\|_2^2\right)$  as  $w \sim \mathcal{N}(0, \mathbf{I})$ . Therefore,

$$\begin{aligned} & \sup_x \mathbb{E}_w \left[ \mathbf{1}_{\{|\langle w, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} \right] \\ &= \sup_x \mathbb{P}_w \left[ |\langle w, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9} \right] \leq \frac{2\ell^{-1/3} C^{-\ell/3} m^{-1/9}}{\sqrt{2\pi} \left\|o_0^{(\ell-1)}(x)\right\|_2}. \end{aligned} \quad (3.70)$$

Now we bound  $\left\|o_0^{(\ell-1)}(x)\right\|_2$  from below. By Lemma 2.5.1, we have with probability at least  $1 - L \exp(O(d \log m) - \Omega(m^{1/3}))$ , for any  $x$  and  $\ell$ ,

$$\begin{aligned} \left\|o_0^{(\ell-1)}(x)\right\|_2^2 &= \langle o_0^{(\ell-1)}(x), o_0^{(\ell-1)}(x) \rangle \\ &\geq \mathbb{E} \left[ \sigma^2 \left( U^{(\ell-1)}(x) \right) \right] - O\left(\frac{\ell C_1^{2\ell}}{m^{1/3}}\right) \\ &\stackrel{(a)}{=} 2^{-\ell} - O\left(\frac{\ell C_1^{2\ell}}{m^{1/3}}\right) = \Omega(C_2^\ell), \end{aligned} \quad (3.71)$$

for some constant  $C_1$  and  $C_2$  where (a) holds by (2.38) which gives  $\mathbb{E} \left[ \sigma^2(U^{(\ell-1)}(x)) \right] = 2^{-\ell+1}$ .

Plugging (3.71) into (3.70), we have for any  $\ell \leq L$ , with probability at least  $1 - L \exp(O(d \log m) - \Omega(m^{1/3}))$ ,

$$\sup_x \mathbb{E}_w \left[ \mathbf{1}_{\{|\langle w, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} \right] \leq C_3^\ell m^{-1/9}$$

for some constant  $C_3$ . Combining the above inequality with (3.69), we have with probability  $1 - L \exp(O(d \log m) - \Omega(m^{1/3})) - L \exp(-m^{-1/3})$ ,

$$\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{|\langle w_i^{(\ell)}, o_0^{(\ell-1)}(x) \rangle| \leq \ell^{-1/3} C^{-\ell/3} m^{-1/9}\}} \leq C_3^\ell m^{-1/9} + 2m^{-1/3} = O\left(C_3^\ell m^{-1/9}\right).$$

This completes the proof of Step 2.

### 3.2.4 Proof of Lemma 2.6.4

**Step 1, bounding  $\sup_x \left\| z_0^{(k)}(x) \right\|_\infty$ :** We begin with proving (2.48). In particular, we show with probability  $1 - \exp(-\Omega(C^{-L+k+1}m^{1/36}))$ ,

$$\sup_x \left\| z_0^{(k)}(x) \right\|_\infty \leq m^{1/36}. \quad (3.72)$$

Note that

$$\mathbb{P} \left[ \sup_k \left\| z_0^{(k)} \right\|_\infty \leq m^{1/36} \right] = \mathbb{P} \left[ \sup_k \left\| z_0^{(k)} \right\|_\infty \leq m^{1/36} | \mathcal{E}_0 \right] \mathbb{P} [\mathcal{E}_0]. \quad (3.73)$$

Now we show

$$\mathbb{P} \left[ \sup_k \left\| z_0^{(k)} \right\|_\infty \leq m^{1/36} | \mathcal{E}_0 \right] \geq 1 - \exp \left( O(dL \log m) - \Omega(C^{-L+k+1}m^{1/36}) \right).$$

Plugging the above bound on  $\mathbb{P} \left[ \sup_k \left\| z_0^{(k)} \right\|_\infty \leq m^{1/36} | \mathcal{E}_0 \right]$  and (3.44) into (3.73), we complete the proof of step 1.

Throughout the remaining proof of step 1, we condition on  $\{\mathbf{W}^{(k)}(0)\}_{k=1}^L$  such that  $\mathcal{E}_0$  holds.

Denote

$$\mathbf{P}^{(k+1)}(x) \triangleq \frac{1}{\sqrt{m}} \left[ \mathbf{W}^{(k+1)}(0) \right]^\top \mathbf{D}_0^{(k+1)}(x) \cdots \frac{1}{\sqrt{m}} \left[ \mathbf{W}^{(L)}(0) \right]^\top \mathbf{D}_0^{(L)}(x), \quad (3.74)$$

and hence  $z_0^{(k)} = \mathbf{P}^{(k+1)}a$  from (3.45).

Therefore,

$$\left[ z_0^{(k)}(x) \right]_r = \left\langle a, \frac{1}{\sqrt{m}} p_r^{(k+1)}(x) \right\rangle,$$

where  $\left[ z_0^{(k)}(x) \right]_r$  is the  $r$ -th coordinate of  $z_0^{(k)}(x)$  and  $p_r^{(k+1)}(x)$  is the  $r$ -th row of  $\mathbf{P}^{(k+1)}(x)$ .

Under  $\mathcal{E}_0$ , for any  $k$  and  $x$ , we have

$$\left\| p_r^{(k+1)}(x) \right\|_2 = \left\| \mathbf{P}^{(k+1)}(x) e_1 \right\|_2 \leq \left\| \mathbf{P}^{(k+1)}(x) \right\|_2 \leq c_0^{L-k}. \quad (3.75)$$

where  $e_1 = (1, 0, \dots, 0)^\top \in \mathbb{R}^m$ .

Since  $a$  is independent with  $p_r^{(k+1)}(x)$ , by Hoeffding inequality, we have for any fixed  $x \in \mathbb{S}^{d-1}$ ,

$$\begin{aligned} & \mathbb{P} \left[ \left| \left\langle a, \frac{1}{\sqrt{m}} p_r^{(k+1)}(x) \right\rangle \right| > m^{1/36} \left| \left\{ \mathbf{W}^{(k)} \right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds.} \right] \\ & \leq \exp \left( - \frac{m^{1/18}}{2 \left\| \frac{1}{\sqrt{m}} p_r^{(k+1)}(x) \right\|_2^2} \right) \\ & = \exp \left( -\Omega \left( C^{-L+k} m^{1/18} \right) \right). \end{aligned}$$

Taking union bounds over  $r$ , we have for any fixed  $x$ ,

$$\begin{aligned} \mathbb{P} \left[ \left\| z_0^{(k)}(x) \right\|_\infty > m^{1/36} \left| \left\{ \mathbf{W}^{(k)}(0) \right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds.} \right] & \leq m \exp \left( -\Omega \left( C^{-L+k} m^{1/18} \right) \right) \\ & = \exp \left( \log m - \Omega \left( C^{-L+k} m^{1/18} \right) \right). \end{aligned}$$

Thus, we have

$$\begin{aligned} & \mathbb{P} \left[ \sup_x \left\| z_0^{(k)}(x) \right\|_\infty > m^{1/36} \left| \left\{ \mathbf{W}^{(k)}(0) \right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds.} \right] \\ & = \mathbb{P} \left[ \sup_{\mathbf{P} \in \mathcal{P}^{(k+1)}} \left\| \mathbf{P} a \right\|_\infty > m^{1/36} \left| \left\{ \mathbf{W}^{(k)}(0) \right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds.} \right] \\ & \leq |\mathcal{P}^{(k+1)}| \exp \left( \log m - \Omega \left( C^{-L+k} m^{1/18} \right) \right), \end{aligned} \quad (3.76)$$

where  $\mathcal{P}^{(k+1)}(\mathbf{W}^{(1)}(0), \dots, \mathbf{W}^{(L)}(0)) = \{ \mathbf{P}^{(k+1)}(x) : x \in \mathbb{S}^{d-1} \}$  with  $\mathbf{P}^{(k+1)}(x)$  defined in (3.74).

Now we bound  $|\mathcal{P}^{(k+1)}|$ . Recall the definition of  $\mathcal{D}_L$  in Section 2.5. By definition, there is an injective mapping from  $\mathcal{P}^{(k+1)}$  to  $\mathcal{D}_L$ . Therefore, we have  $|\mathcal{P}^{(k+1)}| \leq |\mathcal{D}_L| \leq m^{dL}$ ,

where the last inequality holds by (3.15).

Plugging this bound on  $|\mathcal{P}^{(k+1)}|$  into (3.76), we get

$$\begin{aligned} & \mathbb{P} \left[ \sup_x \left\| z_0^{(k)}(x) \right\|_\infty > m^{1/36} \mid \left\{ \mathbf{W}^{(k)}(0) \right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds} \right] \\ & \leq m^{dL} \exp \left( \log m - \Omega(C^{-L+k} m^{1/18}) \right) \\ & = \exp \left( -\Omega(C^{-L+k} m^{1/18}) \right). \end{aligned}$$

**Step 2, bounding  $\left\| z_t^{(\ell)}(x) - z_0^{(\ell)}(x) \right\|_2$ :** Now we prove the second inequality (2.49) holds with high probability. Assume (3.72), all three conditions in Lemma 2.6.2 and (2.47) hold, which can be guaranteed with probability at least  $1 - \exp(-\Omega(C^{-L+k} m^{1/36}))$  following Lemma 2.6.2, Lemma 2.6.3 and Step 1 above.

Fix arbitrary  $t$ . We will use an inductive argument on layer to prove (2.49) holds for all  $1 \leq \ell \leq L$ .

By definition,  $z_t^{(L)}(x) = a$  for any  $x$ . Therefore, (2.49) holds at  $\ell = L$ .

Now suppose (2.49) holds at some  $\ell + 1$ , we are going to show (2.49) holds at  $\ell$  as well.

Note that

$$z_t^{(\ell)}(x) = \frac{1}{\sqrt{m}} \left[ \mathbf{W}^{(\ell+1)}(t) \right]^\top \mathbf{D}_t^{(\ell+1)}(x) z_t^{(\ell+1)}(x).$$

Similar to (3.46), by the triangle inequality, we have

$$\begin{aligned} \left\| z_t^{(\ell)}(x) - z_0^{(\ell)}(x) \right\|_2 & \leq \left\| \left[ \mathbf{W}^{(\ell+1)}(t) - \mathbf{W}^{(\ell+1)}(0) \right]^\top \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell+1)}(x) z_t^{(\ell+1)}(x) \right\|_2 \\ & \quad + \left\| \frac{1}{\sqrt{m}} \left[ \mathbf{W}^{(\ell+1)}(0) \right]^\top \mathbf{D}_t^{(\ell+1)}(x) \left( z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x) \right) \right\|_2 \\ & \quad + \left\| \frac{1}{\sqrt{m}} \left[ \mathbf{W}^{(\ell)}(0) \right]^\top \left( \mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2. \end{aligned} \quad (3.77)$$

Now we bound the first term on the right hand side of (3.77). By the triangle inequality,

$$\begin{aligned} \left\| z_t^{(\ell+1)}(x) \right\|_2 & \leq \left\| z_0^{(\ell+1)}(x) \right\|_2 + \left\| z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x) \right\|_2 \\ & \leq c_0^{L-\ell} \sqrt{m} + O \left( C^{2L-\ell-1} m^{17/36} \right) = O \left( C_1^{2L-\ell-1} \sqrt{m} \right). \end{aligned} \quad (3.78)$$

for some constant  $C$  and  $C_1$  where the second inequality holds by (3.50) under  $\mathcal{E}_0$  and the inductive hypothesis.

As a result,

$$\begin{aligned}
& \left\| \left[ \mathbf{W}^{(\ell+1)}(t) - \mathbf{W}^{(\ell+1)}(0) \right]^\top \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell+1)}(x) z_t^{(\ell+1)}(x) \right\|_2 \\
& \leq \left\| \mathbf{W}^{(\ell+1)}(t) - \mathbf{W}^{(\ell+1)}(0) \right\|_2 \frac{1}{\sqrt{m}} \left\| z_t^{(\ell+1)}(x) \right\|_2 \\
& \leq C_1^{2L-\ell-1} m^{1/3}, \tag{3.79}
\end{aligned}$$

where the last inequality holds by (2.44) and  $R_t \leq m^{1/3}$  from Lemma 2.6.2, and the above bound of  $\left\| z_t^{(\ell+1)}(x) \right\|_2$ .

To bound the second term, note that under  $\mathcal{E}_0$  and the inductive hypothesis, we have

$$\begin{aligned}
& \left\| \frac{1}{\sqrt{m}} \left[ \mathbf{W}^{(\ell+1)}(0) \right]^\top \mathbf{D}_t^{(\ell+1)}(x) \left( z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x) \right) \right\|_2 \\
& \leq \frac{1}{\sqrt{m}} \left\| \mathbf{W}^{(\ell+1)}(0) \right\|_2 \left\| \mathbf{D}_t^{(\ell+1)}(x) \right\|_2 \left\| z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x) \right\|_2 \\
& = O \left( C^{2L-\ell} m^{17/36} \right). \tag{3.80}
\end{aligned}$$

To bound the last term on the right hand side of (3.77), note that under  $\mathcal{E}_0$ ,

$$\begin{aligned}
& \left\| \frac{1}{\sqrt{m}} \left[ \mathbf{W}^{(\ell)}(0) \right]^\top \left( \mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2 \\
& \leq \frac{1}{\sqrt{m}} \left\| \mathbf{W}^{(\ell)}(0) \right\|_2 \left\| \left( \mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2 \\
& \leq c_0 \left\| \left( \mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2.
\end{aligned}$$

By definition, we have

$$\begin{aligned}
& \left\| \left( \mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2 \\
& \leq \sqrt{\sum_{i=1}^m \left( \mathbf{1}_{\{\langle w_i^{(\ell+1)}(t), o_t^{(\ell)}(x) \rangle \geq 0\}} - \mathbf{1}_{\{\langle w_i^{(\ell+1)}(0), o_0^{(\ell)}(x) \rangle \geq 0\}} \right)^2 \left[ z_0^{(\ell+1)}(x) \right]_i^2} \\
& \leq \left\| z_0^{(\ell)}(x) \right\|_\infty \sqrt{\left\| S_t^{(\ell+1)} \right\|_\infty} \\
& = O\left(C^{2L-\ell-1} m^{17/36}\right), \tag{3.81}
\end{aligned}$$

where the last equality holds by (2.47) and the assumption  $\sup_{x,k} \left\| z_0^{(k)}(x) \right\|_\infty \leq m^{1/36}$ .

Therefore,

$$\left\| \frac{1}{\sqrt{m}} \left[ \mathbf{W}^{(\ell)}(0) \right]^\top \left( \mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2 = O\left(C^{2L-\ell} m^{17/36}\right).$$

Plugging (3.79), (3.80) and the above displayed equation into the right hand side of (3.77), we complete the proof of Step 2.

## 3.3 Proof of lemmas in Section 2.7

### 3.3.1 Proof of Lemma 2.7.2

Recall from (2.21) that  $\Phi^{(\ell)}(x, x') \triangleq \mathbb{E} \left[ \sigma \left( U^{(\ell-1)}(x) \right) \sigma \left( U^{(\ell-1)}(x') \right) \right] q_L^{(\ell)}(x, x')$  where  $U^{(\ell)}(x)$  is defined in (2.15) and  $q_L^{(\ell)}$  is defined in (2.20).

**Step 1,  $\Phi$  is positive semi-definite:** We begin with showing  $\Phi$  is positive semi-definite (PSD). Since  $\Phi = \sum_{\ell=1}^L \Phi^{(\ell)}$ , by (A.2) and (A.3) of Lemma A.3.1, it suffices to show both  $\mathbb{E} \left[ \sigma \left( U^{(\ell)}(x) \right) \sigma \left( U^{(\ell)}(x') \right) \right]$  and  $q_L^{(\ell)}(x, x')$  are PSD kernels.

Denote  $G(x, x') \triangleq \mathbb{E} \left[ \sigma \left( U^{(\ell)}(x) \right) \sigma \left( U^{(\ell)}(x') \right) \right]$ . Here, we apply Lemma A.3.2 to prove

$G(x, x')$  is PSD. To begin with, we show

$$\mathbb{E} \left[ \sigma^2(U^{(\ell)}(x)) \sigma^2(U^{(\ell)}(x')) \right] < \infty. \quad (3.82)$$

By definition (2.15), we have

$$\mathbb{E} \left[ \left( U^{(\ell)}(x) \right)^2 \right] = \left[ \Sigma^{(\ell-1)} \right]_{11} \leq \mathbb{E} \left[ \left( U^{(\ell-1)}(x) \right)^2 \right]$$

where the last inequality holds since  $\sigma^2(U^{(\ell-1)}(x)) \leq (U^{(\ell-1)}(x))^2$ .

Since for any  $x \in \mathbb{S}^{d-1}$ ,  $\mathbb{E} \left[ (U^{(0)}(x))^2 \right] = \|x\|_2^2 \leq 1$ , we get

$$\mathbb{E} \left[ \left( U^{(\ell)}(x) \right)^2 \right] \leq 1, \forall \ell.$$

By Cauchy-Schwartz inequality, we have

$$\mathbb{E} \left[ U^{(\ell)}(x) U^{(\ell)}(x') \right] \leq \sqrt{\mathbb{E} \left[ \left( U^{(\ell)}(x) \right)^2 \right] \mathbb{E} \left[ \left( U^{(\ell)}(x') \right)^2 \right]} \leq 1.$$

Thus, for any  $x, y \in \mathbb{S}^{d-1}$ ,

$$\begin{aligned} \mathbb{E} \left[ \sigma^2(U^{(\ell)}(x)) \sigma^2(U^{(\ell)}(y)) \right] &\leq \mathbb{E} \left[ \left( U^{(\ell)}(x) U^{(\ell)}(y) \right)^2 \right] \\ &= \mathbb{E} \left[ \left( U^{(\ell)}(x) \right)^2 \right] \mathbb{E} \left[ \left( U^{(\ell)}(y) \right)^2 \right] + 2 \mathbb{E} \left[ U^{(\ell)}(x) U^{(\ell)}(y) \right] \\ &\leq 3, \end{aligned}$$

where the equality holds by Isserlis' Theorem [Iss18].

By Cauchy-Schwartz inequality, for any  $g \in L_2(\mu)$ , we get

$$\begin{aligned} &\int \int \mathbb{E} \left[ \left| g(x) \sigma(U^{(\ell)}(x)) \sigma(U^{(\ell)}(y)) g(y) \right| \right] d\mu(x) d\mu(y) \\ &\leq \int \int g^2(x) g^2(y) d\mu(x) d\mu(y) \int \int \mathbb{E} \left[ \left( \sigma(U^{(\ell)}(x)) \sigma(U^{(\ell)}(y)) \right)^2 \right] d\mu(x) d\mu(y) < \infty, \end{aligned}$$

where the last inequality holds by (3.82) and the fact that  $g \in L_2(\mu)$ .

As a result, by Fubini Theorem, we have

$$\begin{aligned}
& \int \int g(x)G(x, y)g(y)d\mu(x)d\mu(y) \\
&= \int \int \mathbb{E} \left[ g(x)\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(y))g(y) \right] d\mu(x)d\mu(y) \\
&= \mathbb{E} \left[ \int \int g(x)\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(y))g(y)d\mu(x)d\mu(y) \right] \\
&= \mathbb{E} \left[ \left( \int g(x)\sigma(U^{(\ell)}(x))d\mu(x) \right)^2 \right] \geq 0.
\end{aligned}$$

By Lemma A.3.2, we get  $G(x, x')$  is PSD.

Now we show  $q_L^{(\ell)}(x, x')$  is PSD by induction. By definition, we know

$$q_L^{(\ell)}(x, x') = \prod_{k=\ell}^L \frac{\pi - \arccos \rho^{(k)}(x, x')}{2\pi}.$$

Following (A.3) of Lemma A.3.1, it remains to show

$$F^{(k)}(x, x') \triangleq \frac{\pi - \arccos \rho^{(k)}(x, x')}{2\pi}$$

is PSD for any  $k$  where  $\rho^{(k)}(x, x') = \frac{\mathbb{E}[\sigma(U^{(k)}(x))\sigma(U^{(k)}(x'))]}{\sqrt{\mathbb{E}[\sigma^2(U^{(k)}(x))]} \sqrt{\mathbb{E}[\sigma^2(U^{(k)}(x'))]}}$  is defined in (3.25).

Note that we have shown the numerator of  $\rho^{(k)}(x, x')$  is PSD. From (2.38), we know the denominator of  $\rho^{(k)}(x, x')$  is some constant independent of  $x$  and  $x'$ . Therefore,  $\rho^{(k)}(x, x')$  is PSD and hence we have  $\rho^{(k)}(x, x') = \langle \phi(x), \phi(x') \rangle$  for some function  $\phi$ .

Since  $\frac{\pi - \arccos(\langle x, x' \rangle)}{2\pi}$  is PSD [CS09], by (A.4) of Lemma A.3.1, we get  $F^{(k)}$  is PSD.

**Step 2,**  $\|\Phi\|_2 \leq \|\Phi\|_\infty \leq \frac{L}{2}$  The inequality  $\|\Phi\|_2 \leq \|\Phi\|_\infty$  follows from Lemma A.3.4.

To bound  $\|\Phi\|_\infty$ , we follow (2.36) and (2.35) and get  $\|\Phi^{(\ell)}\|_\infty \leq \frac{1}{2}$  for all  $\ell$ .

Since  $\Phi = \sum_{\ell=1}^L \Phi^{(\ell)}$ , we have  $\|\Phi\|_\infty \leq \frac{L}{2}$ .

**Step 3, bounding  $\|\mathbf{K}_t\|_2$  and  $\|\mathbf{Q}_t\|_2$**  By definition, the eigenvalues of  $\mathbf{K}_t$  equals  $1 - \eta_t \lambda_i, i = 1, 2, \dots$  where  $\lambda_i$  is the  $i$ -th largest eigenvalue of  $\Phi$ . Since  $0 \leq \lambda_i \leq \frac{L}{2}$  for all  $i$ ,

with  $\eta_t \leq \frac{2}{L}$ , we have  $0 \leq 1 - \eta_t \lambda_i \leq 1$  for all  $i$ . This shows  $\|\mathbf{K}_t\|_2 \leq 1$  and  $\mathbf{K}_t$  is PSD.

Similarly, we bound  $\|\mathbf{Q}_t\|_2$ . Following Theorem 1 and Proposition 2.6.1, by the triangle inequality, with probability at least  $1 - \exp(-\Omega(C^{-L}m^{1/36}))$ ,

$$\|H_t - \Phi\|_\infty \leq \|H_t - H_0\|_\infty + \|H_0 - \Phi\|_\infty = O\left(\frac{C^L}{m^{1/36}}\right).$$

Therefore, for  $m = \exp(\Omega(L))$  which is guaranteed by (2.41), we have

$$\|\mathbf{H}_t\|_2 \leq \|H_t\|_\infty \leq \|H_t - \Phi\|_\infty + \|\Phi\|_\infty \leq \frac{2L}{3} \quad (3.83)$$

for all  $t$ . By the definition of  $H_t$  in (2.1), we know  $\mathbf{H}_t$  is PSD. As a result, we have

$$0 \leq \lambda_i(\mathbf{H}_t) \leq \frac{2L}{3}.$$

For  $\eta_t \leq \frac{3}{2L}$ , we have  $0 \leq 1 - \eta_t \lambda_i(\mathbf{H}_t) \leq 1, \forall i$ , where  $\lambda_i(\mathbf{H}_t)$  is the  $i$ -th largest eigenvalue of  $\mathbf{H}_t$ . This shows  $\|\mathbf{Q}_t\|_2 \leq 1$  and  $\mathbf{Q}_t$  is PSD.

### 3.3.2 Proof of Lemma 2.7.3

Recall that

$$\epsilon_t(x, x') = f(x; \mathbf{W}(t)) - f(x; \mathbf{W}(t+1)) + \eta_t H_t(x, x') (f^*(x') + u_t - f(x'; \mathbf{W}(t))). \quad (3.84)$$

Here we provide a lower bound of  $\epsilon_t$ . The upper bound of  $\epsilon_t$  can be obtained analogously.

The proof consists of three steps. Firstly, we study the evolution of the prediction value  $f(x; \mathbf{W}(t+1)) - f(x; \mathbf{W}(t))$ . Since the change of the prediction value is driven by the update of weight  $\mathbf{W}(t)$  in (2.10), intuitively we have

$$\begin{aligned} f(x; \mathbf{W}(t+1)) - f(x; \mathbf{W}(t)) &\approx \sum_{\ell=1}^L \left\langle \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}, \mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t) \right\rangle \\ &= \eta_t (\Delta_t(X_t) + u_t) H_t(x, X_t). \end{aligned}$$

To justify the above approximation, we first show

$$f(x; \mathbf{W}(t+1)) - f(x; \mathbf{W}(t)) \leq \sum_{\ell=1}^L \Delta_{\mathbf{W}^{(\ell)}(t)}(x) + \sum_{\ell=1}^{L-1} \mathfrak{A}_t^{(\ell)}(x),$$

for some  $\Delta_{\mathbf{W}^{(\ell)}(t)}(x)$  and  $\mathfrak{A}_t^{(\ell)}(x)$  defined in (3.87) and (3.91).

Then we prove that for each  $\ell$ ,

$$\Delta_{\mathbf{W}^{(\ell)}(t)}(x) = \eta_t (\Delta_t(X_t) + u_t) \left( H_t^{(\ell)}(x, X_t) + \mathfrak{B}_t^{(\ell)}(x, X_t) + \mathfrak{R}_t^{(\ell)}(x, X_t) \right) \quad (3.85)$$

where  $\mathfrak{B}_t^{(\ell)}(x, X_t)$  and  $\mathfrak{R}_t^{(\ell)}(x, X_t)$  are some error terms defined in (3.96) and (3.97).

Following the definition of  $\epsilon_t$ , we have

$$\begin{aligned} \epsilon_t(x) &= f(x; \mathbf{W}(t)) - f(x; \mathbf{W}(t+1)) + \eta_t H_t(x, X_t) (\Delta_t(X_t) + u_t) \\ &\geq -\eta_t (\Delta_t(X_t) + u_t) \sum_{\ell=1}^L \left( \mathfrak{B}_t^{(\ell)}(x, X_t) + \mathfrak{R}_t^{(\ell)}(x, X_t) \right) - \sum_{r=1}^{L-1} \mathfrak{A}_t^{(r)}. \end{aligned} \quad (3.86)$$

To bound  $\epsilon_t$ , it suffices to bound  $\mathfrak{A}_t^{(\ell)}$ ,  $\mathfrak{B}_t^{(\ell)}$  and  $\mathfrak{R}_t^{(\ell)}$ . In short, we show these terms depend on either the change of output  $o_{t+1}^{(\ell)} - o_t^{(\ell)}$  or the change of activation pattern  $\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)$  which are shown to be small with high probability in Section 2.6.

### Analysis of the evolution of $f(x; \mathbf{W}(t))$

For all  $0 \leq \ell \leq L$ , define

$$\mathbf{Z}_t^{(\ell)+}(x) \triangleq \text{diag} \left\{ \mathbf{1} \left\{ [z_t^{(\ell)}(x)]_i \geq 0 \right\} \right\}, \quad \text{and} \quad \mathbf{Z}_t^{(\ell)-}(x) \triangleq \text{diag} \left\{ \mathbf{1} \left\{ [z_t^{(\ell)}(x)]_i < 0 \right\} \right\}.$$

When  $\ell = L$ , since  $z_t^{(L)}(x) = a$  does not change over time,

$$\mathbf{Z}_t^{(L)+}(x) = \text{diag} \{a_i = 1\}, \quad \mathbf{Z}_t^{(L)-}(x) = \text{diag} \{a_i = -1\}, \quad \forall t.$$

Denote

$$\begin{aligned}\Delta_{\mathbf{W}^{(\ell)}(t)}(x) &\triangleq \left[ z_t^{(\ell)}(x) \right]^\top \left[ \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right] \\ &\quad \frac{1}{\sqrt{m}} \left( \mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t) \right) o_{t+1}^{(\ell-1)}(x),\end{aligned}\tag{3.87}$$

and

$$\begin{aligned}\Delta_{o_t^{(\ell)}}(x) &\triangleq \left[ z_t^{(\ell+1)}(x) \right]^\top \left[ \mathbf{Z}_t^{(\ell+1)+}(x) \mathbf{D}_{t+1}^{(\ell+1)}(x) + \mathbf{Z}_t^{(\ell+1)-}(x) \mathbf{D}_t^{(\ell+1)}(x) \right] \\ &\quad \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \left( o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right).\end{aligned}$$

Intuitively,  $\Delta_{\mathbf{W}^{(\ell)}(t)}(x)$  and  $\Delta_{o_t^{(\ell)}}(x)$  capture the change of prediction value  $f(x; \mathbf{W}(t))$  by the change of  $\mathbf{W}^{(\ell)}(t)$  and  $o_t^{(\ell)}(x)$ , respectively.

Note that for any vector  $p, b, e \in \mathbb{R}^m$ , we have

$$\begin{aligned}p^\top (\sigma(b) - \sigma(e)) &= \sum_{i:p_i \geq 0} p_i (\sigma(b_i) - \sigma(e_i)) + \sum_{i:p_i < 0} (-p_i) (\sigma(e_i) - \sigma(b_i)) \\ &\leq \sum_{i:p_i \geq 0} p_i \mathbf{1}_{\{b_i \geq 0\}} (b_i - e_i) + \sum_{i:p_i < 0} (-p_i) \mathbf{1}_{\{e_i \geq 0\}} (e_i - b_i),\end{aligned}\tag{3.88}$$

where the last inequality holds by the fact that  $\sigma(y) - \sigma(x) \leq \mathbf{1}_{\{y \geq 0\}} (y - x)$ .

Therefore, we have

$$\begin{aligned}&f(x; \mathbf{W}(t+1)) - f(x; \mathbf{W}(t)) \\ &= \frac{1}{\sqrt{m}} a^\top \left( \sigma(\mathbf{W}^{(L)}(t+1) o_{t+1}^{(L-1)}(x)) - \sigma(\mathbf{W}^{(L)}(t) o_t^{(L-1)}(x)) \right) \\ &\leq \frac{1}{\sqrt{m}} a^\top \left( \mathbf{Z}_t^{(L)+} \mathbf{D}_{t+1}^{(L)}(x) + \mathbf{Z}_t^{(L)-} \mathbf{D}_t^{(L)}(x) \right) \left( \mathbf{W}^{(L)}(t+1) o_{t+1}^{(L-1)}(x) - \mathbf{W}^{(L)}(t) o_t^{(L-1)}(x) \right), \\ &= \Delta_{\mathbf{W}^{(L)}(t)}(x) + \Delta_{o_t^{(L-1)}}(x),\end{aligned}\tag{3.89}$$

where the last equality holds since

$$\begin{aligned} & \mathbf{W}^{(L)}(t+1)o_{t+1}^{(L-1)}(x) - \mathbf{W}^{(L)}(t)o_t^{(L-1)}(x) \\ &= \left( \mathbf{W}^{(L)}(t+1) - \mathbf{W}^{(L)}(t) \right) o_{t+1}^{(L-1)}(x) + \mathbf{W}^{(L)}(t) \left( o_{t+1}^{(L-1)}(x) - o_t^{(L-1)}(x) \right), \end{aligned}$$

$$\begin{aligned} & \Delta_{\mathbf{W}^{(L)}(t)}(x) \\ &= a^\top \left[ \mathbf{Z}_t^{(L)+}(x) \mathbf{D}_{t+1}^{(L)}(x) + \mathbf{Z}_t^{(L)-}(x) \mathbf{D}_t^{(L)}(x) \right] \frac{1}{\sqrt{m}} \left( \mathbf{W}^{(L)}(t+1) - \mathbf{W}^{(L)}(t) \right) o_{t+1}^{(L-1)}(x), \end{aligned}$$

and

$$\Delta_{o_t^{(L-1)}}(x) = a^\top \left[ \mathbf{Z}_t^{(L)+}(x) \mathbf{D}_{t+1}^{(L)}(x) + \mathbf{Z}_t^{(L)-}(x) \mathbf{D}_t^{(L)}(x) \right] \frac{1}{\sqrt{m}} \mathbf{W}^{(L)}(t) \left( o_{t+1}^{(L-1)}(x) - o_t^{(L-1)}(x) \right).$$

Intuitively, since the change of  $o_t^{(\ell)}$  comes from the update of  $\mathbf{W}^{(\ell)}(t)$  and  $o_t^{(\ell-1)}$ , we can obtain a recursive relation of  $\Delta_{o_t^{(\ell)}}$ . In particular, we show that for all  $\ell$ ,

$$\Delta_{o_t^{(\ell)}}(x) \leq \Delta_{o_t^{(\ell-1)}}(x) + \Delta_{\mathbf{W}^{(\ell)}(t)}(x) + \mathfrak{A}_t^{(\ell)}(x), \quad (3.90)$$

where

$$\begin{aligned} & \mathfrak{A}_t^{(\ell)}(x) \\ & \triangleq \left[ z_t^{(\ell+1)}(x) \right]^\top \mathbf{Z}_t^{(\ell+1)+}(x) \left[ \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right] \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \left( o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right). \quad (3.91) \end{aligned}$$

Note that  $\mathbf{Z}_t^{(\ell)+}(x) + \mathbf{Z}_t^{(\ell)-}(x) = \mathbf{I}$ . Therefore, for any  $1 \leq \ell \leq L$ ,

$$\mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) = \mathbf{D}_t^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)+}(x) \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right). \quad (3.92)$$

Thus, we have

$$\begin{aligned}
\Delta_{o_t^{(\ell)}}(x) &= \left[ z_t^{(\ell+1)}(x) \right]^\top \left[ \mathbf{Z}_t^{(\ell+1)+}(x) \mathbf{D}_{t+1}^{(\ell+1)}(x) + \mathbf{Z}_t^{(\ell+1)-}(x) \mathbf{D}_t^{(\ell+1)}(x) \right] \times \\
&\quad \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \left( o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right) \\
&= \underbrace{\left[ z_t^{(\ell+1)}(x) \right]^\top \mathbf{D}_t^{(\ell+1)}(x) \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \left( o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right)}_{\text{(I)}} \\
&\quad + \underbrace{\left[ z_t^{(\ell+1)}(x) \right]^\top \mathbf{Z}_t^{(\ell+1)+}(x) \left[ \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right] \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \left( o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right)}_{\mathfrak{A}_t^{(\ell)}(x)}.
\end{aligned}$$

From (2.50), we know  $\left[ z_t^{(\ell)}(x) \right]^\top = \left[ z_t^{(\ell+1)}(x) \right]^\top \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell+1)}(x) \mathbf{W}^{(\ell+1)}(t)$ . Thus, we have

$$\begin{aligned}
\text{(I)} &= \left[ z_t^{(\ell)}(x) \right]^\top \left( o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right) \\
&= \frac{1}{\sqrt{m}} \left[ z_t^{(\ell)}(x) \right]^\top \left( \sigma \left( \mathbf{W}^{(\ell)}(t+1) o_{t+1}^{(\ell-1)}(x) \right) - \sigma \left( \mathbf{W}^{(\ell)}(t) o_t^{(\ell-1)}(x) \right) \right) \\
&\stackrel{\text{(i)}}{\leq} \left[ z_t^{(\ell)}(x) \right]^\top \left( \mathbf{Z}^{(\ell)+} \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}^{(\ell)-} \mathbf{D}_t^{(\ell)}(x) \right) \times \\
&\quad \frac{1}{\sqrt{m}} \left( \mathbf{W}^{(\ell)}(t+1) o_{t+1}^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(t) o_t^{(\ell-1)}(x) \right) \\
&= \left[ z_t^{(\ell)}(x) \right]^\top \left( \mathbf{Z}^{(\ell)+} \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}^{(\ell)-} \mathbf{D}_t^{(\ell)}(x) \right) \frac{1}{\sqrt{m}} \left\{ \left( \mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t) \right) o_{t+1}^{(\ell-1)}(x) \right. \\
&\quad \left. + \mathbf{W}^{(\ell)}(t) \left( o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x) \right) \right\} \\
&= \Delta_{\mathbf{W}^{(\ell)}(t)}(x) + \Delta_{o_t^{(\ell-1)}}(x),
\end{aligned}$$

where (i) holds by (3.88) and the last equality holds by the definition of  $\Delta_{\mathbf{W}^{(\ell)}(t)}$  and  $\Delta_{o_t^{(\ell-1)}}$ .

Hence, we get (3.90).

Recursively plugging (3.90) into the right hand side of (3.89), we get

$$f(x; \mathbf{W}(t+1)) - f(x; \mathbf{W}(t)) \leq \sum_{\ell=1}^L \Delta_{\mathbf{W}^{(\ell)}(t)}(x) + \sum_{\ell=1}^{L-1} \mathfrak{A}_t^{(\ell)}(x). \quad (3.93)$$

## Decomposing $\Delta_{\mathbf{W}^{(\ell)}(t)}(x)$

Here we prove (3.85). Plugging (2.10) into (3.87) to replace  $\mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t)$ , we have

$$\begin{aligned}
& \Delta_{\mathbf{W}^{(\ell)}(t)} \\
&= \frac{1}{m} \left[ z_t^{(\ell)}(x) \right]^\top \left[ \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right] \langle o_{t+1}^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle \\
& \quad \eta_t (\Delta_t(X_t) + u_t) \mathbf{D}_t^{(\ell)}(X_t) z_t^{(\ell)}(X_t). \tag{3.94}
\end{aligned}$$

Note that

$$\begin{aligned}
& \left( \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right) \langle o_{t+1}^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle \\
&= \left( \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right) \langle o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x) + o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle \\
&\stackrel{(a)}{=} \left( \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right) \langle o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle \\
& \quad + \mathbf{D}_t^{(\ell)}(x) \langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle + \mathbf{Z}_t^{(\ell)+}(x) \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right) \langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle,
\end{aligned}$$

where (a) holds by (3.92).

Plugging the above equation into (3.94), we have

$$\begin{aligned}
& \Delta_{\mathbf{W}^{(\ell)}(t)} \\
&= \frac{1}{m} \eta_t (\Delta_t(X_t) + u_t) \left\{ \left[ z_t^{(\ell)}(x) \right]^\top \mathbf{D}_t^{(\ell)}(x) \langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle \right. \\
& \quad + \left[ z_t^{(\ell)}(x) \right]^\top \mathbf{Z}_t^{(\ell)+}(x) \left( \mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right) \langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle \\
& \quad \left. + \left[ z_t^{(\ell)}(x) \right]^\top \left( \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right) \langle o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle \right\} \\
& \quad \mathbf{D}_t^{(\ell)}(X_t) z_t^{(\ell)}(X_t) \\
&= \eta_t (\Delta_t(X_t) + u_t) \left( H_t^{(\ell)}(x, X_t) + \mathfrak{B}_t^{(\ell)}(x, X_t) + \mathfrak{R}_t^{(\ell)}(x, X_t) \right), \tag{3.95}
\end{aligned}$$

where

$$H_t^{(\ell)}(x, x') = \frac{1}{m} \left\langle \mathbf{D}_t^{(\ell)}(x) z_t^{(\ell)}(x) \left[ o_t^{(\ell-1)}(x) \right]^\top, \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x') \left[ o_t^{(\ell-1)}(x') \right]^\top \right\rangle$$

from (2.39),

$$\begin{aligned} \mathfrak{B}_t^{(\ell)}(x, X_t) &\triangleq \left[ z_t^{(\ell)}(x) \right]^\top \mathbf{Z}_t^{(\ell)+}(x) \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right) \frac{1}{m} \langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle \\ &\quad \mathbf{D}_t^{(\ell)}(X_t) z_t^{(\ell)}(X_t), \end{aligned} \quad (3.96)$$

and

$$\begin{aligned} \mathfrak{R}_t^{(\ell)}(x, X_t) &\triangleq \left[ z_t^{(\ell)}(x) \right]^\top \left[ \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right] \\ &\quad \frac{1}{m} \langle o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t) \rangle \mathbf{D}_t^{(\ell)}(X_t) z_t^{(\ell)}(X_t). \end{aligned} \quad (3.97)$$

Intuitively,  $\mathfrak{B}_t^{(\ell)}$  captures the error from the change in activation pattern of the  $\ell$ -th hidden layer and  $\mathfrak{R}_t^{(\ell)}$  captures the error from the change of the output  $o_t^{(\ell-1)}$ .

Plugging (3.95) back into (3.93), we have

$$\begin{aligned} &f(x; \mathbf{W}(t+1)) - f(x; \mathbf{W}(t)) \\ &\leq \eta_t (\Delta_t(X_t) + u_t) \sum_{\ell=1}^L \left( H_t^{(\ell)}(x, X_t) + \mathfrak{B}_t^{(\ell)}(x, X_t) + \mathfrak{R}_t^{(\ell)}(x, X_t) \right) + \sum_{\ell=1}^{L-1} \mathfrak{A}_t^{(\ell)}(x). \end{aligned} \quad (3.98)$$

Recall the definition of  $\epsilon_t$  from (3.84). For any  $x \in \mathbb{S}^{d-1}$ , we have

$$\begin{aligned} \epsilon_t(x) &= f(x; \mathbf{W}(t)) - f(x; \mathbf{W}(t+1)) + \eta_t H_t(x, X_t) (\Delta_t(X_t) + u_t) \\ &\geq -\eta_t (\Delta_t(X_t) + u_t) \sum_{\ell=1}^L \left( \mathfrak{B}_t^{(\ell)}(x, X_t) + \mathfrak{R}_t^{(\ell)}(x, X_t) \right) - \sum_{\ell=1}^{L-1} \mathfrak{A}_t^{(\ell)}. \end{aligned}$$

To bound  $\epsilon_t$ , it remains to bound  $\mathfrak{A}_t^{(\ell)}$ ,  $\mathfrak{B}_t^{(\ell)}$  and  $\mathfrak{R}_t^{(\ell)}$ .

Here, we claim that with probability at least  $1 - \exp\left(-\Omega(C_0^{-L} m^{1/36})\right)$  over the weight

$\mathbf{W}(0)$  and the outer weight  $a$  for any sample path  $\{(X_s, y_s)\}_{s=0}^{t-1}$ ,

$$\sup_x |\mathfrak{A}_t^{(\ell)}(x)| = O\left(\frac{\eta_t \ell C_1^L}{m^{1/36}} |\Delta_t(X_t) + u_t|\right), \quad (3.99)$$

$$\sup_{x, x'} |\mathfrak{B}_t^{(\ell)}(x, x')| = O\left(\frac{C_2^{2L}}{m^{1/36}}\right) \quad (3.100)$$

$$\sup_{x, x'} |\mathfrak{R}_t^{(\ell)}(x, x')| = O\left(\frac{C_3^L}{m^{1/6}}\right). \quad (3.101)$$

With the above claims, we have with probability at least  $1 - \exp\left(-\Omega(C_0^{-L} m^{1/36})\right)$ , for any  $x$  and sample path  $\{(X_s, y_s)\}_{s=0}^{t-1}$ ,

$$\epsilon_t(x) \geq -C_5 \eta_t |\Delta_t(X_t) + u_t| \frac{LC_4^{2L}}{m^{1/36}},$$

for some constant  $C_4$  and  $C_5$ .

Analogously, we get with probability at least  $1 - \exp\left(-\Omega(C_0^{-L} m^{1/36})\right)$ ,

$$\epsilon_t(x) \leq C_5 \eta_t |\Delta_t(X_t) + u_t| \frac{LC_4^{2L}}{m^{1/36}}.$$

As a result, we have

$$\begin{aligned} \mathbb{E} [\|\epsilon_t\|_2 \mid \mathbf{W}(0), a] &\leq \sqrt{\mathbb{E} [\|\epsilon_t\|_2^2 \mid \mathbf{W}(0), a]} = \sqrt{\mathbb{E}_{(X_s, u_s)_{s=0}^t} [\mathbb{E}_X [\epsilon_t^2(X) \mid \mathbf{W}(0), a]]} \\ &\leq \sqrt{\mathbb{E}_{(X_s, u_s)_{s=0}^t} \left[ \sup_x \epsilon_t^2(x) \mid \mathbf{W}(0), a \right]} \\ &= O\left(\frac{\eta_t LC^L}{m^{1/36}}\right) \sqrt{\mathbb{E}_{(X_s, u_s)_{s=0}^t} [(\Delta_t(X_t) + u_t)^2 \mid \mathbf{W}(0), a]} = \frac{\eta_t LC^L \sigma_t}{m^{1/36}}, \end{aligned} \quad (3.102)$$

where the last equality holds since

$$\begin{aligned} \mathbb{E}_{(X_s, u_s)_{s=0}^t} [(\Delta_t(X_t) + u_t)^2 \mid \mathbf{W}(0), a] &= \mathbb{E}_{(X_s, u_s)_{s=0}^t} [\Delta_t^2(X_t) \mid \mathbf{W}(0), a] + \mathbb{E} [u_t^2] \\ &= \mathbb{E}_{(X_s, u_s)_{s=0}^{t-1}} [\|\Delta_t\|_2^2 \mid \mathbf{W}(0), a] + \tau^2 = \sigma_t^2. \end{aligned}$$

In the following, we prove (3.99)–(3.101). Throughout the remaining of Section 3.3.2, we assume the conclusions of Lemma 2.6.2–2.6.4 hold which is guaranteed to occur with probability at least  $1 - \exp\left(-\Omega(C_0^{-L} m^{1/36})\right)$ .

### Bounding $\mathfrak{A}_t^{(\ell)}$

Recall the definition of  $\mathfrak{A}_t^{(\ell)}$  from (3.91). Here, we bound  $\sup_x \left| \mathfrak{A}_t^{(\ell)}(x) \right|$ . Fix arbitrary  $x \in \mathbb{S}^{d-1}$ . Note that

$$\begin{aligned}
& \left| \mathfrak{A}_t^{(\ell)}(x) \right| \\
& \leq \underbrace{\left\| \left[ z_t^{(\ell+1)}(x) \right]^\top \mathbf{Z}_t^{(\ell+1)+}(x) \left[ \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right] \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \right\|_2}_{\text{(I)}} \underbrace{\left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2}_{\text{(II)}}.
\end{aligned} \tag{3.103}$$

We first bound (II). Note that the change of  $o_t^{(\ell)}$  comes from the change of  $\mathbf{W}(t)$ . Intuitively, since  $\mathbf{W}(t)$  does not change much by Lemma 2.6.2, we expect that  $o_t$  does not change much. By Lipschitz property of ReLU function and the triangle inequality, we obtain the following layer-wise recursive relation of  $\left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2$ :

$$\begin{aligned}
\left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2 &= \frac{1}{\sqrt{m}} \left\| \sigma\left(\mathbf{W}^{(\ell)}(t+1)o_{t+1}^{(\ell-1)}(x)\right) - \sigma\left(\mathbf{W}^{(\ell)}(t)o_t^{(\ell-1)}(x)\right) \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \left\| \mathbf{W}^{(\ell)}(t+1)o_{t+1}^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(t)o_t^{(\ell-1)}(x) \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \left\| \left( \mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t) \right) \right\|_2 \left\| o_{t+1}^{(\ell-1)}(x) \right\|_2 \\
&\quad + \frac{1}{\sqrt{m}} \left\| \mathbf{W}^{(\ell)}(t+1) \right\|_2 \left\| o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x) \right\|_2.
\end{aligned}$$

Since  $o_{t+1}^{(0)}(x) - o_t^{(0)}(x) = 0$ , by recursively applying the above inequality, we have

$$\begin{aligned}
& \left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2 \\
& \leq \frac{1}{\sqrt{m}} \sum_{s=1}^{\ell} \left\| \prod_{r=s+1}^{\ell} \frac{1}{\sqrt{m}} \mathbf{W}^{(s)}(t+1) \right\|_2 \left\| \mathbf{W}^{(s)}(t+1) - \mathbf{W}^{(s)}(t) \right\|_2 \left\| o_{t+1}^{(s-1)}(x) \right\|_2.
\end{aligned}$$

Plugging (3.58) into the above equation to replace  $\|\mathbf{W}^{(s)}(t+1) - \mathbf{W}^{(s)}(t)\|_2$ , we have

$$\begin{aligned} & \left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2 \\ & \leq \frac{1}{\sqrt{m}} \sum_{s=1}^{\ell} \left\| \left( \prod_{r=s+1}^{\ell} \frac{1}{\sqrt{m}} \mathbf{W}^{(s)}(t+1) \right) \right\|_2 \left\| \eta_t (\Delta_t(X_t) + u_t) \mathbf{V}_{L,t}^{(\ell)}(x) a \left[ o_t^{(\ell-1)}(X_t) \right]^\top \right\|_2 \\ & \quad \left\| o_{t+1}^{(s-1)}(x) \right\|_2, \end{aligned} \quad (3.104)$$

where  $\mathbf{V}_{L,t}^{(\ell)}(x)$  is defined in (2.8).

From (3.59), we have  $\frac{1}{\sqrt{m}} \|\mathbf{W}^{(\ell)}(t+1)\|_2 \leq C$ , and hence,

$$\left\| \mathbf{V}_{L,t}^{(\ell)}(x) \right\|_2 \leq C^{L-\ell} / \sqrt{m}. \quad (3.105)$$

Plugging (3.59), (3.61) and (3.105) into the right hand side of (3.104), we get

$$\left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2 = O \left( \frac{\ell C^{L+\ell} \eta_t}{\sqrt{m}} |\Delta_t(X_t) + u_t| \right). \quad (3.106)$$

Note that although  $X_t$  does not explicitly appear on the left hand side of (3.106), the evolution of  $o_t^{(\ell)}(x)$  depends on  $X_t$  and  $u_t$  through the update of  $\mathbf{W}(t)$ .

Now we bound (I) on the right hand side of (3.103). Note that

$$\begin{aligned} & \left\| \left[ z_t^{(\ell+1)}(x) \right]^\top \mathbf{Z}_t^{(\ell+1)+}(x) \left[ \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right] \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \right\|_2 \\ & \leq \left\| \left[ z_t^{(\ell+1)}(x) \right]^\top \mathbf{Z}_t^{(\ell+1)+}(x) \left[ \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right] \right\|_2 \left\| \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \right\|_2 \\ & \stackrel{(a)}{\leq} C \left\| \left[ z_t^{(\ell+1)}(x) \right]^\top \mathbf{Z}_t^{(\ell+1)+}(x) \left[ \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right] \right\|_2 \\ & \stackrel{(b)}{\leq} C \left\| \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right\|_2 \\ & \leq C \left\| \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right\|_2 + C \left\| \left( \mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right\|_2, \end{aligned} \quad (3.107)$$

where (a) holds by (3.59), (b) holds since  $\mathbf{Z}_t^{(\ell+1)+}$  is a diagonal matrix with diagonal entries 0 or 1 and the last inequality holds by the triangle inequality.

Here we bound

$$\left\| \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right\|_2. \quad (3.108)$$

By Lemma 2.6.3, we know  $\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)$  has very few non-zero diagonal coordinates. However, if  $z_t^{(\ell+1)}(x)$  has large values on those coordinates, (3.108) can still be large. To show such situation does not occur, we crucially decompose the coordinates of  $z_t^{(\ell+1)}(x)$  into  $\mathcal{M}$  and  $\mathcal{M}^c$  where

$$\mathcal{M} = \left\{ i \in [m] : \left| \left[ z_t^{(\ell+1)}(x) \right]_i \right| < 2m^{1/36} \right\}.$$

For coordinate  $i \in \mathcal{M}^c$ , since  $\left| \left[ z_t^{(\ell+1)}(x) \right]_i \right| \geq 2m^{1/36}$  and  $\sup_x \left\| z_0^{(\ell+1)}(x) \right\|_\infty \leq m^{1/36}$ , we know

$$\left| \left[ z_t^{(\ell+1)}(x) \right]_i \right| \leq 2 \left| \left[ z_t^{(\ell+1)}(x) \right]_i - \left[ z_0^{(\ell+1)}(x) \right]_i \right|.$$

Intuitively, the above displayed equation says that for coordinate  $i$  of  $z_t^{(\ell+1)}(x)$  with large absolute value, since the initial value  $\left| \left[ z_0^{(\ell+1)}(x) \right]_i \right|$  is small, the magnitude of  $\left[ z_t^{(\ell+1)}(x) \right]_i$  is of the same order of its deviation from the initial value.

With the bound on  $\left\| z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x) \right\|_2$  in (2.49) from Lemma 2.6.4, we are able to control the contribution of coordinates in  $\mathcal{M}^c$  on (3.108) as follows:

$$\begin{aligned} \sum_{j \in \mathcal{M}^c} \left[ \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right]_j^2 &\leq \sum_{j \in \mathcal{M}^c} \left[ z_t^{(\ell+1)}(x) \right]_j^2 \leq 4 \left\| z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x) \right\|_2^2 \\ &= O(C_1^{4L-2\ell-2} m^{17/18}), \end{aligned} \quad (3.109)$$

for some constant  $C_1$ .

Next, we show the contribution on (3.108) from  $\mathcal{M}$  is small. This is true since all coordinates in  $\mathcal{M}$  have small values and the number of coordinates having nonzero  $\mathbf{D}_{t+1}^{(\ell)}(x)$ –

$\mathbf{D}_0^{(\ell)}(x)$  is small. In particular, we have

$$\begin{aligned}
& \sum_{j \in \mathcal{M}} \left[ \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right]_j^2 \\
& \leq 4m^{1/18} \sum_{j \in \mathcal{M}} \left( \mathbf{1}_{\{\langle w_j^{(\ell)}(t+1), o_t^{(\ell-1)}(x) \rangle \geq 0\}} - \mathbf{1}_{\{\langle w_j^{(\ell)}(0), o_0^{(\ell-1)}(x) \rangle \geq 0\}} \right)^2 \\
& \leq 4m^{1/18} \sup_x S_{t+1}^{(\ell)}(x) = O(C_2^\ell m^{17/18}), \tag{3.110}
\end{aligned}$$

for some constant  $C_2$  where the last equality holds by (2.47) from Lemma 2.6.3.

Combining (3.109) and (3.110), we have

$$\left\| \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right\|_2 = O(C_3^{2L-\ell-1} m^{17/36}) \tag{3.111}$$

for some constant  $C_3$ .

Similarly, we can get  $\left\| \left( \mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right\|_2 = O(C_3^{2L-\ell-1} m^{17/36})$ .

Plugging the above bound on  $\left\| \left( \mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right\|_2$  and (3.111) into (3.107), we get

$$\begin{aligned}
& \left\| \left[ z_t^{(\ell+1)}(x) \right]^\top \mathbf{Z}_t^{(\ell+1)+}(x) \left[ \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right] \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \right\|_2 \\
& = C \left\| \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right) z_t^{(\ell+1)}(x) \right\|_2 \\
& = O(C_4^L m^{17/36}) \tag{3.112}
\end{aligned}$$

for some constant  $C_4$ .

Combining (3.106) and (3.112), we have for any  $x \in \mathbb{S}^{d-1}$ ,

$$|\mathfrak{A}_t^{(\ell)}(x)| = O\left( \frac{\eta_t \ell C^L}{m^{1/36}} |\Delta_t(X_t) + u_t| \right),$$

for some constant  $C$ .

## Bounding $\mathfrak{B}_t^{(\ell)}$ and $\mathfrak{R}_t^{(\ell)}$

As is mentioned in Section 3.3.2,  $\mathfrak{B}_t^{(\ell)}$  captures the error caused by the change of activation pattern. To bound  $\left| \mathfrak{B}_t^{(\ell)}(x, x') \right|$ , we crucially apply (3.112) which bounds  $\left\| \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x) \right) z_t^{(\ell)}(x) \right\|_2$ . To bound  $\mathfrak{R}_t^{(\ell)}$  which captures the error from the change of the output  $o_t^{(\ell-1)}$ , we apply (2.45) from Lemma 2.6.2 which bounds the deviation of  $o_t^{(\ell-1)}$ .

**Bounding  $\left| \mathfrak{B}_t^{(\ell)} \right|$ :** Recall that

$$\begin{aligned} \mathfrak{B}_t^{(\ell)}(x, x') &= \left[ z_t^{(\ell)}(x) \right]^\top \mathbf{Z}_t^{(\ell)+}(x) \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right) \\ &\quad - \frac{1}{m} \langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x') \rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x'). \end{aligned}$$

Fix any  $x$  and  $x' \in \mathbb{S}^{d-1}$ . By Cauchy-Schwartz inequality, we have

$$\begin{aligned} \left| \mathfrak{B}_t^{(\ell)}(x, x') \right| &\leq \left\| \left[ z_t^{(\ell)}(x) \right]^\top \mathbf{Z}_t^{(\ell)+}(x) \left( \mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x) \right) \right\|_2 \\ &\quad \left\| \frac{1}{m} \langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x') \rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x') \right\|_2 \\ &= O \left( C^L m^{17/36} \right) \left\| \frac{1}{m} \langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x') \rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x') \right\|_2 \end{aligned}$$

where the last equality holds by (3.112).

Moreover, applying Cauchy-Schwartz inequality again, we get

$$\begin{aligned} &\left\| \frac{1}{m} \langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x') \rangle \mathbf{D}_t^{(\ell)}(X_t) z_t^{(\ell)}(x') \right\|_2 \\ &\leq \frac{1}{m} \left\| o_t^{(\ell-1)}(x) \right\|_2 \left\| o_t^{(\ell-1)}(x') \right\|_2 \left\| z_t^{(\ell)}(x') \right\|_2 \\ &= O \left( \frac{C_1^L}{\sqrt{m}} \right). \end{aligned}$$

where the last equality holds by (3.61) and (3.78).

As a result, for any  $x, x' \in \mathbb{S}^{d-1}$ , we have

$$\left| \mathfrak{B}_t^{(\ell)}(x, x') \right| \leq O\left(\frac{C_2^{2L}}{m^{1/36}}\right)$$

for some constant  $C_2$ .

**Bounding  $|\mathfrak{A}_t^{(\ell)}|$ :** Recall that

$$\begin{aligned} \mathfrak{A}_t^{(\ell)}(x, x') &= \left[ z_t^{(\ell)}(x) \right]^\top \left[ \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right] \\ &\quad - \frac{1}{m} \langle o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x') \rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x'). \end{aligned}$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned} |\mathfrak{A}_t^{(\ell)}(x, x')| &\leq \left\| \left[ \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right] z_t^{(\ell)}(x) \right\|_2 \\ &\quad \left\| \frac{1}{m} \langle o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x') \rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x') \right\|_2. \end{aligned}$$

By (3.78) we have

$$\left\| \left[ z_t^{(\ell)}(x) \right]^\top \left[ \mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right] \right\|_2 \leq \left\| z_t^{(\ell)}(x) \right\|_2 = O(C^{L-\ell} \sqrt{m}).$$

Further note that

$$\begin{aligned} &\left\| \frac{1}{m} \langle o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x') \rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x') \right\|_2 \\ &\leq \frac{1}{m} \left\| o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x) \right\|_2 \left\| o_t^{(\ell-1)}(x') \right\|_2 \left\| z_t^{(\ell)}(x') \right\|_2 \\ &\leq \frac{C^{2L-1}}{\sqrt{m}} \left( \left\| o_{t+1}^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) - o_t^{(\ell-1)}(x) + o_0^{(\ell-1)}(x) \right\|_2 \right) \\ &\leq \frac{C^{2L-1}}{\sqrt{m}} \left( \left\| o_{t+1}^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right\|_2 + \left\| o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right\|_2 \right) \\ &= O\left(\frac{C^{2L+\ell-2}}{m^{1/2+1/6}}\right). \end{aligned}$$

where the second inequality holds by (3.78) and (3.61), the third one holds by the triangle inequality, and the last equality holds by (2.45) from Lemma 2.6.2.

As a result, for any  $x, x' \in \mathbb{S}^{d-1}$ ,

$$|\mathfrak{R}_t^{(\ell)}(x, x')| = O\left(\frac{C_3^L}{m^{1/6}}\right),$$

for some constant  $C_3$ .

### 3.3.3 Proof of Lemma 2.7.4

Throughout the proof, we assume the conclusions of Lemma 2.7.3 holds. For the ease of presentation, we use  $\mathbb{E}[\cdot]$  to denote the conditional expectation  $\mathbb{E}[\cdot | \mathbf{W}(0), a]$ .

Recall the definition of  $v_t$  from (2.54). We first show

$$\mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s \right\|_2^2 \right] \leq \sum_{s=0}^t \mathbb{E} [\|v_s\|_2^2]. \quad (3.113)$$

For notation simplicity, denote  $F_t$  as the filtration of  $\{X_1, \dots, X_t\}$ . Let  $q_t = \sum_{r=0}^t \prod_{i=r+1}^t \mathbf{Q}_i \circ v_r$  and  $h_t = \mathbf{Q}_t \circ q_{t-1}$ . Thus,  $q_t = v_t + h_t$ . Then

$$\mathbb{E} [\|q_t\|_2^2] = \mathbb{E} [\|v_t + h_t\|_2^2] \stackrel{(a)}{=} \mathbb{E} [\|v_t\|_2^2] + \mathbb{E} [\|h_t\|_2^2] \stackrel{(b)}{\leq} \mathbb{E} [\|v_t\|_2^2] + \mathbb{E} [\|q_{t-1}\|_2^2],$$

where (a) uses the fact that

$$\mathbb{E} [\langle v_t, h_t \rangle] = \mathbb{E} [\mathbb{E} [\langle v_t, h_t \rangle | F_{t-1}]] = \mathbb{E} [\langle \mathbb{E} [v_t | F_{t-1}], h_t \rangle] = 0;$$

(b) follows from  $\|\mathbf{Q}_t\|_2 \leq 1$  by Lemma 2.7.2.

Recursively applying the last displayed equation yields that

$$\mathbb{E} [\|q_t\|_2^2] \leq \sum_{r=0}^t \mathbb{E} [\|v_r\|_2^2].$$

Next, we bound  $\mathbb{E} [\|v_s\|_2^2]$ . Recall from (2.56) that  $\sigma_s^2 = \mathbb{E} [\|\Delta_s\|_2^2] + \tau^2$ .

Note that

$$\begin{aligned} \mathbb{E} [\|v_s\|_2^2] &= \eta_s^2 \mathbb{E} \left[ (H_s(x, X_s) (\Delta_s(X_s) + u_s))^2 - \mathbb{E}_{X_s} [H_s(x, X_s) \Delta_s(X_s)]^2 \right] \\ &= \eta_s^2 \mathbb{E} [H_s^2(x, X_s) (\Delta_s(X_s) + u_s)^2] - \eta_s^2 (\mathbb{E}_{X_s} [H_s(x, X_s) \Delta_s(X_s)])^2 \\ &\leq \eta_s^2 L^2 (\mathbb{E} [\|\Delta_s\|_2^2] + \tau^2) = \eta_s^2 \frac{4L^2}{9} \sigma_s^2, \end{aligned} \tag{3.114}$$

where the inequality holds by (3.83) that gives  $\|H_t\|_2 \leq \|H_t\|_\infty \leq \frac{2L}{3}$ .

Therefore, to control  $\mathbb{E} [\|v_s\|_2^2]$ , we need to bound  $\sigma_t^2$ . We now claim that

$$\sigma_{t+1}^2 \leq \prod_{s=0}^t \left(1 + \frac{\sqrt{44L}}{9} \eta_t\right)^2 \sigma_0^2,$$

when  $m = \Omega(\exp(L^2))$  and  $\eta_t \leq \frac{3}{2L}$  for all  $t$ .

Given the claim, we have

$$\begin{aligned} \eta_r \sigma_r &\leq \frac{\theta}{r+1} \prod_{k=0}^{r-1} \left(1 + \frac{\sqrt{44L\theta}}{9(k+1)}\right) \sigma_0 \\ &\leq \frac{\theta}{r+1} \exp \left( \frac{\sqrt{44L\theta}}{9} (\log(r+1) + 1) \right) \sigma_0 \\ &\leq \theta (r+1)^{\sqrt{44L\theta}/9-1} e^{\sqrt{44L\theta}/9} \sigma_0. \end{aligned} \tag{3.115}$$

Combining (3.115) and (3.114) into (3.113), we have

$$\begin{aligned}
\mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s \right\|_2^2 \right] &\leq L^2 \sum_{s=0}^t \eta_s^2 \sigma_s^2 \\
&\leq L^2 \sum_{r=0}^t \theta^2 (r+1)^{2\sqrt{44}\theta L/9-2} e^{2\sqrt{44}\theta L/9} \sigma_0^2 \\
&\leq L^2 \theta^2 e^{2\sqrt{44}\theta L/9} \sigma_0^2 \left( \frac{1}{1-2\sqrt{44}\theta L/9} + 1 \right) = c_2^2 \sigma_0^2.
\end{aligned}$$

By Cauchy-Schwartz inequality, we have

$$\mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s \right\|_2 \right] \leq \sqrt{\mathbb{E} \left[ \left\| \sum_{s=0}^t \prod_{r=s+1}^t \mathbf{Q}_r \circ v_s \right\|_2^2 \mid \mathbf{W}(0), a \right]} = c_2 \sigma_0,$$

which completes the proof.

Now we prove the claim. Recall  $\Delta_{t+1} = \mathbf{Q}_t \cdot \Delta_t - v_t + \epsilon_t$ . Therefore,

$$\begin{aligned}
\|\Delta_{t+1}\|_2^2 &= \|\mathbf{Q}_t \cdot \Delta_t - v_t + \epsilon_t\|_2^2 \\
&= \|\mathbf{Q}_t \cdot \Delta_t\|_2^2 + \|v_t\|_2^2 + \|\epsilon_t\|_2^2 - 2\langle \mathbf{Q}_t \cdot \Delta_t, v_t \rangle - 2\langle v_t, \epsilon_t \rangle + 2\langle \mathbf{Q}_t \cdot \Delta_t, \epsilon_t \rangle \\
&\leq \|\Delta_t\|_2^2 + \|v_t\|_2^2 + \|\epsilon_t\|_2^2 + 2\|\Delta_t\|_2 \|v_t\|_2 + 2\|v_t\|_2 \|\epsilon_t\|_2 + 2\|\Delta_t\|_2 \|\epsilon_t\|_2,
\end{aligned} \tag{3.116}$$

where the last inequality holds by  $\|\mathbf{Q}_t\|_2 \leq 1$  whenever  $\eta_t \leq 2/L$  and Cauchy-Schwartz inequality.

From (3.102), for  $m$  satisfying (2.41), we can get

$$\mathbb{E} [\|\epsilon_t\|_2^2] \leq O \left( \frac{L^2 C^L \eta_t^2}{m^{1/18}} \sigma_t^2 \right) \leq \frac{L^2 \eta_t^2}{81} \sigma_t^2. \tag{3.117}$$

Taking conditional expectation on both hand sides of (3.116), we have

$$\begin{aligned}
& \sigma_{t+1}^2 \\
& \leq \sigma_t^2 + \frac{4L^2\eta_t^2}{9}\sigma_t^2 + \frac{L^2\eta_t^2}{81}\sigma_t^2 + 2\mathbb{E}[\|\Delta_t\|_2\|v_t\|_2] + 2\mathbb{E}[\|v_t\|_2\|\epsilon_t\|_2] + 2\mathbb{E}[\|\Delta_t\|_2\|\epsilon_t\|_2] \\
& \leq \left(1 + \frac{37L^2\eta_t^2}{81}\right)\sigma_t^2 + 2\sqrt{\mathbb{E}[\|\Delta_t\|_2^2]}\sqrt{\mathbb{E}[\|v_t\|_2^2]} \\
& \quad + 2\sqrt{\mathbb{E}[\|v_t\|_2^2]}\sqrt{\mathbb{E}[\|\epsilon_t\|_2^2]} + 2\sqrt{\mathbb{E}[\|\Delta_t\|_2^2]}\sqrt{\mathbb{E}[\|\epsilon_t\|_2^2]} \\
& \leq \left(1 + \frac{37L^2\eta_t^2}{81}\right)\sigma_t^2 + \frac{2}{9}\eta_t\sigma_t^2 + \frac{2L^2\eta_t^2}{27}\sigma_t^2 + \frac{2L\eta_t}{9}\sigma_t^2 \\
& = \left(1 + \frac{\sqrt{44}L}{9}\eta_t\right)^2\sigma_t^2
\end{aligned}$$

where the first and the third inequalities hold by (3.114) and (3.117) for  $m$  satisfying (2.41) and the second inequality holds by Cauchy-Schwartz inequality.

### 3.4 Proof of Corollary 1

We first show a key intermediate step to prove Corollary 1.

Define the space of homogeneous harmonic polynomials of order  $\ell$  on the sphere as

$$\mathcal{H}_\ell = \left\{ P : \mathbb{S}^{d-1} \rightarrow \mathbb{R} : P(x) = \sum_{|\alpha|=\ell} c_\alpha x^\alpha, \Delta P = 0 \right\}$$

where  $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ ,  $|\alpha| = \sum_{i=1}^d \alpha_i$ ,  $c_\alpha \in \mathbb{R}$  and  $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$  is the Laplacian operator.

Denote for all  $\ell \geq 0$ ,  $\{Y_{\ell,i}\}_{i=1}^{N_\ell}$  as some orthonormal basis of  $\mathcal{H}_\ell$  where  $N_\ell$  is the dimension of  $\mathcal{H}_\ell$ , i.e.,  $\langle Y_{\ell,i}, Y_{\ell,j} \rangle = 0$  for  $i \neq j$ . Moreover, from [DX13, Theorem 1.1.2] for  $\ell \neq \ell'$ ,  $\mathcal{H}_\ell$  and  $\mathcal{H}_{\ell'}$  are orthogonal. Hence,  $\{Y_{\ell,i}\}$  are orthogonal across different  $\ell$  as well.

We now derive in Theorem 3 an expansion for functions with the form  $\mathcal{K}(x, y) =$

$h(\langle x, y \rangle)$ ,  $x, y \in \mathbb{S}^{d-1}$ ,  $d \geq 3$  in terms of  $\{Y_{\ell,i}\}$ ,  $1 \leq i \leq N_\ell$ ,  $\ell \geq 0$ . A similar result is obtained in [SY19] without a full proof. We provide a proof in Section 3.4.1 for completeness.

**Theorem 3.** *Suppose the function  $\mathcal{K}$  has the form  $\mathcal{K}(x, y) = h(\langle x, y \rangle)$  where  $h$  is analytic on  $[-1, 1]$ ,  $x, y \in \mathbb{S}^{d-1}$  and  $d \geq 3$ . Then*

$$\mathcal{K}(x, y) = \sum_{\ell \geq 0} \beta_\ell(h) \sum_{i=1}^{N_\ell} Y_{\ell,i}(x) Y_{\ell,i}(y)$$

where

$$\beta_\ell(h) = \frac{d-2}{2} \sum_{m=0}^{\infty} \frac{h_{\ell+2m}}{2^{\ell+2m} m! \binom{d-2}{2}_{\ell+m+1}} \quad (3.118)$$

with  $h_{\ell+2m}$  is the  $(\ell+2m)$ -th derivative of  $h$  at 0 and  $(\cdot)_n$  is the Pochhammer symbol recursively defined as  $(a)_0 = 1$ ,  $(a)_k = (a+k-1)(a)_{k-1}$  for  $k \geq 1$ .

**Remark 3.4.1.** *The case  $d = 2$  can be analyzed using Fourier analysis. Since this is not of particular interest in our study, we do not provide the analysis here. One can refer to [DX13, Section 1.6] if interested.*

*Proof of Corollary 1.* From [Wan10, Theorem 7.4], we know the polynomial of degree  $\ell^*$  can be projected onto the direct sum of the spaces of homogeneous harmonic polynomials up to degree  $\ell^* + 1$ . Now we claim  $\Phi$  can be expanded in the space of homogeneous harmonic polynomials. With the claim, we have  $\mathcal{R}(f^*, \ell^* + 1) = 0$  which completes the proof.

It remains to prove the claim. Recall the definition of  $\Phi^{(\ell)}$  in (2.21). Here, we show  $\Phi^{(\ell)}(x, x')$  is analytic and can be viewed as a function of  $\langle x, x' \rangle$  only by analyzing  $\mathbb{E} [\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))]$  and  $q_L^{(\ell)}(x, x')$ .

We begin with analyzing  $\mathbb{E} [\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))]$ . By (2.15), we get  $(U^{(\ell)}(x), U^{(\ell)}(x'))$  depends on  $\text{Cov}(\sigma(U^{(\ell)}(x)), \sigma(U^{(\ell)}(x')))$ . Since  $\Sigma^{(0)}$  only depends on  $\langle x, x' \rangle$ , we know

the joint distribution of  $(U^{(1)}(x), U^{(1)}(x'))$  only depends on  $\langle x, x' \rangle$ . Hence,  $\text{Cov}(\sigma(U^{(1)}(x)), \sigma(U^{(1)}(x')))$  only depends on  $\langle x, x' \rangle$ . Following the recursive relationship of  $U^{(\ell)}$ , we get the joint distribution of  $(U^{(\ell)}(x), U^{(\ell)}(x'))$  for all  $\ell \geq 1$  only depends on  $\langle x, x' \rangle$ . Hence,  $\mathbb{E}[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))]$  only depends on  $\langle x, x' \rangle$ . Note that a product of two ReLU functions is analytic. By Fubini Theorem and Leibniz integral rule, we know  $\mathbb{E}[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))]$  is analytic.

Next, we study  $q_L^{(\ell)}(x, x')$  which is defined in (2.19). We have shown the numerator of  $\rho^{(\ell)}(x, x')$  only depends on  $\langle x, x' \rangle$ . By (2.38), we know the denominator of  $\rho^{(k)}(x, x')$  is some constant independent of  $x$  and  $x'$ . Therefore,  $\rho^{(k)}(x, x')$  and hence  $q_L^{(\ell)}(x, x')$  only depends on  $\langle x, x' \rangle$ . Since a composition of analytic functions is analytic and arccos function is analytic, we know  $q_L^{(\ell)}$  is analytic.

Since for any  $\ell$ ,  $\Phi^{(\ell)}$  is analytic and can be viewed as a function of  $\langle x, x' \rangle$ , we know  $\Phi = \sum_{\ell=1}^L \Phi^{(\ell)}$  is also analytic and is a function of  $\langle x, x' \rangle$ .  $\square$

### 3.4.1 Proof of Theorem 3

We begin with a key result that will be used in the proof of Theorem 3.

**Proposition 3.4.1.** *[CI12, Theorem 2, eq (2.1)] Let  $h$  be analytic in  $[-1, 1]$ . Let  $h_n = h^{(n)}(0)$  be  $n$ -th order derivative, then for any  $\alpha > -1, \alpha \neq -\frac{1}{2}$ ,*

$$h(x) = \sum_{n=0}^{\infty} \tilde{h}_n C_n^{\alpha+1/2}(x), \quad x \in [-1, 1] \quad (3.119)$$

where

$$C_n^{\alpha+1/2}(x) = \frac{(2\alpha+1)_n}{n!} \sum_{k=0}^n (-1)^k \binom{n}{k} \frac{(n+2\alpha+1)_k}{(\alpha+1)_k} \left(\frac{1-x}{2}\right)^k,$$

is the Gegenbauer polynomial, and

$$\tilde{h}_n = (\alpha+n+1/2) \sum_{m=0}^{\infty} \frac{h_{n+2m}}{2^{n+2m} m! (\alpha+1/2)_{n+m+1}}, \quad (3.120)$$

with  $h_{n+2m} = h^{(n+2m)}(0)$ , the  $n + 2m$ -th derivative of  $h$  at 0.

**Remark 3.4.2.** Gegenbauer polynomials are orthogonal across different  $n$ , i.e., for  $m \neq n$ ,  $d \geq 3$  and any fixed  $y \in \mathbb{S}^{d-1}$ ,  $\left\langle C_n^{\frac{d-2}{2}}(\langle \cdot, y \rangle), C_m^{\frac{d-2}{2}}(\langle \cdot, y \rangle) \right\rangle_{\mathbb{S}^{d-1}} = 0$ . The proof is based on the orthogonality of  $\mathcal{H}_\ell$ . One can check [DX13, Corollary 2.8] for a detailed proof.

The form of  $\beta_\ell(h)$  in (3.118) depends on the specific function  $h$ . For the ease of presentation, we abbreviate  $\beta_\ell(h)$  as  $\beta_\ell$ . Now we proceed to the proof of Theorem 3.

*Proof of Theorem 3.* From [DX13, eq(2.8)], we know for any  $l \geq 0$ ,

$$\frac{\ell + \lambda}{\lambda} C_\ell^\lambda(\langle x, y \rangle) = \sum_{i=1}^{N_\ell} Y_{\ell,i}(x) Y_{\ell,i}(y) \quad (3.121)$$

where  $\lambda = \frac{d-2}{2}$ ,  $x, y \in \mathbb{S}^{d-1}$ .

Plugging (3.121) into (3.119) and note that  $\alpha + 1/2 = \lambda = \frac{d-2}{2}$ , we get

$$h(\langle x, y \rangle) = \sum_{\ell \geq 0} \tilde{h}_\ell \frac{\lambda}{\ell + \lambda} \sum_{i=1}^{N_\ell} Y_{\ell,i}(x) Y_{\ell,i}(y) = \beta_\ell \sum_{i=1}^{N_\ell} Y_{\ell,i}(x) Y_{\ell,i}(y)$$

where

$$\beta_\ell = \tilde{h}_\ell \frac{\lambda}{\ell + \lambda} = \frac{d-2}{2} \sum_{m=0}^{\infty} \frac{h_{\ell+2m}}{2^{\ell+2m} m! \left(\frac{d-2}{2}\right)_{\ell+m+1}}.$$

□

# Chapter 4

## Robust GD in Federated Learning

### 4.1 Introduction

Federated Learning (FL) is an emerging model training paradigm, wherein the central parameter server (PS) trains a model by communicating with distributed clients while keeping the training data stored locally at the clients [MMR<sup>+</sup>17, KMA<sup>+</sup>21]. While opening up a world of new opportunities for training machine learning models without compromising data privacy, FL faces two core challenges. First, local data distributions across different clients are highly heterogeneous and moreover, the data volume is highly unbalanced. Thus training a single global model for all clients may not work well. Instead, it is more desirable to train a personalized model for each client based on the client’s own dataset as well as the datasets of other clients. Second, when training with thousands of clients, the learner lacks enough administrative power over those external clients and thus the system is susceptible to adversarial attacks. In particular, some external clients may be hacked by a system adversary and behave maliciously. This chapter aims to tackle data heterogeneity and adversarial attack simultaneously.

A popular approach to deal with data heterogeneity and model personalization is clustering. In practice, different clients often fall into different groups according to their characteristics. As such, it is natural to cluster clients into different groups and train a model for each group. However, the underlying clusters are often hidden and thus the central problem is how to estimate the model parameters and the clusters simultaneously. This problem is known as Clustered Federated learning [SMS20].

Significant progress has been made in this direction [MMRS20, XLS<sup>+</sup>21, LLV21, GCYR20, SXY22]. For example, the recent work [SXY22] proposes a two-phase federated learning algorithm to learn clustered personalized models in the context of mixed regression problems and proves the global convergence from any initialization. However, most of the existing work in clustered federated learning is restricted to the failure-free setting without adversarial attacks.

A separate line of work is devoted to the study of the adversarial attack under the Byzantine model[BEMGS17, CSX17, YCKB18, SX19, KHJ21, FGG<sup>+</sup>22]. Various robust gradient aggregators and Byzantine-resilient FL algorithms have been developed. However, the existing literature mostly focuses on the homogeneous setting with IID data distributions and balanced data partition. In fact, data heterogeneity poses significant challenges to the design of robust gradient aggregators, as the local gradients are no longer close to a single population gradient, rendering it hard to distinguish between the statistical perturbation from the Byzantine one. A few more recent works [DD21, KHJ22, AFG<sup>+</sup>23] have extended the results to the non-IID settings by imposing strong assumptions, e.g., the dissimilarity among local gradients is bounded. However, these assumptions often do not hold in practice, for example, the local gradients of clients across different clusters can differ a lot, violating the bounded gradient dissimilarity assumptions. Moreover, all the above work focuses on training a single global model and does not consider model personalization. As we will see in numerical experiments in Section 4.5.1, ignoring the cluster structure and training a single model will suffer significant performance degradation.

To the best of our knowledge, [GHYR19] is the only work in the literature that considers clustered federated learning and Byzantine attack simultaneously. A three-stage algorithm is proposed with provable performance guarantees. However, a number of restricted assumptions are postulated. For example, every client needs to locally compute an initial model estimate that is sufficiently accurate and distributed

as sub-Gaussian. Moreover, the data volume across all clients is assumed to be balanced. These assumptions rarely hold in practice; in fact, the authors of [GHYR19] raise the open question of weakening the sub-Gaussian assumption along with a better initialization scheme.

In this chapter, we design a Byzantine-resilient FL algorithm that can accurately recover the clusters and estimate the model parameters for each cluster. Our algorithm works without any distributional assumptions and draws upon several innovative ideas:

- We leverage the existence of a few anchor clients, which correspond to active clients or clients who are specially recruited by the PS in practice. While the majority of the clients have a limited number of data points, it is reasonable to assume that the anchor clients have a sufficient number of data points so that their local model estimates are relatively accurate. Thus, we utilize the local model estimates at anchor clients as noisy “seeds” to cluster the anchor clients and obtain some coarse estimates of the cluster centers, which are in turn used for classifying all clients. This significantly relaxes the initialization requirement in [GHYR19], which demands the local model estimates are accurate for all clients.
- We design a new robust clustering algorithm that works without distributional assumptions. In particular, we first construct a graph by thresholding the pairwise distances among the local model estimates of anchor clients and then apply a semidefinite programming relaxation (SDP) to robustly cluster them. The thresholding is critical to suppress the influence of Byzantine clients and the SDP further improves the clustering accuracy and robustness. We provide a deterministic condition in terms of the graph connectivity for the method to accurately cluster non-Byzantine anchor clients, which could be of independent

interest.

- To deal with the unbalancedness of data partition across clients and improve the final estimation accuracy, we employ the idea of batching and geometric median of means. In particular, within each estimated cluster, we randomly partition the clients into a specific number of batches, so that the majority of the batches do not contain any misclassified or Byzantine clients. Thus the geometric median of means of all batches gives a robust aggregator of local gradients, which can then be used to iteratively refine the model estimates via distributed gradient descent.

We show that our algorithm can achieve a misclassification rate on the order of  $\alpha \triangleq K\epsilon/\rho$ , where  $K$  is the number of clusters,  $\rho$  is the proportional size of the smallest cluster, and  $\epsilon$  is an upper bound on the fraction of Byzantine clients at each communication round. Moreover, for both strongly convex and non-convex objectives, we show that our algorithm can estimate the underlying model parameter for each cluster  $k$  with an error on the order of  $\sigma_k^2 \max\{\alpha M_k, 1\} \log(M_k)/N_k$ , where  $M_k$  and  $N_k$  are the number of clients and number of data points in cluster  $k$ , respectively, and  $\sigma_k^2$  is the variance of the local gradients in cluster  $k$ . Our results are applicable even when there are clients with only  $O(1)$  data points, and significantly improve the accuracy of clustering and model estimation in [GHYR19] when the data partition is unbalanced and the fraction of Byzantine clients  $\epsilon$  is not too small.

We also provide numerical studies to validate the performance of our algorithm. We show that when the data partition is highly unbalanced, our algorithm performs much better than the algorithm from [GHYR19] both in clustering and model parameter estimation. In addition, we vary the separation of cluster centers and show that as cluster structure becomes more evident, our algorithm attains much lower loss than the existing Byzantine-resilient algorithms that ignore the cluster structure.

## 4.2 Problem Setup

Consider a federated learning (FL) setting with one central parameter server (PS) and  $M$  distributed clients, where PS can communicate with all clients in synchronous communication rounds. To model data heterogeneity and personalization in FL, following [SMS20] we assume each client falls into one of the  $K$  clusters, that is, each client  $i$  is associated with a hidden cluster label  $z_i \in [K]$ . Denote the true cluster as  $\mathcal{C}_k = \{i : z_i = k\}$  and its proportional size as  $\rho_k = |\mathcal{C}_k|/M$ . Clients with the same cluster are assumed to have the same data distribution. In particular, at any communication round, client  $i \in \mathcal{C}_k$  can draw a batch of  $n_i$  fresh data points from the unknown distribution  $\mathcal{D}_k$ . Here, the batch size  $n_i$  can vary significantly across clients  $i$ , as the local datasets can be highly unbalanced and the clients may have different levels of activeness in practice. Denote the total number of data points within cluster  $k$  as  $N_k = \sum_{i \in \mathcal{C}_k} n_i$ .

Our goal is to estimate the hidden cluster structure and learn the minimizer  $\theta_k$  of the population loss for each cluster, i.e.,

$$\theta_k \in \operatorname{argmin}_{\theta} \{L_k(\theta) \triangleq \mathbb{E}_{X \sim \mathcal{D}_k} [l(\theta; X)]\},$$

where  $l(\theta; x)$  is some function measuring the loss induced by the data point  $x$  under the model parameter  $\theta$ . Throughout the chapter, we assume  $L_k$  is  $\beta$ -Lipschitz and refer to  $\theta_k$  as the center of cluster  $k$ .

In practice, the FL systems are often massive in scale and hence exhibit some unpredictability such as machine crashes/failures, stalled processes, and malicious attacks. To model such unpredictability, we assume the existence of a Byzantine adversary who can control up to  $\epsilon$  fraction of clients in each communication round [CSX17]. The Byzantine adversary is assumed to have complete knowledge of the system in-

cluding all the information of all clients and the programs run in the system at each communication round. Clients suffering Byzantine attacks denoted as Byzantine clients, can lie during the communication round and can collude with each other. The set of Byzantine clients may change across different communication rounds.

Here, we focus on the case of  $\rho \triangleq \min_k \rho_k > \epsilon$ . If  $\rho < \epsilon$ , the Byzantine adversary can corrupt an entire cluster, which makes it impossible to learn the cluster center.

### 4.3 Algorithm and Theoretical Guarantees

Since the underlying clusters are hidden and the communication from the clients to PS is subject to Byzantine attacks, a key challenge is how to *simultaneously and securely* estimate the clusters and the cluster centers. To tackle this challenge, we propose a two-phase method outlined in Algorithm 1. The detailed algorithm descriptions are deferred to the next section.

---

**Algorithm 1** Byzantine-Resilient Clustered Federated Learning

---

**Input:** local estimate  $\{\theta_i^{(0)}\}_{i \in \mathcal{A}}$  from anchor clients

**Output:**  $K$  Clusters  $\{\widehat{\mathcal{C}}_k\}_{k \in [K]}$  and estimations of cluster centers  $\widehat{\theta}_k, k \in [K]$

- 1: PS constructs a graph  $A$  via truncating the pairwise distances among  $\{\theta_i^{(0)}\}_{i \in \mathcal{A}}$  and then use an SDP to cluster the anchor clients into  $K$  groups (Algorithm 2).
  - 2: Applying the Iterative Filtering algorithm within each estimated cluster, PS computes and broadcasts the coarse estimates  $\{\widehat{\theta}_k^{(0)}\}_{k \in [K]}$  of cluster centers;
  - 3: Every client finds the cluster label that produces the minimum local empirical loss at  $\{\widehat{\theta}_k^{(0)}\}_{k \in [K]}$  and sends the estimated cluster label to PS (Algorithm 4);
  - 4: PS and all clients collectively run Byzantine Gradient Descent to refine the estimation of  $\theta_k$  iteratively within each estimated cluster (Algorithm 5).
- 

In phase 1 (Steps 1 and 2 of Algorithm 1), we obtain some coarse estimates of cluster centers by leveraging the existence of a few anchor clients. Anchor clients, denoted by  $\mathcal{A}$ , represent active clients in practice who have access to enough data points so that their local model estimates  $\theta_i^{(0)}$  are relatively accurate. These local model estimates serve as noisy “seeds” and we assume they are given as inputs.

In practice, such local model estimates can be generated by, for example, solving the regularized empirical risk minimization based on the local data. We first design a graph-based semi-definite programming method to robustly cluster the anchor clients, even if at most an  $\epsilon$ -fraction of  $\{\theta_i^{(0)}\}_{i \in \mathcal{A}}$  is corrupted by the Byzantine adversary. Then we apply a standard iterative filtering algorithm [SCV17, Algorithm 1] within each estimated cluster to obtain a coarse estimator of the cluster center despite the presence of Byzantine errors. Note that the use of anchor clients has appeared in the previous work [SXY22]; however, they do not consider SDP clustering and Byzantine clients as we do.

In phase 2 (Steps 3 and 4 of Algorithm 1), we first utilize the coarse estimates to classify all clients. Intuitively, if the coarse estimates are relatively close to the true cluster centers, the local empirical losses evaluated at  $\widehat{\theta}_k^{(0)}$  for most clients in cluster  $k$  are likely to be smaller than their local empirical losses evaluated at  $\widehat{\theta}_j^{(0)}$  for  $j \neq k$ . Thus, we can accurately estimate the cluster label for most clients by finding the one that minimizes the local empirical loss among  $\{\widehat{\theta}_k^{(0)}\}_{k \in [K]}$ . Then we refine the estimate of the cluster center within each obtained cluster. Here, significant challenges arise, as an estimated cluster may contain both misclassified and Byzantine clients and moreover the data volume over clients are highly unbalanced. To resolve these challenges, we crucially exploit the idea of batching and *geometric median of means* [CSX17]. We randomly partition the clients into batches so that more than half of the batches are free of misclassified and Byzantine clients and moreover the data points are more evenly distributed over the batches. Then we use the geometric median of the averaged gradients from each batch to robustly aggregate the gradients and update the estimates iteratively via distributed gradient descent. Note that a similar batching idea has been used in [KHJ22] for variance reduction.

### 4.3.1 Theoretical Guarantees of Phase 1

Recall that  $\mathcal{A}$  denotes the collection of anchor clients. Denote  $\mathcal{A}_g$  as the set of good clients in phase 1 and  $\mathcal{A}_b \triangleq \mathcal{A} \setminus \mathcal{A}_g$  as the collection of clients suffering Byzantine attacks. Let  $m_k = |\mathcal{A}_g \cap \mathcal{C}_k|$  denote the number of good anchor clients in cluster  $k$ . For the ease of presentation, we assume  $\min_k m_k / |\mathcal{A}| \geq \rho$ ; if not, we just define  $\rho = \min_k m_k / |\mathcal{A}|$  and the analysis follows. Throughout the chapter we use  $C$  to denote absolute constant whose value may vary in lines.

Intuitively, the local estimates at anchor clients are expected to be similar within the same cluster and dissimilar across different clusters. However, due to the existence of the Byzantine adversary, an adversarially chosen  $\epsilon$ -fraction of the local estimates may be arbitrarily corrupted. To tolerate these Byzantine errors, in Step 1 PS thresholds the pairwise distances among all the received local model estimates from anchor clients and construct graph  $A$ . In particular, given some appropriate threshold  $\tau$ ,  $\mathbf{A}(i, j) = 1$  if  $\|\theta_i^{(0)} - \theta_j^{(0)}\|_2 < \tau$ ,  $i \neq j$  and  $\mathbf{A}(i, j) = 0$  otherwise. After thresholding, the graph  $\mathbf{A}$  is less sensitive to the Byzantine errors. For the graph-based SDP clustering to work, we need graph  $\mathbf{A}$  to be more densely connected within the clusters. To this end, define the assortativity and disassortivity as

$$a_{\text{in}} = \min_{i \in \mathcal{A}_g} \frac{1}{m_{z_i}} \sum_{j \in \mathcal{A}_g: z_j = z_i} \mathbf{A}(i, j), \quad a_{\text{out}} = \max_{i \in \mathcal{A}_g} \max_{k \neq z_i} \frac{1}{m_k} \sum_{j \in \mathcal{A}_g: z_j = k} \mathbf{A}(i, j). \quad (4.1)$$

A higher value of  $a_{\text{in}}$  and a lower value of  $a_{\text{out}}$  implies that the anchor clients are more densely connected within the clusters and loosely connected across the clusters. The following assumption quantifies the difference needed on  $a_{\text{in}}$  and  $a_{\text{out}}$  to correctly cluster the anchor clients.

**Assumption 1.** (*graph connectivity*)

$$a_{\text{in}} > \frac{1}{2} + c + (1/\rho + c') a_{\text{out}}, \quad (4.2)$$

where  $c \equiv c(K, \epsilon/\rho)$  and  $c' \equiv c'(K, \epsilon/\rho)$  are two small constants depending on  $K$  and  $\epsilon/\rho$ .

Note that  $a_{\text{in}} > 1/2$  is necessary; otherwise, a cluster may split into two disjoint sub-clusters even without Byzantine clients, in which case there is no hope of reliably recovering the cluster structure. Both  $c$  and  $c'$  decay as  $\epsilon/\rho$  goes to 0. In particular, when  $\epsilon = 0$ , both  $c$  and  $c'$  will be 0.

With Assumption 1, we now utilize the following SDP to recover the cluster label for anchor clients:

$$\max \langle \mathbf{X}, \mathbf{A} - \lambda \mathbf{J} - \gamma \mathbf{I} \rangle \quad \text{s.t. } \mathbf{X} \succeq \mathbf{0}, \mathbf{X} \geq \mathbf{0}, \mathbf{X}(i, i) \leq 1, \forall i, \quad (4.3)$$

where  $\mathbf{I}$  and  $\mathbf{J}$  denote the identity and all-one matrices, respectively; and  $\mathbf{X} \succeq \mathbf{0}$  is the semi-definite constraint. Here  $0 < \lambda < 1$  and  $\gamma > 0$  are two tuning parameters. Intuitively, the introduction of  $\gamma$  promotes the low-rank structure of the solution. The parameter  $\lambda$  is the key to inducing the desired cluster structure of the solution. Intuitively,  $[\mathbf{A} - \lambda \mathbf{J}](i, j)$  is positive if  $\mathbf{A}(i, j) = 1$  and negative otherwise. This creates an incentive for the optimal solution  $\mathbf{X}(i, j)$  to be positive if  $\mathbf{A}(i, j) = 1$  and 0 otherwise. Since we expect  $\mathbf{A}$  to have a large number of 1's within clusters and a few 1's between clusters, the introduction of  $\lambda$  helps to recover the cluster structure.

**Theorem 4.** *Suppose  $\epsilon/\rho = O(1/K)$ . Under Assumption 1, there exist choices of  $\lambda$*

and  $\gamma$  such that optimal solutions of SDP (4.3) must be of the form

$$\widehat{\mathbf{X}} = \mathbf{P}^\top \begin{pmatrix} \mathbf{J}_{m_1} & \cdots & \mathbf{0} & * \\ \vdots & \ddots & \vdots & * \\ \mathbf{0} & \cdots & \mathbf{J}_{m_k} & * \\ * & \cdots & * & * \end{pmatrix} \mathbf{P}, \quad (4.4)$$

where  $\mathbf{P}$  is some unknown permutation matrix capturing the permutation of anchor client indices,  $\mathbf{J}_{m_k}$  denotes the  $m_k \times m_k$  all-one matrix, and  $*$  denotes any arbitrary values.

In short, the solution  $\widehat{\mathbf{X}}$  ensures that rows and columns corresponding to good anchor clients within the same cluster form an all-one block. With such a nice structure, we can obtain an accurate clustering of anchor clients by greedily comparing the rows of  $\widehat{\mathbf{X}}$  (cf. Algorithm 3).

A similar result is obtained in [CL15] with (4.3) for robust community detection under the stochastic block model, i.e., *independent* edge formation with a high chance of forming an edge within the same cluster and low chance between clusters. In contrast, Theorem 4 works deterministically for arbitrary graphs as long as Assumption 1 holds and hence can be applied in more settings. To prove Theorem 4, we carefully construct dual variables which together with primal solutions (4.4) is shown to satisfy the KKT conditions and hence certifies the optimality of (4.4) under Assumption 1.

In passing, we remark that SDP has been extensively used for clustering with outliers. However, the existing literature often imposes distributional assumptions (cf. Section Section 4.6 in the supplementary file for a more detailed discussion). Thus, our deterministic result for robust clustering could be of independent interest.

After recovering the cluster structure of anchor clients, PS computes the coarse estimates  $\{\widehat{\theta}_k\}_{k \in [K]}$  in Step 2 using the iterative filtering algorithm [SCV17, Algorithm

1], which is a standard robust mean estimator. Denote the variance of local estimates as

$$\nu_k^2 \triangleq \left\| \frac{1}{m_k} \sum_{i \in \mathcal{A}_g \cap \mathcal{C}_k} \left( \theta_i^{(0)} - \bar{\theta}_k \right) \left( \theta_i^{(0)} - \bar{\theta}_k \right)^\top \right\|_{\text{op}},$$

where  $\|\cdot\|_{\text{op}}$  denotes the matrix spectral norm. It follows from [SCV17, Proposition 16] that as long as  $\epsilon/\rho < 1/4$ , the outputs of the iterative filtering algorithm [SCV17, Algorithm 1] satisfy

$$\left\| \widehat{\theta}_k^{(0)} - \bar{\theta}_k \right\|_2 \leq C \sqrt{\frac{\epsilon}{\rho}} \nu_k$$

for some universal constant  $C > 0$ .

### 4.3.2 Theoretical Guarantees of Phase 2

The coarse estimates of cluster centers obtained in Phase 1 play a crucial role in Step 3. In particular, each client estimates its cluster label  $\widehat{z}_i$  by minimizing local empirical losses evaluated at these estimates, i.e.,

$$\widehat{z}_i = \arg \min_{k \in [K]} \frac{1}{n_i} \sum_{j=1}^{n_i} l \left( \widehat{\theta}_k^{(0)}; X_{ij} \right),$$

where  $\{X_{ij} : j \in [n_i]\}$  denotes the sample of  $n_i$  data points drawn by client  $i$  independently according to distribution  $\mathcal{D}_{z_i}$ .

To ensure small misclassification errors, we want  $\nu_k$  to be relatively small so that  $L_k(\widehat{\theta}_k^{(0)})$  is still smaller than  $L_k(\widehat{\theta}_j^{(0)})$  for  $j \neq k$ . The following assumption formalizes the control of the fluctuation.

**Assumption 2.** (*fluctuation of estimates from anchor clients*) *There exists a universal constant  $C > 0$  such that*

$$\nu_k^2 < \frac{C\rho}{\beta^2\epsilon} \cdot \Gamma_k^2, \tag{4.5}$$

where  $\Gamma_k \triangleq \max_{j \neq k} \left\{ L_k \left( \bar{\theta}_j^{(0)} \right) - L_k \left( \bar{\theta}_k^{(0)} \right) \right\}$  is the gap of the local population loss.

In addition, the random fluctuation of local empirical losses themselves also affects the misclassification rate. Intuitively, if the local empirical loss is highly volatile, the misclassification rate will be high even with a large loss gap  $\Gamma_k$ . The following assumption quantifies the control of the volatility.

**Assumption 3.** (*fluctuation of local empirical loss*)

$$\frac{\log(M_k)}{M_k} < \frac{CKF_k}{\Gamma_k^2} \leq \frac{\epsilon}{\rho_k}, \quad \forall k \in [K]$$

where  $F_k = \frac{1}{M_k} \sum_{i: z_i=k} \frac{\tau_k^2}{n_i}$  and  $\tau_k^2 = \sup_{\theta} \mathbb{E}_{X \sim \mathcal{D}_k} [(l(\theta; X) - L_k(\theta))^2]$ .

Note that  $F_k/\Gamma_k^2$  can be interpreted as the noise-to-signal ratio, which is typically small when each cluster has enough data points per client on average.

Now we present our guarantee on the success of clustering of all clients in Step 3.

**Theorem 5.** *Suppose Assumption 1–3 hold. Step 3 of Algorithm 1 outputs  $\{\widehat{\mathcal{C}}_k\}_{k \in [K]}$  such that with probability tending to 1 as  $M_k$  diverges, there exists a permutation  $\pi$  ensuring*

$$\frac{|\widehat{\mathcal{C}}_{\pi(k)} \cap \mathcal{C}_k^c|}{|\widehat{\mathcal{C}}_{\pi(k)}|} \leq \frac{K\epsilon}{\rho_k} \equiv \alpha_k, \quad \forall k \in [K], \quad (4.6)$$

Here,  $\alpha_k$  is the worst case bound when all misclassified clients and Byzantine clients join the same cluster. In practice, we expect the misclassification errors to spread out across all clusters.

With the guarantee of Theorem 5, we can then apply Byzantine Gradient Descent (cf. Algorithm 5) separately within each output cluster to iteratively refine the cluster center estimates.

In order to ensure sufficient data for the refinement, we make the following assumptions.

**Assumption 4.** (*Preserving data from correctly classified clients*) For any  $k \in [K]$ ,

$$\frac{CKM_k\tau_k^2}{\Gamma_k^2} < N_k,$$

for a universal constant  $C > 0$ .

The above assumption ensures that for every cluster  $k$ , the expected number of data on misclassified clients, which is at most  $C'KM_k\tau_k^2/\Gamma_k^2$ , only accounts for a fraction of the total number of data points within the cluster.

Recall that to resolve the challenges of unbalanced data, we randomly partition the clients into batches in Step 4 to ensure more evenly distributed data over the batches. The following assumption gives the maximum level of unbalancedness we can tolerate.

**Assumption 5.** (*Tolerance of unbalancedness*) For any  $k \in [K]$ ,

$$N_k > C\epsilon Mn_{\max}, \quad \frac{N_k}{n_{\max}} \geq C' \log M_k \quad \text{and} \quad (n_{\max} - n_{\min})^2 < \frac{C'N_k^2}{\alpha_k M_k^2 \log M_k}.$$

for some universal constant  $C$  and  $C'$ , where  $n_{\max} = \max_{i \in [M]} n_i$  and  $n_{\min} = \min_{i \in [M]} n_i$ .

The first inequality ensures that even if the Byzantine adversary controls the clients with the largest amount of data within the cluster, we still have a sufficient number of data for the refinement step. The second inequality also rules out the extreme situation when most of the data are stored on very few clients.

Now we explain the intuition of the last inequality. With random partition, the *mean* number of data points in each batch is exactly the same. We need to further control the deviation of the data volume from the mean across all batches. The last inequality is introduced to circumvent the extreme situation where there exists a batch only containing data-scarce clients so that its data volume is much smaller than the mean.

Now we present the performance guarantee of Algorithm 1 on model estimation for both strongly-convex and non-convex functions.

Denote  $\sigma_k^2 = \sup_{\theta} \mathbb{E}_{X \sim \mathcal{D}_k} [\|\nabla l(\theta; X) - \nabla L_k(\theta)\|_2^2]$ .

**Theorem 6.** *Suppose Assumptions 1–5 hold. With probability tending to 1 as  $M_k$  diverges, the output  $\widehat{\theta}_k^{(T)}$  of Algorithm 1 after  $T = O(\log M_k)$  iterations of Step 4 satisfies the followings:*

- If  $L_k$  is  $\mu$ -strongly convex and  $\zeta$ -smooth,

$$\left\| \widehat{\theta}_k^{(T)} - \theta_k \right\|_2^2 \leq \left( 1 - \frac{\mu^2}{8\zeta^2} \right)^T \left\| \widehat{\theta}_k^{(0)} - \theta_k \right\|_2^2 + \frac{C' \sigma_k^2 \log M_k \max\{\alpha_k M_k, 1\}}{\mu^2 N_k}; \quad (4.7)$$

- If  $L_k$  is non-convex and  $\zeta$ -smooth,

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\| \nabla L_k \left( \widehat{\theta}_k^{(t)} \right) \right\|_2^2 \leq \frac{2\zeta}{T} \left( L_k \left( \widehat{\theta}_k^{(0)} \right) - L_k \left( \theta_k \right) \right) + \frac{C' \sigma_k^2 \log M_k \max\{\alpha_k M_k, 1\}}{N_k}. \quad (4.8)$$

Note that the number of batches needed in Step 4 is proportional to  $\max\{\alpha_k M_k, 1\}$ . Hence,  $N_k / \max\{\alpha_k M_k, 1\}$  is roughly the average number of data within one batch. When there is no misclassification within the cluster, i.e.,  $\alpha_k M_k < 1$ , we only need to form a single batch; thus Algorithm 1 achieves  $\widetilde{O}\left(\frac{\sigma_k^2}{N_k}\right)$  estimation error which matches the optimal estimation error rate in the centralized failure-free setting up to a logarithmic factor. When there is a misclassification error, the estimation error rate is deteriorated by a factor no more than  $\alpha_k M_k$ .

Now we compare Theorem 6 with [GHYR19, Theorem 2]. Note that [GHYR19, Theorem 2] imposes various restricted assumptions including the sub-Gaussianity of the local estimates, the sub-exponentiality of gradients, and the relatively good initial labeling. Under these assumptions, it is shown in [GHYR19, Theorem 2] that the algorithm proposed therein achieves  $\widetilde{O}\left(\frac{\alpha_k^2 \sigma_k^2}{n_{\min}}\right)$  estimation error, where  $n_{\min}$  is the min-

imum number of local data points across all clients. Our result is better in two major aspects. First of all, Theorem 6 does not require any distributional assumption on local estimates or gradients and hence is more general. Secondly, Algorithm 1 does not require a sufficient number of data on each client and hence handles the unbalancedness of data much better. In particular, when  $n_{\min} = O(\alpha_k N_k \log(N_k) / M_k \log(M_k))$ , Algorithm 1 achieves a much smaller estimation error.

## 4.4 Byzantine Federated Learning Algorithm

In this section, we discuss each step of Algorithm 1 in more detail.

### 4.4.1 Phase 1: Coarse Estimation of Cluster Centers

#### Step 1: Initial Clustering of Anchor Clients

Due to the unbalanced data among clients, estimates from some local clients can be unreliable due to insufficient data. To resolve this issue, we utilize anchor clients who have a sufficient number of data to estimate the cluster centers. Intuitively, estimates from anchor clients are expected to scatter around the true cluster centers. As long as the cluster centers are moderately separated, we can utilize semi-definite programming (SDP) to accurately and robustly cluster anchor clients. In particular, we propose the following algorithm to estimate the clusters.

Recall that we use  $\mathcal{A}$  to denote the collection of anchor clients. Throughout this subsection, for ease of presentation, we re-index clients in  $\mathcal{A}$  from 1 to  $m$ , where  $m = |\mathcal{A}|$ .

---

**Algorithm 2** SDP-based Clustering

---

**Input:**  $\{\theta_j^{(0)}\}_{j \in [m]}$ ,  $\lambda, \gamma, \tau$

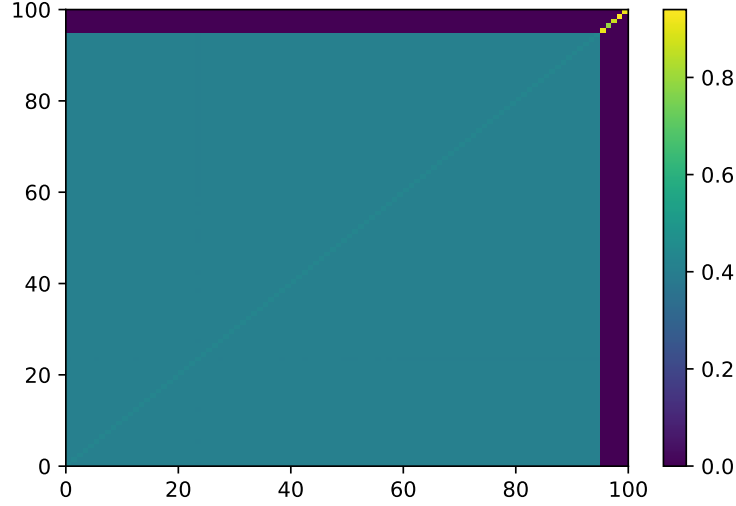
**Output:**  $\widehat{\mathbf{X}}$

- 1: Compute the pairwise distance matrix  $\mathbf{D}$  such that  $\mathbf{D}(i, j) = \|\theta_i^{(0)} - \theta_j^{(0)}\|_2$ ;
  - 2: **Thresholding:** Construct graph  $\mathbf{A}$  such that  $\mathbf{A}(i, i) = 0$  and  $\mathbf{A}(i, j) = \mathbf{1}_{\{\mathbf{D}(i, j) \leq \tau\}}, i \neq j$ ;
  - 3: **SDP:** Output an optimal solution  $\widehat{\mathbf{X}}$  to semidefinite programming (4.3);
  - 4: **Clustering:** Cluster anchor clients based on rows of  $\widehat{\mathbf{X}}$  via Algorithm 3.
- 

Note that we cannot directly use pairwise distance matrix  $\mathbf{D}$  in semi-definite programming (4.3). As shown in Figure 4.1, if the local model estimates of the Byzantine clients are extremely far from those of good clients, SDP (4.3) with  $\mathbf{D}$  as input will mistakenly assign all good clients as one single cluster. As such, we introduce the **Thresholding** step to limit the impact of Byzantine clients. With an appropriate threshold chosen, the adjacency matrix  $\mathbf{A}$  tends to have more edges connection within the same clusters and fewer edge connections across different clusters.

After obtaining  $\mathbf{A}$ , one naïve attempt is to compare the rows of  $\mathbf{A}$  and assign clients with very similar rows to the same cluster. However, even in the absence of the Byzantine adversary, such a method demands  $a_{\text{in}}$  to be close to 1 and  $a_{\text{out}}$  to be close to 0, in order to guarantee the accurate clustering for all good anchor clients, where  $a_{\text{in}}$  and  $a_{\text{out}}$  are the assortativity and assortativity defined in (4.1). The **SDP** Step can be viewed as a denoising procedure applied to  $\mathbf{A}$ . Even when Byzantine adversary corrupts an  $\epsilon$  fraction of clients and  $a_{\text{in}}, a_{\text{out}}$  are only moderately separated, we can show that optimal solutions of the SDP  $\widehat{\mathbf{X}}$  must have a desired diagonal-block form (4.4). In particular, rows and columns corresponding to good anchor clients within the same cluster form an all-one block.

Given the nice structure of  $\widehat{\mathbf{X}}$ , we can now recover the clusters of good anchor clients simply by comparing the rows of  $\widehat{\mathbf{X}}$ . Specifically, if good client  $i$  and  $j$  belong to the same cluster, then  $\widehat{\mathbf{X}}(i)$  and  $\widehat{\mathbf{X}}(j)$  differ by at most  $\epsilon m$  coordinates, where  $\widehat{\mathbf{X}}(i)$  denotes the  $i$ -th row of  $\widehat{\mathbf{X}}$ . In contrast,  $\widehat{\mathbf{X}}(i)$  and  $\widehat{\mathbf{X}}(j)$  differ by at least  $2\rho m$  coordinates if they belong to different clusters. As long as  $2\rho$  is much larger than  $\epsilon$ , we are able to tell if two clients belong



**Figure 4.1:** Solution to (4.3) with the adjacency matrix  $\mathbf{A}$  replaced by the negative pairwise distance matrix  $-\mathbf{D}$ . Color represents the value of each entry of the solution. Input data is the local model estimates of anchor clients in the numerical study from Section 4.5.1 with  $\chi = 0.9$  under Gaussian attack. Good clients are indexed from 0 to 94 and Byzantine clients are indexed from 95 to 99. The plot shows that rows and columns of good clients form a block matrix with identical entry values, and hence the SDP fails to extract the true underlying clusters.

to the same cluster or not. Based on this intuition, we propose Algorithm 3 to recover the cluster label  $z_i$  for good clients  $i$ .

---

**Algorithm 3** Clustering based on counting common neighbors

---

**Input:**  $\{\widehat{\mathbf{X}}(i)\}_{i \in [m]}$

**Output:**  $K$  Clusters  $\{\widetilde{\mathcal{C}}_k\}_{k \in [K]}$

- 1: Construct a graph  $G = ([m], E)$  such that  $i$  and  $j$  are connected if  $\widehat{\mathbf{X}}(i)$  and  $\widehat{\mathbf{X}}(j)$  differ by no more than  $\epsilon m$  coordinates
  - 2: Pick a unassigned node  $v$  with degree more than  $2\epsilon m + 1$  and assign all nodes that share more than  $\epsilon m + 1$  neighbors with  $v$  in the same cluster (including  $v$ )
  - 3: Continue the above process until all nodes with degrees more than  $2\epsilon m + 1$  are assigned
  - 4: Output the  $K$  clusters of the largest sizes.
- 

The following lemma shows that Algorithm 3 correctly clusters all non-Byzantine clients once  $\widehat{\mathbf{X}}$  has the desired form (4.4).

**Lemma 4.4.1.** *Suppose  $\widehat{\mathbf{X}}$  is of the form (4.4). Then there exists a permutation  $\pi$  such that the output  $\widetilde{\mathcal{C}}_k$  of Algorithm 3 satisfies*

$$\mathcal{A}_g \cap \mathcal{C}_k \subseteq \widetilde{\mathcal{C}}_{\pi(k)}, \forall k \in [K].$$

The proof of Lemma 4.4.1 follows from simple case-by-case analysis and is deferred to Section 5.4.

## Step 2: Estimation of Cluster Centers

Recall from Section 4.3 that  $\theta_i^{(0)}$ 's are local model estimates from anchor clients. With the accurate clusters obtained by Algorithm 2, we now utilize  $\{\theta_i^{(0)} : i \in \widetilde{\mathcal{C}}_{\pi(k)}\}$  to obtain a good estimate to  $\theta_k$ . Note that  $\widetilde{\mathcal{C}}_{\pi(k)}$  may contain Byzantine clients; thus we cannot simply use the average over  $\widetilde{\mathcal{C}}_{\pi(k)}$  as our estimator. Instead, we apply a robust mean estimator, known as iterative filtering.

**Theorem 7.** *[SCV17, Proposition 16] If  $\epsilon/\rho_k < 1/4$ , the output of [SCV17, Algorithm 1] satisfies*

$$\|\widehat{\theta}_k^{(0)} - \bar{\theta}_k\|_2 \leq C \sqrt{\frac{\epsilon}{\rho_k}} \sigma_A.$$

for some universal constant  $C > 0$  where  $\bar{\theta}_k = \frac{1}{|\mathcal{A}_g \cap \widetilde{\mathcal{C}}_{\pi(k)}|} \sum_{i \in \mathcal{A}_g \cap \widetilde{\mathcal{C}}_{\pi(k)}} \theta_i^{(0)}$ .

Intuitively, Theorem 7 says that even though there are  $\epsilon/\rho_k$  fraction of Byzantine clients, we are still able to output an estimator that is close to  $\bar{\theta}_k$ . Note that  $\bar{\theta}_k$  the average of  $\theta_i^{(0)}$  from clients in  $\mathcal{A}_g \cap \mathcal{C}_k$ . With a sufficient number of data points on anchor clients from cluster  $k$ , we expect  $\bar{\theta}_k$  to be very close to  $\theta_k$ . As a result, we obtain a coarse estimation  $\widehat{\theta}_k^{(0)}$  to  $\theta_k$  as well.

The key of the iterative fitting algorithm [CSV17, Algorithm 1] lies on approximating each  $\theta_i^{(0)}$  as an average of  $(1 - \epsilon)m$  other  $\theta_j^{(0)}$ 's within the same cluster. Intuitively, if  $\theta_i^{(0)}$  indeed belongs to this cluster, it should be similar to other  $\theta_j^{(0)}$ 's within the same cluster. Hence, we expect a small approximation error. In contrast, if we observe a large

approximation error,  $\theta_i^{(0)}$  is likely to be an outlier and thus should be removed. Iteratively applying the above filtering procedure, we limit the impact of Byzantine clients.

## 4.4.2 Phase 2: Clustering and Refinement of the Estimation

### Step 3: Robust Clustering for All Clients

Recall that in Phase 1, the estimator of  $\theta_k$  only utilizes the local data from a few anchor clients. In order to obtain a more accurate estimator of  $\theta_k$ , we want to utilize data from all clients. To this end, we need to first recover the cluster labels for all clients. In Phase 1, we have obtained a relatively accurate estimator  $\widehat{\theta}_k^{(0)}$  which is closer to  $\theta_k$  than any other  $\theta_j, j \neq k$ . Therefore, local clients in cluster  $k$  is likely to have smaller local empirical losses  $l_i(\widehat{\theta}_k^{(0)})$  than  $l_i(\widehat{\theta}_j^{(0)})$  for all  $j \neq k$ , where

$$l_i(\theta) \triangleq \frac{1}{n_i} \sum_{j=1}^{n_i} l(\theta; X_{ij})$$

is defined as the client  $i$ 's local empirical loss function. Hence, each client can accurately estimate its cluster label by minimizing its local empirical loss at all estimated model parameters.

---

#### Algorithm 4 Classification for all clients

---

**Input:**  $\{\widehat{\theta}_k^{(0)}\}_{k \in [K]}$

**Output:**  $\{\widehat{\mathcal{C}}_k\}_{k \in [K]}$

- 1: Parameter Server (PS) broadcasts  $\{\widehat{\theta}_k^{(0)}\}_{k \in [K]}$
- 2: Each client  $i \in [m]$  determines the cluster label that minimizes the local empirical loss

$$\widehat{z}_i = \arg \min_{k \in [K]} l_i(\widehat{\theta}_k^{(0)}) \quad (4.9)$$

- 3: Output cluster  $\widehat{\mathcal{C}}_k = \{i : \widehat{z}_i = k\}$
-

The clustering accuracy of Algorithm 4 is established in Theorem 5. Here we provide a proof sketch of Theorem 5. We need to crucially control the number of misclassified clients within each  $\mathcal{C}_k$ , which according to (4.9) is given by

$$Q_k = \sum_{i:z_i=k} \mathbf{1}_{\{l_i(\widehat{\theta}_k^{(0)}) > \min_{j \neq k} l_i(\widehat{\theta}_j^{(0)})\}}.$$

Since all the above indicator variables are mutually independent, a simple application of Bernstein inequality shows that  $Q_k$  concentrates on its mean. To further bound the mean of  $Q_k$ , it suffices to bound the probability of client  $i$  in cluster  $k$  being clustered accurately, i.e.,

$$\mathbb{P} \left[ l_i(\widehat{\theta}_k^{(0)}) < \min_{j \neq k} l_i(\widehat{\theta}_j^{(0)}) \right]. \quad (4.10)$$

Observe that if the local empirical losses are close to the population losses, that is,

$$l_i(\widehat{\theta}_k^{(0)}) < L_k(\widehat{\theta}_k^{(0)}) + \Delta, \quad \text{and} \quad l_i(\widehat{\theta}_j^{(0)}) > L_k(\widehat{\theta}_j^{(0)}) - \Delta, \quad (4.11)$$

for some  $\Delta < \frac{L_k(\widehat{\theta}_k^{(0)}) - L_k(\widehat{\theta}_j^{(0)})}{2}$ , then we must have  $l_i(\widehat{\theta}_k^{(0)}) < l_i(\widehat{\theta}_j^{(0)})$ . Hence, we can bound (4.10) by combining the tail probability of the events in (4.11) with a union bound over  $j$ .

The detailed proof is provided in Section 5.2.

#### Step 4: Refining Estimation

In this final step, we refine the model parameter estimation  $\widehat{\theta}_k^{(0)}$  within the estimated cluster  $\widehat{\mathcal{C}}_{\pi(k)}$  via a distributed gradient descent. In particular, upon receiving a global model update, each client in cluster  $\widehat{\mathcal{C}}_{\pi(k)}$  first draws fresh samples from  $\mathcal{D}_k$  to compute a stochastic gradient and then sends the gradient to the PS. The PS then aggregates all the received local gradients and runs a step of gradient descent to update the global model estimate.

However, due to the existence of the Byzantine adversary, we cannot simply take the average of all received local gradients. In particular, a single Byzantine client can completely

skew the average and thus foils the algorithm. Thus, it is crucial to design a robust gradient aggregator. Many of the existing robust gradient aggregators such as those in [YCKB18, GHYR19, DD21] crucially rely on the fact that all good clients have a sufficient amount of data so that their local stochastic gradients are sufficiently accurate. In an unbalanced data setting where some clients may have a very limited number of data, their local stochastic gradients are highly volatile and hence these robust gradient aggregators suffer from a large estimation error. For example, numerical experiments in Section 4.5.2 verify that [GHYR19, Algorithm 1] has a large estimation error when the data is unbalanced.

To resolve the challenge of unbalanced data, we propose Algorithm 5. The key idea is to batch the clients together so that within each batch, there are sufficiently many data points to yield accurate estimation. To implement the batching, one straightforward way is to independently assign each client to one of the batches uniformly at random. Unfortunately, because of the independent assignment, the batch size will be different, and even worse there may exist an empty batch with a certain probability. To circumvent this issue, we instead partition clients into equal-sized batches uniformly at random. However, under such a random partition, the assignments of clients are no longer independent. This dependency brings some technical difficulty in the analysis. In particular, to show the number of data points in each batch concentrates on its mean, we need to apply certain concentration inequality for the sum of correlated random variables.

In passing, we remark that a similar idea of batching has been used in [CSX17, KHJ22] to reduce the variance of the local gradients in the heterogeneous setting. However, both works assume the even distribution of data among clients so the number of data points in each batch is identical. In addition to batching, another idea known as nearest-neighbor mixing [AFG<sup>+</sup>23] was proposed recently to reduce the variance of the local gradients. In particular, each local gradient  $\nabla l_i$  is mixed with a large number of gradients that are close to  $\nabla l_i$ . However, [AFG<sup>+</sup>23] requires the fixed Byzantine clients across communication rounds, which does not hold in our setting.

In the following, we provide the convergence guarantee for both strongly-convex and non-convex smooth loss  $L_k$  under Algorithm 5. Recall that  $\sigma_k^2$  is the variance of the gradient

of loss function  $l$ .

---

**Algorithm 5** Byzantine-resilient Gradient Descent

---

**Input:** Initial estimator  $\widehat{\theta}_k^{(0)}$ , cluster  $\widehat{\mathcal{C}}_{\pi(k)}$ , number of batches  $b$ , step size  $\eta$ , number of iterations  $T$

**Output:**  $\widehat{\theta}_k^{(T)}$

- 1: PS partitions  $\widehat{\mathcal{C}}_{\pi(k)}$  into  $b$  batches  $\{\mathcal{B}_j\}_{j=1}^b$  of equal sizes uniformly at random
- 2: **For**  $t = 1, 2, \dots, T$  **do**:

3: Parameter Server

4: Broadcast  $\widehat{\theta}_k^{(t)}$  to all clients in  $\widehat{\mathcal{C}}_{\pi(k)}$ ;

5: Wait to receive gradients from local client  $i$ :

$$\widehat{g}_i^{(t)} = \begin{cases} g_i^{(t)} & \text{if } i \text{ is good,} \\ * & \text{if } i \text{ is Byzantine, or not receive anything} \end{cases}$$

- 6: Compute the geometric median of the means:

$$\mathcal{G}_k^{(t)} \leftarrow \text{GM} \left( \left\{ \widehat{g}_j^{(t)} \right\}_{j \in [b]} \right), \quad \text{where } \bar{g}_j^{(t)} = \frac{1}{\sum_{i:i \in \mathcal{B}_j} n_i} \sum_{i:i \in \mathcal{B}_j} \widehat{g}_i^{(t)}.$$

- 7: Update:  $\widehat{\theta}_k^{(t+1)} = \widehat{\theta}_k^{(t)} - \eta \mathcal{G}_k^{(t)}$ .

8: Client  $i$  in  $\widehat{\mathcal{C}}_{\pi(k)}$

9: Draw fresh sample of  $n_i$  data points  $\{X_{ir}^{(t)}\}_{r \in [n_i]}$  from  $\mathcal{D}_k$

10: Send gradient  $g_i^{(t)} = \sum_{r=1}^{n_i} \nabla l(\widehat{\theta}_k^{(t)}; X_{ir}^{(t)})$  to PS

---

**Theorem 8.** Assume  $L_k$  is  $\mu$ -strongly convex and  $\zeta$ -smooth. Let  $b = \max\{2.5\alpha_k|\widehat{\mathcal{C}}_{\pi(k)}|, 1\}$  where  $\alpha_k$  as the misclassification rate of cluster  $\widehat{\mathcal{C}}_{\pi(k)}$  including Byzantine clients. With probability at least  $1 - T \exp(-Cb \log \log M_k)$  where  $C > 0$  is some universal constant, Algorithm 5 with step size  $\eta = \frac{\mu}{2\zeta^2}$  outputs

$$\left\| \widehat{\theta}_k^{(t)} - \theta_k \right\|_2 \leq \left( 1 - \frac{\mu^2}{8\zeta^2} \right)^t \left\| \widehat{\theta}_k^{(0)} - \theta_k \right\|_2 + \frac{C\sqrt{\log M_k}\sigma_k}{\mu\sqrt{W_k}}, \quad \forall t = 1, 2, \dots, T, \quad (4.12)$$

where  $W_k = \min_{j \in [b]} \sum_{i \in \mathcal{B}_j} n_i$ .

**Theorem 9.** Assume  $L_k$  is  $\zeta$ -smooth. Let  $b = \max\{2.5\alpha_k\widehat{M}_k, 1\}$ . With probability at least

$1 - T \exp(-Cb \log \log M_k)$ , Algorithm 5 with step size  $\eta = \frac{1}{\zeta}$  ensures

$$\frac{1}{T} \sum_{t=0}^{T-1} \left\| \nabla L_k(\hat{\theta}_k^{(t)}) \right\|_2^2 \leq \frac{2\zeta}{T} \left( L_k(\hat{\theta}_k^{(0)}) - L_k(\theta_k) \right) + \frac{114 \log M_k \sigma_k^2}{W_k}. \quad (4.13)$$

Both Theorem 8 and 9 show the tradeoff on the choice of  $b$ . With a larger  $b$ , we can tolerate more Byzantine clients, i.e., larger  $\alpha_k$ . However, the estimation error, i.e., the second term on the right hand side of (4.12) is larger due to smaller  $W_k$ .

Here, we explain the choice of the number of batches  $b$ . Recall that  $\alpha_k$  is an upper bound of the misclassification rate of  $\hat{\mathcal{C}}_{\pi(k)}$ . When  $\alpha_k$  is non-negligible, we take the number of the batches  $b$  to be strictly larger than  $2\alpha_k |\hat{\mathcal{C}}_{\pi(k)}|$  to limit the impact from Byzantine and misclassified clients. In particular, we ensure more than half of the batches are free of Byzantine and misclassified clients. When there are no misclassified or Byzantine clients in  $\hat{\mathcal{C}}_{\pi(k)}$ , only one batch is needed, i.e.,  $b = 1$ , and hence Algorithm 5 reduces to the standard gradient descent training.

To prove both theorems, we show the convergence of the estimation error deterministically under  $\cap_t \Omega_t$  and then show  $\cap_t \Omega_t$  occurs with high probability, where

$$\Omega_t \equiv \Omega_t(\Delta, \xi; b) = \left\{ \sum_{j=1}^b \mathbf{1} \left\{ \left\| \frac{1}{\sum_{i \in \mathcal{B}_j} n_i} \sum_{i \in \mathcal{B}_j} g_i^{(t)} - \nabla L_k(\hat{\theta}_k^{(t)}) \right\|_2 \leq \Delta \right\} \geq b(1 - \xi) + \alpha_k \widehat{M}_k \right\}, \quad (4.14)$$

for some constant  $\xi \in \left( \frac{\alpha_k \widehat{M}_k}{b}, \frac{1}{2} \right)$ . In short,  $\Omega_t(\Delta, \xi; b)$  denotes the event that there are at least  $b(1 - \xi) + \alpha_k \widehat{M}_k$  batches whose average gradients are close to the true gradient  $\nabla L_k(\hat{\theta}_k^{(t)})$  in the Byzantine-free case. Here, the lower bound of  $\xi$  is to ensure  $b(1 - \xi) + \alpha_k \widehat{M}_k$  is smaller than  $b$ , i.e.  $\Omega_t$  is not an empty event. Since there are at most  $\alpha_k \widehat{M}_k$  batches containing Byzantine clients, under  $\Omega_t$  we are left with at least  $(1 - \xi)b > b/2$  batches in which there is no Byzantine client and the average gradient is close to the true gradient. By the standard robustness property of the geometric median, this ensures the geometric median of means  $\mathcal{G}_t$  is close to  $\nabla L_k(\hat{\theta}_k^{(t)})$  (cf. Lemma 5.3.1).

In view of (4.12) and (4.13), both estimation errors crucially depend on the minimum

number of data among batches  $W_k$ . The following lemma shows that with high probability,  $W_k = \Omega\left(\frac{N_k}{b}\right)$ . As a result, we ensure an approximately even distribution of data across batches.

**Lemma 4.4.2.** *Under Assumption 4 and 5, with probability  $1 - \exp(-\Omega(\log M_k))$ ,*

$$W_k = \Omega\left(\frac{N_k}{\max\{\alpha_k M_k, 1\}}\right).$$

To prove Lemma 4.4.2, we need to bound from below the number of data in each batch  $j$ , that is,

$$Y_j = \sum_{i \in \mathcal{C}_k \cap \widehat{\mathcal{C}}_{\pi(k)}} n_i \mathbf{1}_{\{i \in \mathcal{B}_j\}}.$$

Here,  $Y_j$  involves two sources of randomness. The first one comes from the freshly drawn data from clients which determines  $\widehat{\mathcal{C}}_{\pi(k)}$ . The second one is the random partition of  $\widehat{\mathcal{C}}_{\pi(k)}$  into  $\mathcal{B}_j$ 's.

To obtain the desired lower bound shown in the lemma, we first condition on  $\widehat{\mathcal{C}}_{\pi(k)}$  and show  $Y_j$  concentrates on its conditional mean  $\mathbb{E}[Y_j | \widehat{\mathcal{C}}_{\pi(k)}]$ . One issue here is that  $\mathbf{1}_{\{i \in \mathcal{B}_j\}}$  and  $\mathbf{1}_{\{r \in \mathcal{B}_j\}}$  are correlated because of the equal size constraint of  $\mathcal{B}_j$ 's. Thus, one cannot apply vanilla concentration inequalities for the sum of independent random variables. To resolve the dependency issue, we turn to the concentration inequality given in Lemma A.1.5, which works for the sum of correlated random variables in our case.

It remains to bound  $\mathbb{E}[Y_j | \widehat{\mathcal{C}}_{\pi(k)}]$ . With the random partition, we know

$$\mathbb{E}[Y_j | \widehat{\mathcal{C}}_{\pi(k)}] = H_k/b, \quad \text{where } H_k = \sum_{i \in \mathcal{C}_k} n_i \mathbf{1}_{\{i \in \widehat{\mathcal{C}}_{\pi(k)}\}}$$

is the total number of data from correctly classified clients. Intuitively, with a small  $\alpha_k$ , we expect that most of the data from clients in  $\mathcal{C}_k$  is maintained in  $\mathcal{C}_k \cap \widehat{\mathcal{C}}_{\pi(k)}$  and most clients in  $\mathcal{C}_k$  are correctly classified. In particular, we show  $H_k \geq \Omega(N_k)$  and  $\widehat{M}_k = O(M_k)$ . As a result,

$$\frac{H_k}{b} = \frac{H_k}{\max\{2.5\alpha_k \widehat{M}_k, 1\}} \geq \frac{CN_k}{\max\{\alpha_k M_k, 1\}}$$

for some universal constant  $C > 0$ . The detailed proof is provided in Section 5.3.2.

## 4.5 Numerical Studies

In this section, we provide numerical studies to corroborate our theoretical findings. More extensive numerical studies are deferred to the supplementary material.

Consider a federated learning system with  $M = 4000$  clients and model parameter dimension  $d = 30$ . Clients belong to three unknown clusters with proportional cluster sizes equal to  $\rho_1 = 0.3$ ,  $\rho_2 = 0.3$ , and  $\rho_3 = 0.35$ , respectively. The rest  $\epsilon = 0.05$  fraction of clients is assumed to be controlled by the Byzantine adversary. For simplicity, we assume the set of Byzantine clients is fixed across all communication rounds. In each communication round, each client  $i$  draws a sample of  $n_i$  data points  $\{(X_{ij}, y_{ij})\}_{j \in [n_i]}$ , where  $X_{ij}$  are independently generated from Gaussian distribution  $\mathcal{N}(0, \mathbf{I}_d)$  and  $y_{ij} = X_{ij}\theta_{z_i} + \xi_{ij}$  for  $\xi_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.64)$ . The model parameters  $\{\theta_k\}$  and local data volume  $n_i$  will be specified later.

We consider two representative types of Byzantine attacks mentioned in [XKG18]. In the Gaussian attack, each Byzantine client outputs a random Gaussian vector with variance  $200\mathbf{I}_d$ . In the Bit-flip attack, each Byzantine client flips the 22th, 30th, 31th, and 32th bits of the binary representation of each coordinate of the output vector.

### 4.5.1 Comparison with Methods Ignoring Cluster Structure

As mentioned in Section 4.1, one recent line of work [DD21, KHJ22, AFG<sup>+</sup>23] deals with the data heterogeneity by assuming the dissimilarity among local gradients is bounded while ignoring the cluster structure. Here, we compare the performance of the algorithms therein with Algorithm 1.

We consider model parameters with a varying level of separation  $(\chi\theta_1, \chi\theta_2, \chi\theta_3)$  by choosing different values of  $\chi$ , where  $\theta_1 = \frac{1}{2}\mathbf{1} - \frac{1}{2}e_1$ ,  $\theta_2 = 2e_1$ ,  $\theta_3 = -\mathbf{1}$ , with  $\mathbf{1} \in \mathbb{R}^d$  denoting the all one vector and  $e_1 = (1, 0, \dots, 0)^\top \in \mathbb{R}^d$ . As  $\chi$  increases, the cluster centers are more

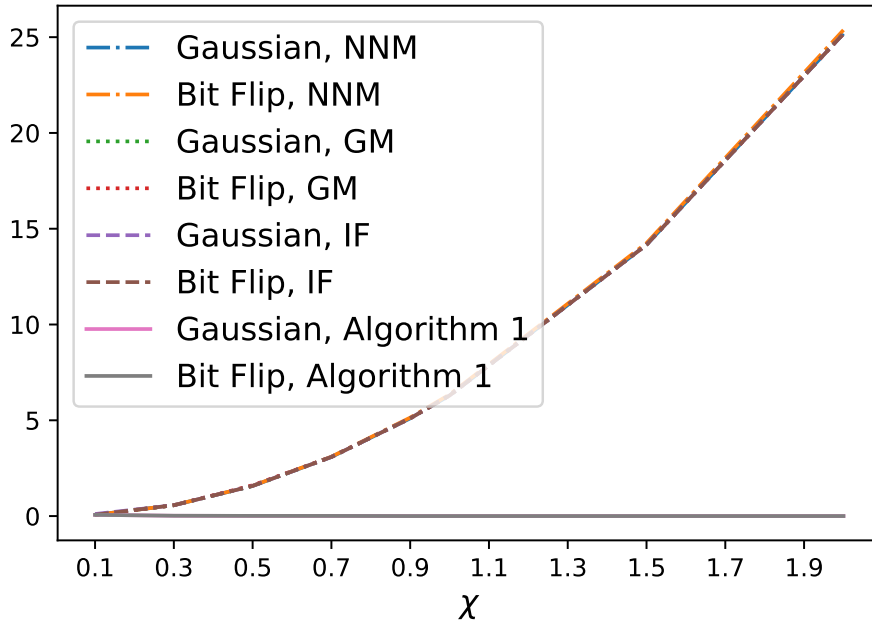
and more separated, indicating a clearer cluster structure. To model the unbalancedness of data, we assume the number of data  $n_i$  on each client  $i$  are i.i.d. generated from the Poisson distribution with mean 23.

We consider three benchmark algorithms. Denote GM as the algorithm proposed in [KHJ22] which groups the clients into batches and then applies the geometric median of means to update the estimate of the model parameters. Denote NNM as the algorithm proposed in [AFG<sup>+</sup>23] which first applies the nearest-neighbor mixing for local gradients and then applies geometric median to update the estimate of the model parameters. Denote IF as the iterative filtering algorithm used in [DD21].

To implement Algorithm 1, we pick 100 anchor clients uniformly at random from all clients with more than 35 number of data points. Then the Byzantine adversary randomly selects  $\epsilon = 5\%$  of anchor clients to attack. All algorithms are run for 20 iterations with step size 0.1.

We evaluate the algorithms using the weighted population loss, that is,  $L(\{\hat{\theta}_k\}_{k \in [K]}) = \sum_{k \in [K]} \rho_k L_k(\hat{\theta}_k)$ , where  $L_k(\theta) = \frac{1}{2} \|\theta - \theta_k\|_2^2$  is the population loss on cluster  $k$ , and  $\hat{\theta}_k$  is the model estimator on cluster  $k$ . Since the three benchmark algorithms ignore the cluster structure and only output a single model  $\hat{\theta}$ , their performance is measured by taking  $\hat{\theta}_k = \hat{\theta}$  for all  $k \in [K]$ .

In Figure 4.2, we vary the separation level  $\chi$  and plot the weighted population losses achieved by different algorithms. We observe that when  $\chi$  gets larger, i.e., the cluster structure is more and more evident, our algorithm performs significantly better than all the other algorithms. When  $\chi$  is small, although the cluster structure is not very evident, our algorithm still performs as well as the other algorithms. Note that the two curves for Algorithm 1 are very close to each other, so visually we only see one line. This is because Algorithm 1 is able to learn the true model parameters of all clusters well regardless of the attack types. Also, the curves for NNM, GM, and IF are close to each other. This is because the estimators  $\hat{\theta}$  under all three algorithms converge to the same global model parameter.



**Figure 4.2:** Weighted population loss with varying cluster separation.

## 4.5.2 Comparison with Methods Utilizing Cluster Structure

As aforementioned, [GHYR19] proposes an algorithm taking both cluster structure and Byzantine attack into account. Here, we compare that algorithm with Algorithm 1.

### Synthetic data

To simulate the practical scenario where the majority of the clients have limited data while a few have relatively large amounts of data points, we assume that 90% of the clients  $i$  have  $n_i$  i.i.d. generated from  $\max\{\text{Poisson}(2), 1\}$  while the remaining 10% have 35 number of data points. Such setup models the practical setting where many clients have very few data points and only a small fraction of clients have a sufficient number of data for estimation. The cluster centers are set to be  $(0.9\theta_1, 0.9\theta_2, 0.9\theta_3)$  where  $(\theta_1, \theta_2, \theta_3)$  is defined as before.

We adopt the same experiment setup in Section 4.5.1. In particular, there are 95 good anchor clients and 5 anchor clients under Byzantine attack. Recall that [GHYR19, Algorithm 1] needs a good initial clustering as input. Here we assign 70% of the clients

within each cluster with the true cluster label and the other 30% to one another cluster.<sup>1</sup> In particular, we mislabel 30% clients in cluster 1 to cluster 2, 30% clients in cluster 2 to cluster 3, and 30% clients in cluster 3 to cluster 1 correspondingly.

Table 4.1 summarizes the misclassification rates attained by both algorithms. It can be seen that the misclassification rate of our algorithm is significantly lower than [GHYR19, Algorithm 1]. The underlying reason for such a large performance gap is that [GHYR19, Algorithm 1] demands the local model estimates are accurate for all clients. In contrast, by leveraging anchor clients who have a sufficient number of data points to produce relatively accurate local estimates, our algorithm significantly relaxes the initialization requirement.

**Table 4.1:** Comparison of misclassification errors under Gaussian attack

	Misclassification Error		
	Cluster 1	Cluster 2	Cluster 3
Our algorithm	14.49%	14.43%	10.94%
[GHYR19, Algorithm 1]	27.26%	40.64%	31.69%

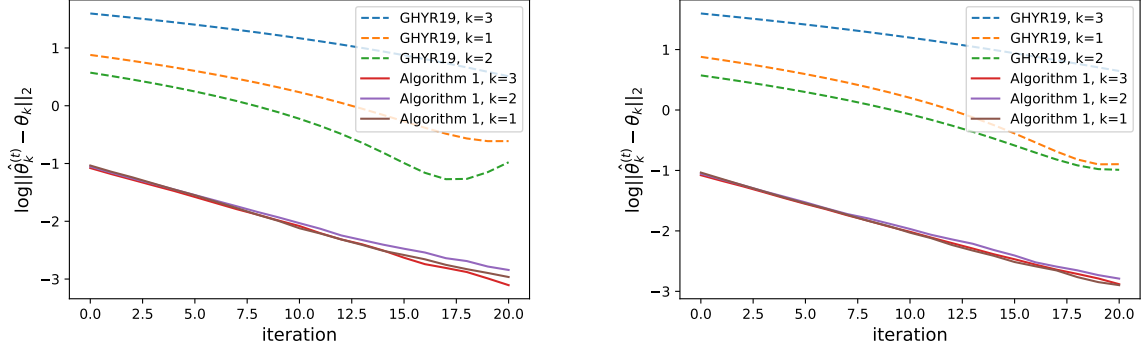
Figure 4.3 plots the estimation error  $\|\widehat{\theta}_k^{(t)} - \theta_k\|_2$  in log scale over 20 iterations under both algorithms. For both attack types, our algorithm achieves a much better estimation of the model parameters than [GHYR19, Algorithm 1]. In particular, under both Gaussian and Fit Flip attacks, the estimation errors of our Algorithm 1 for all clusters converge to zero exponentially fast. This confirms our theoretical prediction in Theorem 6. On the contrary, the estimation errors of [GHYR19, Algorithm 1] are much larger and decay at slow rates.

## Real data

We further compare our algorithm with [GHYR19, Algorithm 1] on the *set2.test.txt* from Yahoo Learning to Ranking dataset [Yah]. We treat the data as unsupervised and ignore the label for this experiment. The dataset contains 103174 query-document pair (denote as a query) with  $d = 595$  features whose values are normalized between 0 and 1. Each query

---

<sup>1</sup>In the numerical study of [GHYR19], the authors assign correct labels for 60% of the clients within each cluster. Thus, the initialization in our setting is even more favorable.



(a) Gaussian Attack

(b) Bit Flip Attack

**Figure 4.3:** Comparison of estimation errors  $\|\hat{\theta}_k^{(t)} - \theta_k\|_2$  under different Byzantine attacks.

is randomly sampled from the query logs of the Yahoo! search engine. Due to the privacy concern, how each feature is computed is not disclosed. In short, the features include not only document information such as the summary statistics of the number of words in various fields and the type of document but also measures of the link between the query and the document. These include but are not limited to textual similarity and topical similarity between the query and the document. A more detailed overview of the dataset can be found in [CC11].

We adopt a similar setup to the one from [GHYR19]. In particular, we first create a graph by forming an edge between query  $i$  and  $j$  if the Euclidean distance of their corresponding feature vectors is lower than 3.415. Then we apply the Louvain algorithm [BGLL08] to cluster the graph. We keep the  $K = 6$  largest clusters  $\{\mathcal{C}_k\}_{k=1}^K$  and remove the remaining clusters. The average of the feature vectors for queries in each cluster is viewed as the true cluster center  $\{\theta_k\}_{k=1}^K$ . We use the empirical distribution of feature vectors for queries within  $\mathcal{C}_k$  as  $\mathcal{D}_k$ .

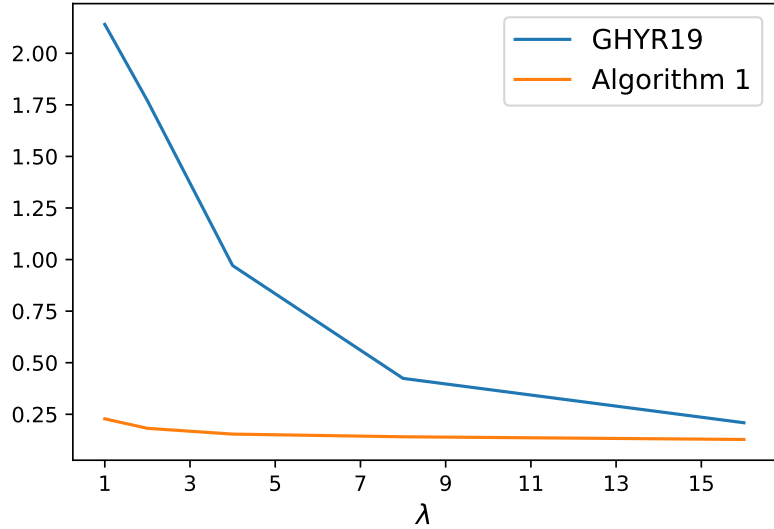
We generate 4000 clients such that the number of clients in each cluster is proportional to  $(0.2, 0.15, 0.2, 0.15, 0.15, 0.15)$ . To model the unbalanced data situation, 90% of the clients  $i$  have  $n_i$  i.i.d. generated from  $\max\{\text{Poisson}(\lambda), 1\}$  while the remaining 10% have 40 number of data points. At each communication round, each client in cluster  $k$  randomly draws  $n_i$  feature vectors from  $\mathcal{D}_k$ , where each feature vector can be viewed as a noisy observation of

the cluster center.

We adopt the same type of Byzantine attack used in [GHYR19]: the Byzantine adversary corrupts the vector sent by the client by adding  $P - 0.5\mathbf{1}_d$  vector where each coordinate of  $P$  is an independent Bernoulli(0.5) random variable. With such an attack, the Byzantine adversary can corrupt approximately half of the coordinates. Since each feature value is within 0 and 1, adding  $P - 0.5\mathbf{1}_d$  incurs significant perturbation.

Similar to Section 4.5.2, we randomly select 100 anchor clients and assume 5 of them are under Byzantine attack for Algorithm 1. We assign 60% of the clients within each cluster with the true cluster label and the other 40% to one another cluster for [GHYR19, Algorithm 1].

To evaluate the performance, we compare the sum of the deviations to the true cluster centers, that is,  $\sum_{k=1}^K \|\hat{\theta}_k^{(T)} - \theta_k\|_2$  where  $T = 20$ . Figure 4.4 summarizes the result for various values of  $\lambda$ . As shown in the figure, when  $\lambda$  is large, i.e., the number of data points



**Figure 4.4:** Comparison of estimation error  $\sum_{k=1}^K \|\hat{\theta}_k - \theta_k\|_2$  for various  $\lambda$

on each client is large, our algorithm performs as well as [GHYR19, Algorithm 1]. When the number of data points on most clients is getting smaller and smaller, our algorithm performs significantly better than [GHYR19, Algorithm 1].

## 4.6 Further Discussions on SDP Relaxations

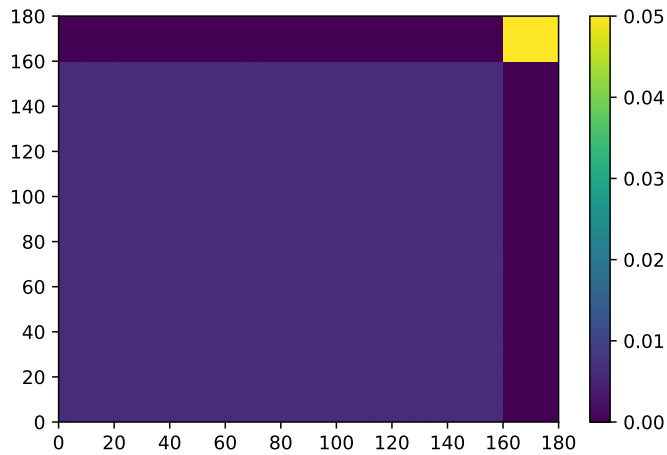
SDP has been extensively used for clustering with outliers [CL15, MVW17, SCV17, CSV17, BK20, LLL<sup>+</sup>20, DKK<sup>+</sup>22]. Most of the previous works impose distributional assumptions on input data and prove SDP works with a high probability. For example, [BK20] utilizes the sum of squares (SOS) proofs and proposes an SDP to robustly cluster data generated from sub-Gaussian mixture models with equal weights. From a high level, their idea is based on moment matching and thus crucially requires distributional assumptions.

There are only a handful of recent works that study SDP clustering in deterministic settings without distributional assumptions. In a distribution-free setting, an SDP-related algorithm that can be used for clustering with the presence of outliers is proposed in [CSV17, Algorithm 4]. However, such an algorithm may output more than the actual number of clusters. In particular, clients from the same true cluster can possibly be divided into multiple sub-clusters under [CSV17, Algorithm 4]. As a consequence, the true cluster structure cannot be recovered. In addition, when the size of some output cluster of [CSV17, Algorithm 4] is smaller than the number of Byzantine clients, Byzantine clients can corrupt an entire output cluster, preventing us from learning the cluster center.

Another SDP-related algorithm proposed in a distribution-free setting is Peng and Wei’s relaxation [LLL<sup>+</sup>20] which uses the pairwise distance matrix and the cluster number as input. It has been shown to solve the  $k$ -means clustering under certain deterministic conditions when there is no outlier. However, when there exist Byzantine clients who are far away from all good clients, the solution of Peng and Wei’s relaxation groups the good clients as a single cluster and the outliers as the other one. Hence, Peng and Wei’s relaxation fails to recover the true cluster in the presence of outliers.

Now we provide a numerical example to illustrate the aforementioned limitation of Peng and Wei’s relaxation. Consider two clusters, each of which contains 80 good clients. Each good client from cluster  $k$  outputs a Gaussian random vector  $\mathcal{N}(\mu_k, \mathbf{I}_d)$  where  $\mu_1 = 0$  and  $\mu_2 = 3$ , for  $k = 1, 2$ . In addition, there are 20 clients under Byzantine attack. Each

Byzantine client outputs a Gaussian random vector  $\mathcal{N}(20, \mathbf{I}_d)$ . Peng and Wei’s relaxation is then applied to the pairwise distance matrix of the output from all clients, and the solution is presented in Figure 4.5. Here, we denote index 0 to 79 for clients in cluster 1 and index 80 to 159 for clients in cluster 2. The remaining 20 indices represent Byzantine clients. Color represents the value of each entry. In Figure 4.5, it can be seen that rows and columns of good clients form a block matrix with identical entry values. As a result, the solution of Peng and Wei’s relaxation fails to recover the true cluster structure.



**Figure 4.5:** Solution to Peng and Wei’s relaxation with the presence of outliers. Color represents the value of the entry. Index 0 to 79 denotes good clients from cluster 1, and index 80 to 159 denotes good clients from cluster 2. The remaining 20 indices represent Byzantine clients.

## 4.7 Crucial Role of Initial Label in [GHYR19]

A key step of [GHYR19, Algorithm 1] is the trimmed  $K$ -mean clustering. It requires an initial clustering as the input and iteratively updates the clustering. In particular, at each round, given the estimated cluster labels at the previous round, it updates the cluster labels as follows. First, it computes the geometric median of the estimates from clients with the estimated label  $k$ , denoted by  $\tilde{\theta}_k$ . Then, the algorithm updates the cluster center by taking

the average of the local estimates falling inside a ball centered at  $\tilde{\theta}_k$ . Lastly, the algorithm updates the cluster label of each client by assigning the label corresponding to the nearest cluster center.

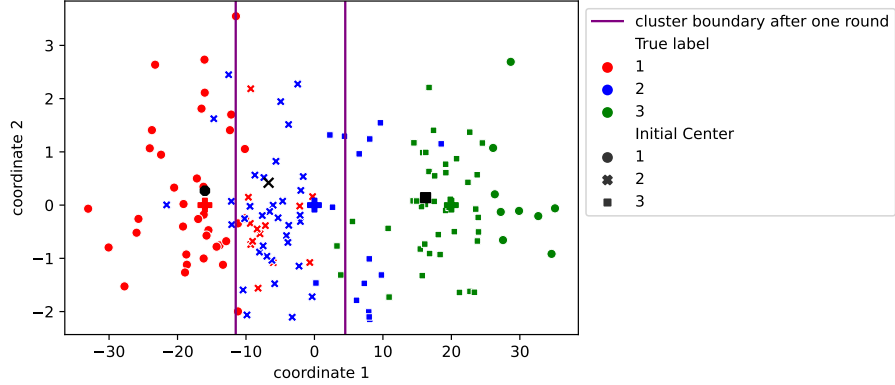
Note that the performance of [GHYR19, Algorithm 1] crucially relies on  $\tilde{\theta}_k$ . If  $\tilde{\theta}_k$  is far from the true cluster center  $\theta_k$ , a large fraction of the included clients to compute the trimmed mean can come from other clusters. Thus, the updated cluster center could be far from  $\theta_k$ . As a result, the updated clustering label suffers from a high misclassification error.

Reasons such as noisy local estimates due to insufficient data among clients or inaccurate initial labels can all lead to a bad  $\tilde{\theta}_k$ . Here, we use a simple mixture-of-Gaussian example to illustrate the influence of one specific initialization scheme on  $\tilde{\theta}_k$ . In particular, under this scheme, even the overall accuracy of the initial label is high, the obtained  $\tilde{\theta}_k$  is still far from  $\theta_k$ .

We generate 150 local estimates evenly from three normal distributions  $\mathcal{N}(-16, 8)$ ,  $\mathcal{N}(0, 8)$  and  $\mathcal{N}(20, 8)$ . Figure 4.6 shows the initial clustering and the clustering after the first round of the trimmed  $K$ -mean clustering from [GHYR19]. Note that if we were plotting all the points in one dimension, the figure would be cluttered with overlapping points. Thus, for better visualization, we add an auxiliary  $\mathcal{N}(0, 1)$  for the data point as coordinate 2.

In Figure 4.6, we use color to denote the true cluster label: red for  $\mathcal{C}_1$ , blue for  $\mathcal{C}_2$ , and green for  $\mathcal{C}_3$ . The large colored crosses represent the true cluster centers  $\theta_k, k \in \{1, 2, 3\}$ . Different shapes are used to denote the initial cluster label: circle for  $\hat{\mathcal{C}}_1$ , x for  $\hat{\mathcal{C}}_2$ , and square for  $\hat{\mathcal{C}}_3$ . Here, within each true cluster, we only assign incorrect labels to the clients with large local estimates, represented by red x, blue square and green circles. More than 70% of clients in each cluster are labelled accurately, illustrated by a large number of red circles, blue x, and green squares. In short, most clients in  $\hat{\mathcal{C}}_k$  are from  $\mathcal{C}_k$ .

The black shape points represent  $\tilde{\theta}_k, k = 1, 2, 3$ . It can be seen that both  $\tilde{\theta}_2$  and  $\tilde{\theta}_3$  are smaller than the corresponding true cluster centers (colored crosses). This is because under the above initialization scheme, though  $\hat{\mathcal{C}}_k$  and  $\mathcal{C}_k$  has a large overlap, when  $k \in \{2, 3\}$  the



**Figure 4.6:** An example illustrating the limitation of [GHYR19] in handling unbalanced data. Colors represent the true cluster labels. Crosses of the corresponding color denote the true cluster centers. Shape denotes the initial label of each client where we assign true labels to more than 70% of clients in each cluster. Black shape points denote  $\{\tilde{\theta}_k\}$ . The second coordinate is added only for illustration purposes. The purple vertical lines represent the cluster boundaries after one round of the algorithm.

estimates of clients in  $\widehat{\mathcal{C}}_k - \mathcal{C}_k$  is much smaller than the estimates of clients in  $\mathcal{C}_k - \widehat{\mathcal{C}}_k$  where  $A - B$  denotes the collection of clients in cluster  $A$  but not in  $B$ . Therefore, the geometric median  $\tilde{\theta}_k$  is much smaller than  $\theta_k$  when  $k \in \{2, 3\}$ . When  $k = 1$ , estimates of clients in  $\widehat{\mathcal{C}}_1 - \mathcal{C}_1$  and  $\mathcal{C}_1 - \widehat{\mathcal{C}}_1$  are both large. As a result,  $\tilde{\theta}_1$  does not deviate much from  $\theta_1$ .

After running the trimmed mean and the re-labeling step, the updated labels of clients are illustrated by the regions separated by vertical purple lines. In particular, clients in the left region form  $\widehat{\mathcal{C}}_1$ , clients in the middle form  $\widehat{\mathcal{C}}_2$ , and the remaining form  $\widehat{\mathcal{C}}_3$ . We see a relatively small misclassification error from  $\widehat{\mathcal{C}}_1$  and  $\widehat{\mathcal{C}}_3$ , indicated by a small number of non-red points in the left region and a small number of non-green points in the right region. However, we see a large number of red points in  $\widehat{\mathcal{C}}_2$ , indicating a high misclassification error in  $\widehat{\mathcal{C}}_2$ . As we explained before, this high misclassification error in  $\widehat{\mathcal{C}}_2$  is due to the fact that  $\tilde{\theta}_2$  is far from  $\theta_2$ .

# Chapter 5

## Proofs in Chapter 4

In this chapter, we present the proof of the theoretical guarantees of each step of Algorithm 1.

### 5.1 Proof of Theorem 4

The proof of Theorem 4 follows from the two lemmas below. In short, Lemma 5.1.1 shows that under condition (5.1), there exist choices of tuning parameters  $\lambda$  and  $\gamma$  in the specific range given in (5.2) and (5.3). Lemma 5.1.2 shows that as long as  $\lambda$  and  $\gamma$  fall into the ranges in (5.2) and (5.3), the output of Algorithm 2 is of the desired form (4.4).

**Lemma 5.1.1.** *If  $\epsilon < \frac{\rho}{C\sqrt{K}}$  and*

$$a_{\text{in}} > \frac{1}{2} + \frac{5}{C\sqrt{K}} + \left[ \max \left\{ \frac{1}{2} + \frac{1}{2\rho}, 2 \right\} + \frac{2\epsilon}{\rho} \right] a_{\text{out}} + \frac{2\sqrt{2}+1}{2} \left( 1 + \frac{\epsilon}{\rho} \right) \frac{\epsilon}{\rho} \quad (5.1)$$

*for some large constant  $C$  and  $C'$ . Then there exist  $\lambda$  and  $\gamma$  such that*

$$\frac{4\epsilon}{\rho} + 2a_{\text{out}} < \lambda < \frac{\kappa}{1 + \epsilon/\rho} \quad (5.2)$$

*and*

$$\frac{\rho + 2C\sqrt{K}/5 + CK\epsilon/5}{\rho - CK\epsilon/5} \epsilon m < \gamma < \left( a_{\text{in}} - \left( 1 + \frac{\epsilon}{\rho} \right) \lambda \right) \rho n - \frac{2}{m}, \quad (5.3)$$

*where*

$$\kappa \triangleq \min \left\{ a_{\text{in}} - \left( \frac{\rho + 2C\sqrt{K}/5 + CK\epsilon/5}{\rho - CK\epsilon/5} \right) \frac{\epsilon}{\rho} - \frac{2}{\rho m^2}, 2a_{\text{in}} - 1 - a_{\text{out}} \frac{1 - \rho - \epsilon}{\rho} - \frac{10}{C\sqrt{K}} \right\}.$$

**Lemma 5.1.2.** *With tuning parameters  $\lambda$  and  $\gamma$  satisfying (5.2) and (5.3), the output of (4.3) must be of the form (4.4).*

### 5.1.1 Proof of Lemma 5.1.1

Here, we show that under (5.1), there must exist  $0 < \lambda < 1$  and  $\gamma > 0$  such that (5.2) and (5.3) hold.

To begin with, we show the lower bound of  $\lambda$  in (5.2) is below 1. Since  $a_{\text{in}} < 1$ , (5.1) requires

$$2a_{\text{out}} < \frac{1}{2} - \frac{(2\sqrt{2} + 1)\epsilon}{\rho}.$$

Hence, we have

$$\frac{(2\sqrt{2} + 1)\epsilon}{\rho} + 2a_{\text{out}} < 1.$$

To complete the validation of (5.2), it remains to show

$$\left(1 + \frac{\epsilon}{\rho}\right) \left(\frac{(2\sqrt{2} + 1)\epsilon}{\rho} + 2a_{\text{out}}\right) < a_{\text{in}} - \left(\frac{\rho + 2C\sqrt{K}/5 + C\sqrt{K}\epsilon/5}{\rho - CK\epsilon/5}\right) \frac{\epsilon}{\rho} - \frac{2}{\rho m^2}, \quad (5.4)$$

$$\left(1 + \frac{\epsilon}{\rho}\right) \left(\frac{(2\sqrt{2} + 1)\epsilon}{\rho} + 2a_{\text{out}}\right) < 2a_{\text{in}} - 1 - a_{\text{out}} \left(\frac{1}{\rho} - 1 - \epsilon\right) - \frac{10}{C\sqrt{K}}. \quad (5.5)$$

Since  $\epsilon < \frac{\rho}{KC'}$ , we have  $\frac{\epsilon}{\rho} < \frac{1}{C'K}$  and  $CK\epsilon < \frac{C\rho}{C'} \leq \rho$  for  $C' > C$ .

Hence,

$$\begin{aligned} \left(\frac{\rho + 2C\sqrt{K}/5 + CK\epsilon/5}{\rho - CK\epsilon/5}\right) \frac{\epsilon}{\rho} &\leq \left(\frac{5}{4} + \frac{C\sqrt{K}}{2\rho} + \frac{CK\epsilon}{4\rho}\right) \frac{\epsilon}{\rho} \\ &\leq \frac{5}{4C'K} + \frac{C}{2C'\sqrt{K}\rho} + \frac{C}{4(C')^2K} \\ &\leq \frac{1}{2} + \frac{1}{C'K} \left(\frac{5}{4} + \frac{C}{4C'}\right) \\ &\leq \frac{1}{2} + \frac{5}{C\sqrt{K}} - \frac{2}{\rho m^2} \end{aligned} \quad (5.6)$$

where the third inequality holds when  $C' > \frac{C}{\sqrt{K}\rho}$  and the last inequality holds by  $C' > \frac{3}{10\sqrt{K}}C$ .

Under (5.1), we have

$$\begin{aligned} a_{\text{in}} &> \frac{1}{2} + \frac{5}{C\sqrt{K}} + \left(2 + \frac{2\epsilon}{\rho}\right) a_{\text{out}} + \frac{(2\sqrt{2}+1)\epsilon}{\rho} \left(1 + \frac{\epsilon}{\rho}\right) \\ &> \left(\frac{\rho + 2C\sqrt{K}/5 + CK\epsilon/5}{\rho - CK\epsilon/5}\right) \frac{\epsilon}{\rho} + \left(1 + \frac{\epsilon}{\rho}\right) \left(2a_{\text{out}} + \frac{(2\sqrt{2}+1)\epsilon}{\rho}\right) + \frac{2}{\rho m^2}, \end{aligned}$$

where the last inequality holds by (5.6).

As a result, (5.4) holds.

Similarly, (5.1) yields

$$a_{\text{in}} > \frac{1}{2} + \frac{5}{C\sqrt{K}} + \left(\frac{1}{2} + \frac{1}{2\rho} + 2\frac{\epsilon}{\rho}\right) a_{\text{out}} + \frac{2\sqrt{2}+1}{2} \left(1 + \frac{\epsilon}{\rho}\right) \frac{\epsilon}{\rho},$$

which is equivalent as (5.5).

To show (5.3) is valid, we need

$$\frac{\rho + 2C\sqrt{K}/5 + CK\epsilon/5}{\rho - CK\epsilon/5} \epsilon < \left(a_{\text{in}} - \left(1 + \frac{\epsilon}{\rho}\right) \lambda\right) \rho - \frac{2}{m^2},$$

which is equivalent as

$$\lambda > \frac{1}{1 + \epsilon/\rho} \left(a_{\text{in}} - \left(\frac{\rho + 2C\sqrt{K}/5 + CK\epsilon/5}{\rho - CK\epsilon/5}\right) \frac{\epsilon}{\rho} - \frac{2}{\rho m^2}\right).$$

This is guaranteed by (5.2) which has been validated above.

### 5.1.2 Proof of Lemma 5.1.2

**Notation** Define  $I$  as the set of non-Byzantine clients and  $I^c$  as the set of Byzantine clients. Throughout the section, for any matrix  $\mathbf{M}$ , we use  $\mathbf{M}_{A,B}$  to denote the submatrix of  $\mathbf{M}$  whose rows correspond to clients in set  $A$  and columns correspond to clients in set  $B$ . For the ease of presentation, we use  $\mathbf{M}_{ij}$  to denote  $\mathbf{M}_{\mathcal{C}_i, \mathcal{C}_j}$  for any  $i, j$ . Denote  $q = \epsilon m$ .

We first construct a candidate solution  $\mathbf{X}^*$ . Then we show that any solution to (4.3) must satisfy  $\widehat{\mathbf{X}}_{I,I} = \mathbf{X}_{I,I}^*$ .

Through permutation of client index, we can organize clients so that client with index  $\sum_{j=1}^k m_j + 1$  to  $\sum_{j=1}^{k+1} m_j$  belongs to cluster  $k + 1$  and client with index  $m - q + 1$  to  $m$  are Byzantine. Capturing this intuition, we write the corresponding adjacency matrix  $\mathbf{A}$  as

$$\mathbf{A} = \mathbf{P} \begin{pmatrix} \mathbf{K} & \mathbf{Z} \\ \mathbf{Z}^\top & \mathbf{W} \end{pmatrix} \mathbf{P}^\top, \quad (5.7)$$

where  $\mathbf{P} \in \mathbb{R}^{m \times m}$  is some unknown permutation matrix representing the client index permutation,  $\mathbf{K} = \mathbf{A}_{I,I}$  captures the connectivity among non-Byzantine clients,  $\mathbf{Z} = \mathbf{A}_{I,I^c} \in \{0, 1\}^{m \times q}$  is some arbitrary matrix and  $\mathbf{W} = \mathbf{A}_{I^c,I^c} \in \{0, 1\}^{q \times q}$  is some arbitrary symmetric matrix with diagonal entries 0;

WLOG, we assume  $\mathbf{P} = \mathbf{I}$  for the analysis.

**Primal solution construction** Denote

$$\gamma \mathbf{I} + \lambda \mathbf{J} - \mathbf{A} = \begin{pmatrix} \gamma \mathbf{I} + \lambda \mathbf{J} - \mathbf{K}_{11} & \cdots & \lambda \mathbf{J} - \mathbf{K}_{1K} & \tilde{\mathbf{Z}}_1 \\ \vdots & \ddots & \vdots & \vdots \\ \lambda \mathbf{J} - \mathbf{K}_{1K}^\top & \cdots & \gamma \mathbf{I} + \lambda \mathbf{J} - \mathbf{K}_{KK} & \tilde{\mathbf{Z}}_K \\ \tilde{\mathbf{Z}}_1^\top & \cdots & \tilde{\mathbf{Z}}_K^\top & \tilde{\mathbf{W}} \end{pmatrix}, \quad (5.8)$$

where  $\tilde{\mathbf{Z}}_i = \lambda \mathbf{J} - \mathbf{Z}_i$  and  $\tilde{\mathbf{W}} = \gamma \mathbf{I} + \lambda \mathbf{J} - \mathbf{W}$ .

Now we construct a candidate solution to (4.3):

$$\mathbf{X}^* = \mathbf{V} \mathbf{V}^\top = \begin{pmatrix} \mathbf{J}_{m_1} & \cdots & \mathbf{0} & \mathbf{1}_{m_1} x_1^\top \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{J}_{m_K} & \mathbf{1}_{m_K} x_K^\top \\ x_1 \mathbf{1}_{m_1}^\top & \cdots & x_K \mathbf{1}_{m_K}^\top & \sum_{i=1}^K x_i x_i^\top \end{pmatrix},$$

where

$$\mathbf{V} = \begin{pmatrix} \mathbf{1}_{m_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{1}_{m_K} \\ x_1 & \cdots & x_K \end{pmatrix} \in \mathbb{R}^{m \times K}. \quad (5.9)$$

To see what  $\{x_i\}_{i=1}^K$  is, note that the objective value of (4.3) satisfies

$$\begin{aligned} \langle \mathbf{X}^*, \mathbf{A} - \lambda \mathbf{J} - \gamma \mathbf{I} \rangle &= \text{Tr}(\mathbf{X}^* (\mathbf{A} - \lambda \mathbf{J} - \gamma \mathbf{I})) \\ &= \sum_{i=1}^K \langle \mathbf{K}_{ii} - \lambda \mathbf{J} - \gamma \mathbf{I}, \mathbf{J} \rangle - \sum_{i=1}^K \left[ 2 \langle x_i, \tilde{\mathbf{Z}}_i^\top \mathbf{1}_{m_i} \rangle + x_i^\top \tilde{\mathbf{W}} x_i \right]. \end{aligned}$$

Therefore, in order for  $\mathbf{X}^*$  to maximize  $\langle \mathbf{X}^*, \mathbf{A} - \lambda \mathbf{J} - \gamma \mathbf{I} \rangle$ , we need  $x_1, x_2, \dots, x_K$  to be the solution of the following program (5.10).

$$\begin{aligned} \min_{y_i \in \mathbb{R}^q, i=1,2,\dots,K} \quad & \sum_{i=1}^K \langle y_i, \tilde{\mathbf{Z}}_i^\top \mathbf{1}_{m_i} \rangle + \frac{1}{2} \sum_{i=1}^K y_i^\top \tilde{\mathbf{W}} y_i \\ \text{subject to} \quad & y_i \geq 0, i \in [K] \\ & \sum_{i=1}^K y_i (e_j e_j^\top) y_i \leq 1, \forall j \in [q]. \end{aligned} \quad (5.10)$$

The two constraints of (5.10) are necessary for  $\mathbf{X}^*$  to be feasible.

Under some mild conditions, the solution  $x_1, \dots, x_K$  exist uniquely.

**Lemma 5.1.3.** [CL15, Lemma 6.9] *If  $\gamma \geq 2q$  and  $0 < \lambda < 1$ , then there is a unique solution  $\{x_i\}_{i=1}^K$  to (5.10). Moreover, there are unique vectors  $\{\alpha_i\}_{i=1}^K, \xi \geq 0 \in \mathbb{R}^q$  such that*

$$\tilde{\mathbf{W}} x_i + \tilde{\mathbf{Z}}_i^\top \mathbf{1}_{m_i} = \alpha_i - \text{diag}\{\xi\} x_i, i \in [K] \quad (5.11)$$

$$\xi_j \left( 1 - \sum_{i=1}^K x_i^\top (e_j e_j^\top) x_i \right) = 0, j \in [q] \quad (5.12)$$

$$\langle x_i, \alpha_i \rangle = 0, i \in [K]. \quad (5.13)$$

Intuitively,  $\{\alpha_i\}_{i=1}^K$  and  $\xi$  are dual variables to the constraints of (5.10) in order, and

(5.11)–(5.13) are the KKT conditions of (5.10).

**Sufficient Condition for the optimality of  $\mathbf{X}^*$**  The following lemma provides a sufficient condition to ensure  $\mathbf{X}^*$  is optimal.

**Lemma 5.1.4.** [CL15, Lemma 6.10] *Let  $\alpha_i$  and  $\xi$  are defined as in Lemma 5.1.3. If there exists symmetric matrices  $\mathbf{S} \in \mathbb{R}^{m \times m}$  and  $\Psi_{ii} \in \mathbb{R}^{m_i \times m_i}$  and matrices  $\Phi_{ij} \in \mathbb{R}^{m_i \times m_j}$  such that  $\mathbf{S}$  equals*

$$\begin{pmatrix} \gamma \mathbf{I}_{m_1} + \lambda \mathbf{J}_{m_1} - \mathbf{K}_{11} + \Psi_{11} & \cdots & \lambda \mathbf{J}_{m_1, m_K} - \mathbf{K}_{1K} - \Phi_{1K} & \tilde{\mathbf{Z}}_1 - \frac{1}{m_1} \mathbf{1}_{m_1} \alpha_1^\top \\ \vdots & \ddots & \vdots & \vdots \\ \lambda \mathbf{J}_{m_K, m_1} - \mathbf{K}_{1K}^\top - \Phi_{1K}^\top & \cdots & \gamma \mathbf{I}_{m_K} + \lambda \mathbf{J}_{m_K} - \mathbf{K}_{KK} + \Psi_{KK} & \tilde{\mathbf{Z}}_K - \frac{1}{m_K} \mathbf{1}_{m_K} \alpha_K^\top \\ \tilde{\mathbf{Z}}_1^\top - \frac{1}{m_1} \alpha_1 \mathbf{1}_{m_1}^\top & \cdots & \tilde{\mathbf{Z}}_K^\top - \frac{1}{m_K} \alpha_K \mathbf{1}^\top & \tilde{\mathbf{W}} + \text{diag}\{\xi\} \end{pmatrix}, \quad (5.14)$$

$\Psi_{ii} > 0$ ,  $\Phi_{ij} > 0$ ,  $\mathbf{S}\mathbf{V} = \mathbf{0}$  and  $\mathbf{S} \succeq \mathbf{0}$ , then any minimizer  $\hat{\mathbf{X}}$  to (4.3) must be of the form

$$\hat{\mathbf{X}} = \mathbf{X}^* + \mathbf{H}, \quad (5.15)$$

where  $\mathbf{H}$  is arbitrary symmetric matrix with  $\mathbf{H}_{I,I} = 0$ . Moreover,  $\mathbf{X}^*$  is a solution to (4.3).

It suffices to construct  $\Psi_{ii}$  and  $\Phi_{ij}$  to satisfy the conditions in Lemma 5.1.4 with our choice of  $\gamma$  and  $\lambda$ .

To motivate the construction of  $\Psi_{ii}$  and  $\Phi_{ij}$ , we first verify the condition  $\mathbf{S}\mathbf{V} = \mathbf{0}$ .

**Proof of  $\mathbf{S}\mathbf{V} = \mathbf{0}$ :** For the ease of presentation, denote  $I^c$  as the cluster  $K + 1$ . Denote  $\mathbf{S}_{i,*} \in \mathbb{R}^{m_i \times m}$  as the rows of  $\mathbf{S}$  corresponding to clients in cluster  $i$  and  $v_{jr}$  as the coordinates of  $v_j$  corresponding to cluster  $\mathcal{C}_r$ .

To prove  $\mathbf{S}\mathbf{V} = \mathbf{0}$ , it suffices to show

$$[\mathbf{S}\mathbf{V}]_{ij} = \mathbf{S}_{i,*} v_j = \sum_{r=1}^{K+1} \mathbf{S}_{ir} v_{jr} = 0, \forall j \in \{1, \dots, K\}, i \in [K + 1]. \quad (5.16)$$

From (5.14), we have

$$\mathbf{S}_{ir} = \begin{cases} \gamma \mathbf{I}_{m_i} + \lambda \mathbf{J}_{m_i} - \mathbf{K}_{ii} + \boldsymbol{\Psi}_{ii} & r = i, i \in [K] \\ \lambda \mathbf{J}_{m_i, m_r} - \mathbf{K}_{ir} - \boldsymbol{\Phi}_{ir} & r > i, i, r \in [K] \\ \lambda \mathbf{J}_{m_i, m_r} - \mathbf{K}_{ri}^\top - \boldsymbol{\Phi}_{ri}^\top & r < i, i, r \in [K] \\ \tilde{\mathbf{Z}}_i - \frac{1}{m_i} \mathbf{1}_{m_i} \alpha_i^\top & i \in [K], r = K + 1 \\ \tilde{\mathbf{Z}}_r^\top - \frac{1}{m_r} \alpha_r \mathbf{1}_{m_r}^\top & i = K + 1, r \in [K] \\ \widetilde{\mathbf{W}} + \text{diag} \{ \xi \} & i = r = K + 1 \end{cases}$$

From the definition of  $\mathbf{V}$  in (5.9), we have

$$v_{jr} = \begin{cases} \mathbf{1}_{m_r} & r = j \\ 0 & r \neq j \\ x_j & r = K + 1 \end{cases}$$

Therefore, (5.16) is equivalent to

$$\gamma \mathbf{1}_{m_i} + \lambda m_i \mathbf{1}_{m_i} - \mathbf{K}_{ii} \mathbf{1}_{m_i} + \boldsymbol{\Psi}_{ii} \mathbf{1}_{m_i} + \tilde{\mathbf{Z}}_i x_i = 0, \quad i = j \in [K] \quad (5.17)$$

$$\lambda m_j \mathbf{1}_{m_i} - \mathbf{K}_{ij} \mathbf{1}_{m_j} - \boldsymbol{\Phi}_{ij} \mathbf{1}_{m_j} + \left( \tilde{\mathbf{Z}}_i - \frac{1}{m_i} \mathbf{1}_{m_i} \alpha_i^\top \right) x_j = 0, \quad i < j, i, j \in [K] \quad (5.18)$$

$$\lambda m_j \mathbf{1}_{m_i} - \mathbf{K}_{ji}^\top \mathbf{1}_{m_j} - \boldsymbol{\Phi}_{ji}^\top \mathbf{1}_{m_j} + \left( \tilde{\mathbf{Z}}_i - \frac{1}{m_i} \mathbf{1}_{m_i} \alpha_i^\top \right) x_j = 0, \quad i > j, i, j \in [K] \quad (5.19)$$

$$\tilde{\mathbf{Z}}_j^\top \mathbf{1}_{m_j} - \alpha_j + \left( \widetilde{\mathbf{W}} + \text{diag} \{ \xi \} \right) x_j = 0, \quad i = K + 1, j \in [K]. \quad (5.20)$$

Note that (5.20) automatically holds by (5.11).

To ensure (5.17) holds, we construct

$$\boldsymbol{\Psi}_{ii} = \text{diag} \left\{ \mathbf{K}_{ii} \mathbf{1}_{m_i} - \tilde{\mathbf{Z}}_i x_i \right\} - (\lambda m_i + \gamma + \delta) \mathbf{I}_{m_i} + \frac{\delta}{m_i} \mathbf{J}_{m_i}, \quad (5.21)$$

for any  $0 < \delta < 2$ .

Here the  $\frac{\delta}{m_i} \mathbf{J}_{m_i}$  is imposed to ensure the strict positivity of off-diagonal entries of  $\Psi_{ii}$ . Now we construct  $\Phi$  to satisfy (5.18) and (5.19). Note that (5.18) is equivalent as

$$\Phi_{ij} \mathbf{1}_{m_j} = \lambda m_j \mathbf{1}_{m_i} - \mathbf{K}_{ij} \mathbf{1}_{m_j} + \left( \tilde{\mathbf{Z}}_i - \frac{1}{m_i} \mathbf{1}_{m_i} \alpha_i^\top \right) x_j \triangleq g_{ij}, \forall i < j. \quad (5.22)$$

Similarly, (5.19) is equivalent as

$$\Phi_{ij}^\top \mathbf{1}_{m_i} = \lambda m_i \mathbf{1}_{m_j} - \mathbf{K}_{ij}^\top \mathbf{1}_{m_i} + \left( \tilde{\mathbf{Z}}_j - \frac{1}{m_j} \mathbf{1}_{m_j} \alpha_j^\top \right) x_i \triangleq g_{ji}, \forall i < j. \quad (5.23)$$

Inspired by [CL15], we construct the following

$$\Phi_{ij} = \frac{1}{m_j} g_{ij} \mathbf{1}_{m_j}^\top + \frac{1}{m_i} \mathbf{1}_{m_i} g_{ji}^\top - \frac{v_{ij}}{m_i m_j} \mathbf{J}_{m_i, m_j}, \quad (5.24)$$

with

$$v_{ij} = \mathbf{1}_{m_i}^\top g_{ij} = g_{ji}^\top \mathbf{1}_{m_i}.$$

Now we show  $\Phi_{ij}$  satisfies (5.22) and (5.23), which completes the verification of  $\mathbf{S}\mathbf{V} = \mathbf{0}$ .

From (5.24), we have

$$\Phi_{ij} \mathbf{1}_{m_j} = g_{ij} + \frac{1}{m_i} \mathbf{1}_{m_i} g_{ji}^\top \mathbf{1}_{m_i} - \frac{v_{ij}}{m_i} \mathbf{1}_{m_i} = g_{ij}, \quad (5.25)$$

where the last equality follows from the definition of  $v_{ij}$ .

Similarly, we can get  $\Phi_{ij}^\top \mathbf{1}_{m_i} = g_{ji}$ .

**Proof of  $\Psi_{ii} > 0$ :** From the definition of  $a_{\text{in}}$  and  $a_{\text{out}}$ , for any  $i \in [K]$ , we have

$$\begin{aligned} [\mathbf{K}_{ii} \mathbf{1}](s) &\geq a_{\text{in}} m_i, \forall s \in [m_i] \\ [\mathbf{K}_{ij} \mathbf{1}](s) &\leq a_{\text{out}} m_j, \forall s \in [m_i], j \neq i. \end{aligned} \quad (5.26)$$

From (5.8), we know each entry of  $\tilde{\mathbf{Z}}_i$  is either  $\lambda$  and  $\lambda - 1$ . Since  $x_i$  is the solution to

(5.10), each coordinate of  $x_i$  is between 0 and 1. In together with  $\lambda \leq 1$ , we have

$$\tilde{\mathbf{Z}}_i x_i \leq \lambda q.$$

Hence, each diagonal entry of  $\Psi_{ii}$  is at least  $a_{\text{in}} m_i - \lambda q - \lambda m_i - \gamma - \left(1 - \frac{1}{m_i}\right) \delta$ .

As a result,

$$\begin{aligned} & a_{\text{in}} m_i - \lambda q - \lambda m_i - \gamma - \left(1 - \frac{1}{m_i}\right) \delta \\ & > a_{\text{in}} \rho_i m - \lambda \epsilon m - \lambda \rho_i m - \gamma - \delta \\ & \geq a_{\text{in}} \rho m - \lambda \epsilon m - \lambda \rho m - \gamma - \frac{2}{m} \\ & > a_{\text{in}} \rho m - \lambda \epsilon m - \lambda \rho m - \frac{2}{m} - \left(a_{\text{in}} - \left(1 + \frac{\epsilon}{\rho}\right) \lambda - \frac{2}{\rho m^2}\right) \rho m \\ & > 0 \end{aligned}$$

where the second inequality hold by  $a_{\text{in}} \geq \lambda$  implied by (5.2) and the last inequality follows the bounds of  $\gamma$  in (5.3) and the fact that  $\delta < 2/m$ .

**Proof of  $\Phi_{ij} > 0$ :** To show  $\Phi_{ij} > 0$ , we utilize the following proposition.

**Proposition 5.1.5.** *If  $1 > \lambda > \frac{(1+2\sqrt{2})\epsilon}{\rho} + 2a_{\text{out}}$ , we have  $\Phi_{ij} > 0$  for all  $i \neq j$ .*

Any  $\lambda$  satisfying (5.2) ensures the conditions of Proposition 5.1.5 hold. Hence  $\Phi_{ij} > 0$ .

**Proof of  $\mathbf{S} \succeq \mathbf{0}$ :** Since  $\mathbf{S}\mathbf{V} = \mathbf{0}$ ,  $\mathbf{S}$  has rank at most  $m - K$ . To prove  $\mathbf{S}$  is PSD, it is sufficient to show  $\lambda_{m-K}(\mathbf{S}) > 0$ . One natural approach is to show the diagonal dominance of  $\mathbf{S}$ , i.e.,  $\mathbf{S}(i, j) \geq \sum_{j \neq i} |\mathbf{S}(i, j)|, \forall i$ . However, it is challenging to directly verify the diagonal dominance of  $\mathbf{S}$  due to the impact of Byzantine clients through  $\tilde{\mathbf{Z}}_i$  and  $\tilde{\mathbf{W}}$ .

To overcome the challenge and limit the influence of Byzantine clients through  $\Phi_{ij}$ , we want to transform the analysis of  $\mathbf{S}$  to some matrix  $\hat{\mathbf{S}}$  whose spectrum is under better control. In particular,  $\hat{\mathbf{S}}$  should not be influenced by  $\Phi_{ij}$ .

As the first step, we introduce  $\tilde{\mathbf{S}}$  which is inspired by [CL15] to facilitate the verification of the diagonal dominance of the good clients:

$$\tilde{\mathbf{S}} = \mathbf{S} + \mathbf{\Gamma}_1 + \mathbf{\Gamma}_2, \quad (5.27)$$

where

$$\mathbf{\Gamma}_1 = \begin{pmatrix} \left(1 - \lambda - \frac{\delta}{m_1}\right) \mathbf{J}_{m_1} & \cdots & -\lambda \mathbf{J}_{m_1, m_K} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ -\lambda \mathbf{J}_{m_K, m_1} & \cdots & \left(1 - \lambda - \frac{\delta}{m_K}\right) \mathbf{J}_{m_K} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

and

$$\mathbf{\Gamma}_2 = \begin{pmatrix} \mathbf{0} & \cdots & \mathbf{\Phi}_{1K} & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{\Phi}_{1K}^\top & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Intuitively,  $\mathbf{\Gamma}_1$  is designed to facilitate the verification of diagonal dominance easier, and  $\mathbf{\Gamma}_2$  plays the role of removing the impact of  $\mathbf{\Phi}_{ij}$ , both for rows corresponding to good clients.

We now bound  $\lambda_{m-K}(\mathbf{S})$  using the spectrum of  $\tilde{\mathbf{S}}$ .

Define

$$\mathbf{Q} = \begin{pmatrix} \frac{1}{\sqrt{l_1}} \mathbf{1}_{m_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{m_2}} \mathbf{1}_{m_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{m_K}} \mathbf{1}_{m_K} \\ 0 & 0 & \cdots & 0. \end{pmatrix} \in \mathbb{R}^{m \times K},$$

and  $\mathbf{Q}_\perp \in \mathbb{R}^{m \times (m-K)}$  such that  $\mathbf{U} = [\mathbf{Q}_\perp, \mathbf{Q}] \in \mathbb{R}^{m \times m}$  is an orthogonal matrix.

Since each column of  $\mathbf{\Gamma}_1$  is a linear combination of  $\mathbf{Q}$ , and each column of  $\mathbf{Q}_\perp$  is per-

pendicular to all columns of  $\mathbf{Q}$ , we have

$$\mathbf{Q}_\perp^\top \Gamma_1 = \mathbf{0}. \quad (5.28)$$

By (5.24),  $\Phi_{ij} = \mathbf{1}a_{ij}^\top + b\mathbf{1}_{ij}^\top$  for some vector  $a_{ij}$  and  $b_{ij}$ . Thus, we have

$$\Gamma_2 = \underbrace{\begin{pmatrix} \mathbf{0} & \cdots & \mathbf{1}_{m_1}a_{1K}^\top & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_{m_K}b_{1K}^\top & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix}}_{\Gamma_2^{(1)}} + \underbrace{\begin{pmatrix} \mathbf{0} & \cdots & b_{1K}\mathbf{1}_{m_K}^\top & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots \\ a_{1K}\mathbf{1}_{m_1}^\top & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix}}_{\Gamma_2^{(2)}}. \quad (5.29)$$

Since each column of  $\Gamma_2^{(1)}$  is a linear combination of columns of  $\mathbf{Q}$ , by the orthogonality between  $\mathbf{Q}_\perp$  and  $\mathbf{Q}$ , we have  $\mathbf{Q}_\perp^\top \Gamma_2^{(1)} = \mathbf{0}$ .

Similarly, since each row of  $\Gamma_2^{(2)}$  is a linear combination of columns of  $\mathbf{Q}$ , we have  $\Gamma_2^{(2)} \mathbf{Q}_\perp = \mathbf{0}$ .

Hence, we have

$$\mathbf{Q}_\perp^\top \Gamma_2 \mathbf{Q}_\perp = \mathbf{0}$$

Combining the above equality with (5.28), we have

$$\mathbf{Q}_\perp^\top (\Gamma_1 + \Gamma_2) \mathbf{Q}_\perp = \mathbf{0}.$$

Therefore,

$$\lambda_{m-K}(\mathbf{S}) = \lambda_{n-K}(\mathbf{U}^\top \mathbf{S} \mathbf{U}) \geq \lambda_{m-K}(\mathbf{Q}_\perp^\top \mathbf{S} \mathbf{Q}_\perp) = \lambda_{m-K}(\mathbf{Q}_\perp^\top \tilde{\mathbf{S}} \mathbf{Q}_\perp).$$

where the inequality holds by the fact that  $\mathbf{U}^\top \mathbf{S} \mathbf{U} = \begin{pmatrix} \mathbf{Q}_\perp^\top \mathbf{S} \mathbf{Q}_\perp & \mathbf{Q}_\perp^\top \mathbf{S} \mathbf{Q} \\ \mathbf{Q}^\top \mathbf{S} \mathbf{Q}_\perp & \mathbf{Q}^\top \mathbf{S} \mathbf{Q} \end{pmatrix}$  and Lemma A.4.2 (Cauchy Interlacing Theorem).

Similarly, note that  $\mathbf{U}^\top \tilde{\mathbf{S}} \mathbf{U} = \begin{pmatrix} \mathbf{Q}_\perp^\top \tilde{\mathbf{S}} \mathbf{Q}_\perp & \mathbf{Q}_\perp^\top \tilde{\mathbf{S}} \mathbf{Q} \\ \mathbf{Q}^\top \tilde{\mathbf{S}} \mathbf{Q}_\perp & \mathbf{Q}^\top \tilde{\mathbf{S}} \mathbf{Q} \end{pmatrix}$ . Again by Lemma A.4.2 (Cauchy Interlacing Theorem), we have

$$\lambda_{m-K} \left( \mathbf{Q}_\perp^\top \tilde{\mathbf{S}} \mathbf{Q}_\perp \right) \geq \lambda_m \left( \mathbf{U}^\top \tilde{\mathbf{S}} \mathbf{U} \right) = \lambda_m \left( \tilde{\mathbf{S}} \right).$$

In short, to show  $\lambda_{m-K}(\mathbf{S}) > 0$ , it is sufficient to show  $\tilde{\mathbf{S}}$  is PD.

Through  $\tilde{\mathbf{S}}$ , we have better control of the rows corresponding to good clients. To control the rows corresponding to Byzantine clients, we utilize the observation that for any diagonal matrix  $\mathbf{D}$  with positive diagonal entries,  $\tilde{\mathbf{S}} \succ \mathbf{0}$  is equivalent to  $\mathbf{D} \tilde{\mathbf{S}} \mathbf{D} \succ \mathbf{0}$ .

With  $\mathbf{D} = \begin{pmatrix} w \mathbf{I}_{m-q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}$  for some  $0 < w < 1$ , we introduce

$$\hat{\mathbf{S}} = \begin{pmatrix} w^2 \tilde{\mathbf{S}}_{I,I} & w \tilde{\mathbf{S}}_{I,I^c} \\ w \tilde{\mathbf{S}}_{I^c,I} & \tilde{\mathbf{S}}_{I^c,I^c} \end{pmatrix}.$$

In short, for rows in  $I^c$ , the off-diagonal entries corresponding to columns in  $I$  shrink by a factor of  $w$ . Even though the off-diagonal entries in  $(I^c, I^c)$ -block is still the same, the verification of diagonal dominance for rows in  $I^c$  is easier as the size of  $I$  is much larger than  $I^c$ .

As a cost, for rows in  $I$ , the off diagonal entries corresponding to columns in  $I^c$  gets inflated by a factor of  $1/w$  compared to the diagonal entries. However, since  $|I^c| = \epsilon m$ , the impact of the inflation from  $I^c$  columns is limited.

The following Proposition provides a sufficient condition to ensure  $\hat{\mathbf{S}}$  is diagonally dominant, i.e.  $\hat{\mathbf{S}}(i, i) > \sum_{j \neq i} |\hat{\mathbf{S}}(i, j)|, \forall i$ . Hence, by Lemma A.4.3 (Gershgorin Theorem),  $\hat{\mathbf{S}}$  is PD. This further implies  $\tilde{\mathbf{S}}$  is PD.

**Proposition 5.1.6.** *Let  $0 < w < 1$  and*

$$\hat{\mathbf{S}} = \begin{pmatrix} w \mathbf{I}_{m-q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \tilde{\mathbf{S}} \begin{pmatrix} w \mathbf{I}_{m-q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}.$$

We have  $\widehat{\mathbf{S}}$  is diagonally dominant if the following two conditions hold,

$$2a_{\text{in}} - 1 > \frac{\lambda\epsilon}{\rho} + \lambda\rho_k + a_{\text{out}} \frac{1 - \epsilon - \rho}{\rho} + \frac{2\epsilon}{\rho w}, \quad (5.30)$$

$$\gamma > \frac{q-1}{1 - \sqrt{K}w} + \frac{w}{1 - \sqrt{K}w} \left( 2(m-q) + \sqrt{K}(q-1) \right) \lambda. \quad (5.31)$$

Equation (5.30) is to ensure the diagonal dominance of the first  $m-q$  rows while (5.31) is to ensure the diagonal dominance of the last  $q$  rows. The term  $\frac{2\epsilon}{\rho w}$  in (5.30) is the cost to pay for the inflation from  $I^c$  columns. To ensure the diagonal dominance of clients in  $I$  is maintained, particularly for the clients in the smallest cluster,  $\frac{1}{w}$  need to be smaller compared to  $\frac{\rho}{\epsilon}$ . Meanwhile,  $w$  cannot be too large in view of (5.31). This is because  $\gamma$  which appears in the diagonal of  $I^c$  rows, needs to be smaller than  $O(\rho m)$  due to the condition  $\Psi_{ii} > 0$ .

It turns out that  $w = \frac{C\sqrt{K}\epsilon}{5\rho}$  for some large constant  $C$  is enough to satisfy both restrictions on  $w$ . With such  $w$ , (5.30) and (5.31) are equivalent as

$$2a_{\text{in}} - 1 > \frac{\lambda\epsilon}{\rho} + \lambda + a_{\text{out}} \left( \frac{1}{\rho} - \frac{\epsilon}{\rho} - 1 \right) + \frac{10}{C\sqrt{K}}, \quad (5.32)$$

$$\gamma > \frac{q-1}{1 - \frac{CK\epsilon}{5\rho}} + \frac{C\sqrt{K}\epsilon/5}{\rho - CK\epsilon/5} \left( 2(m-q) + \sqrt{K}(q-1) \right) \lambda. \quad (5.33)$$

Now we show (5.32) and (5.33) hold for any  $\gamma$  and  $\lambda$  satisfying (5.2) and (5.3). A rearrangement of (5.32) yields

$$\left( 1 + \frac{\epsilon}{\rho} \right) \lambda < 2a_{\text{in}} - 1 - a_{\text{out}} \left( \frac{1}{\rho} - \frac{\epsilon}{\rho} - 1 \right) + \frac{10}{C\sqrt{K}}.$$

Since the right hand side above is bigger than  $\kappa$ , any  $\lambda$  satisfying (5.2) ensures that (5.32) holds.

Plugging  $q = em$  into the right hand side of (5.33), we have

$$\frac{q-1}{1 - \frac{CK\epsilon}{5\rho}} + \frac{C\sqrt{K}\epsilon/5}{\rho - CK\epsilon/5} \left( 2(m-q) + \sqrt{K}(q-1) \right) \lambda \leq \frac{\rho + 2C\sqrt{K}/5 + CK\epsilon/5}{\rho - CK\epsilon/5} em.$$

Since  $\gamma$  is bigger than  $\frac{\rho+2C\sqrt{K}/5+CK\epsilon/5}{\rho-CK\epsilon/5}\epsilon m$  under (5.3), we ensure (5.33) is satisfied.

## 5.2 Proof of Theorem 5

Fix any  $k \in [K]$ . WLOG, assume client  $1, 2, \dots, M_k$  belong to cluster  $k$ .

For the ease of presentation, denote the loss of client  $i$  evaluated at  $\widehat{\theta}_a^{(0)}$  as

$$l_i(a) = \frac{1}{n_i} \sum_{r=1}^{n_i} l(\widehat{\theta}_a^{(t)}; x_{ir}), \forall a \in [K].$$

Denote  $Q_k = \sum_{i=1}^{M_k} \mathbf{1}_{\{l_i(k) > \min_{j \neq k} l_i(j)\}}$  as the number of clients in cluster  $k$  clustered incorrectly.

Since client data are mutually independent, by Bernstein inequality, with probability at least  $1 - \exp(-\Omega(\log M_k))$ ,

$$\begin{aligned} \frac{Q_k}{M_k} &\leq \frac{\mathbb{E}[Q_k]}{M_k} + \frac{C\sqrt{\text{Var}(Q_k) \log M_k}}{M_k} + \frac{C \log M_k}{M_k} \\ &\leq \frac{\mathbb{E}[Q_k]}{M_k} + \frac{C\sqrt{\text{Var}(Q_k) \log M_k}}{M_k} + \frac{\epsilon}{3\rho_k}, \end{aligned} \quad (5.34)$$

where the last inequality holds under Assumption 3.

We now bound the first two terms on the right-hand side of (5.34).

Denote

$$p_k \triangleq \mathbb{E}[Q_k] = \sum_{i=1}^{M_k} \mathbb{P} \left[ l_i(k) > \min_{j \neq k} l_i(j) \right].$$

To bound  $p_k$ , we need to bound  $\mathbb{P}[l_i(k) > \min_{j \neq k} l_i(j)]$  for all  $i$ .

Fix arbitrary  $i$ . Since  $L_k$  is  $\beta$ -Lipchitz, we have

$$\left| L_k(\widehat{\theta}_j^{(0)}) - L_k(\bar{\theta}_j) \right| \leq \beta \left\| \theta_j^{(0)} - \bar{\theta}_j \right\|_2 = \beta D_j, \forall j \neq k, \quad (5.35)$$

where  $D_k = \left\| \widehat{\theta}_k^{(0)} - \bar{\theta}_k \right\|_2$ .

Thus,

$$\begin{aligned}
L_k(\widehat{\theta}_k^{(0)}) + \frac{\Gamma_k - \beta(D_k + D_j)}{2} &\leq L_k(\bar{\theta}_k) + \frac{\Gamma_k + \beta D_k - \beta D_j}{2} \\
&\leq L_k(\bar{\theta}_j) - \frac{\Gamma_k - \beta D_k + \beta D_j}{2} \\
&\leq L_k(\widehat{\theta}_j^{(0)}) - \frac{\Gamma_k - \beta(D_k + D_j)}{2}, \tag{5.36}
\end{aligned}$$

where  $\Gamma_k = \min_{j \neq k} L_k(\bar{\theta}_j) - L_k(\bar{\theta}_k)$ ; the first and the last inequalities hold by (5.35) and the second inequality holds by the definition of  $\Gamma_k$ .

Therefore, when

$$l_i(k) \leq L_k(\widehat{\theta}_k^{(0)}) + \frac{\Gamma_k - \beta(D_j + D_k)}{2}, \quad \text{and} \quad l_i(j) \geq L_j(\widehat{\theta}_j^{(0)}) - \frac{\Gamma_k - \beta(D_j + D_k)}{2},$$

we must have

$$l_i(k) < l_i(j). \tag{5.37}$$

As a result, we have

$$\begin{aligned}
&\mathbb{P} \left[ l_i(k) \geq \min_{j \neq k} l_i(j) \right] \\
&\leq \mathbb{P} \left[ l_i(k) \geq L_k(\widehat{\theta}_k^{(0)}) + \frac{\Gamma_k - \beta(D_j + D_k)}{2} \right] + \sum_{j \neq k} \mathbb{P} \left[ l_i(j) \leq L_j(\widehat{\theta}_j^{(0)}) - \frac{\Gamma_k - \beta(D_j + D_k)}{2} \right].
\end{aligned}$$

Note that

$$\mathbb{E} [l_i(k)] = L_k(\widehat{\theta}_k^{(0)}), \quad \text{and} \quad \text{Var} (l_i(k)) \leq \frac{\tau_k^2}{n_i}.$$

By Chebyshev's inequality, we have

$$\mathbb{P} \left[ |l_i(a) - L_k(\widehat{\theta}_a^{(0)})| \leq \frac{\Gamma_k - \beta(D_a + D_k)}{2} \right] \leq \frac{4\tau_k^2}{(\Gamma_k - \beta(D_j + D_k))^2 n_i}, \forall a \in [K]. \tag{5.38}$$

Hence,

$$\begin{aligned} \mathbb{P} \left[ l_i(k) > \min_{j \neq k} l_i(j) \right] &\leq \sum_{a=1}^K \mathbb{P} \left[ |l_i(a) - L_k(\widehat{\theta}_a^{(0)})| < \frac{\Gamma_k - \beta(D_a + D_k)}{2} \right] \\ &\leq \frac{8K\tau_k^2}{(\Gamma_k - \beta(D_k + \min_j D_j))^2 n_i}. \end{aligned} \quad (5.39)$$

From Theorem 7, we have

$$D_k \leq \sqrt{\frac{\epsilon}{16C\rho_k}} \nu_k^2 \leq \frac{\min_j \Gamma_j}{4\beta}, \forall k \in [K]$$

where the last inequality holds under Assumption 2.

Plugging the above bounds of  $D_k$  back into (5.39), we have

$$\mathbb{P} \left[ l_i(k) > \min_{j \neq k} l_i(j) \right] \leq \frac{32K\tau_k^2}{\Gamma_k^2 n_i}. \quad (5.40)$$

Summing up the above inequality over  $i$ , we have

$$\frac{p_k}{M_k} = \frac{1}{M_k} \sum_{i=1}^{M_k} \mathbb{P} \left[ v_i(k) > \min_{j \neq k} v_i(j) \right] \leq \frac{1}{M_k} \sum_{i=1}^{M_k} \frac{32K\tau_k^2}{\Gamma_k^2 n_i} = \frac{32KF_k}{\Gamma_k^2} \leq \frac{\epsilon}{3\rho_k}, \quad (5.41)$$

where the last inequality holds under Assumption 3.

Now we bound the second term on the right-hand side of (5.34). Since  $Q_k$  is a sum of independent Bernoulli, we have  $\text{Var}(Q_k) \leq p_k$ .

Therefore,

$$\frac{\sqrt{\log M_k \text{Var}(Q_k)}}{M_k} \leq \frac{\sqrt{\log M_k p_k / M_k}}{\sqrt{M_k}} \leq \frac{\sqrt{32KF_k / \Gamma_k^2} \cdot \sqrt{\log M_k}}{\sqrt{M_k}} \leq \frac{\epsilon}{3\rho_k}, \quad (5.42)$$

where the last inequality holds under Assumption 3.

Plugging (5.41) and (5.42) into (5.34), we have  $Q_k/M_k \leq \epsilon/\rho_k$ .

Taking union bounds over  $k$ , we have with probability at least  $1 - K \exp(-\Omega(\log(\rho M)))$ ,

$$Q_k/M_k \leq \epsilon/\rho_k, \forall k. \quad (5.43)$$

As a result,

$$\begin{aligned} \min_{\pi} \frac{|\widehat{\mathcal{C}}_{\pi(k)} \cap \mathcal{C}_k^c|}{|\widehat{\mathcal{C}}_{\pi(k)}|} &= \frac{|\widehat{\mathcal{C}}_{\pi(k)} \cap \mathcal{C}_k^c|}{|\widehat{\mathcal{C}}_{\pi(k)} \cap \mathcal{C}_k| + |\widehat{\mathcal{C}}_{\pi(k)} \cap \mathcal{C}_k^c|} \\ &\leq \frac{\sum_{j \neq k} \epsilon/\rho_j \cdot M_j + \epsilon M}{(1 - \epsilon/\rho_k) M_k + \sum_{j \neq k} \epsilon/\rho_j \cdot M_j + \epsilon M} \\ &= \frac{K\epsilon}{\rho_k + K\epsilon} \leq \frac{K\epsilon}{\rho_k}, \end{aligned}$$

where the second to last inequality holds by dividing both numerator and denominator by  $M$ .

## 5.3 Proof of Theorem 6

The proof of Theorem 6 is completed by plugging the bound of  $W_k$  from Lemma 4.4.2 into Theorem 8 and 9.

### 5.3.1 Proof of Theorem 8 and 9

To prove both theorems, a crucial step is to control the deviation of the geometric median of means  $\mathcal{G}_k^{(t)}$  from the true population gradient  $\nabla L_k(\widehat{\theta}_k^{(t)})$ . With small  $\left\| \mathcal{G}_k^{(t)} - \nabla L_k(\widehat{\theta}_k^{(t)}) \right\|_2$ , we can approximate the refinement process by gradient descent and hence prove the guarantee.

The following lemma bounds  $\left\| \mathcal{G}_k^{(t)} - \nabla L_k(\widehat{\theta}_k^{(t)}) \right\|_2$  under  $\Omega_t$  where  $\Omega_t(\Delta, \xi; b)$  is defined in (4.14).

**Lemma 5.3.1.** *[CSX17, Lemma 2] For any  $\Delta > 0$ , under  $\Omega_t(\Delta, \xi; b)$ ,*

$$\left\| \mathcal{G}_k^{(t)} - \nabla L_k(\widehat{\theta}_k^{(t)}) \right\|_2 \leq C_{\xi} \Delta,$$

where  $C_{\xi} = \frac{2-2\xi}{1-2\xi}$ .

We now prove Theorem 8 and 9.

*Proof of Theorem 8.* By Lemma 5.3.1, under  $\cap_{t=0}^{T-1} \Omega_t \left( \frac{C\sqrt{\log M_k} \sigma_k}{\sqrt{W_k}}, \xi; b \right)$ , we have

$$\sup_{t \in [T-1]} \left\| \mathcal{G}_k^{(t)} - \nabla L_k \left( \widehat{\theta}_k^{(t)} \right) \right\|_2 \leq \frac{CC_\xi \sqrt{\log M_k} \sigma_k}{\sqrt{W_k}}. \quad (5.44)$$

Thus, under  $\cap_{t=0}^{T-1} \Omega_t \left( \frac{C\sqrt{\log M_k} \sigma_k}{\sqrt{W_k}}, \xi; b \right)$ , for any  $1 \leq t \leq T$ ,

$$\begin{aligned} \left\| \widehat{\theta}_k^{(t)} - \theta_k \right\|_2 &= \left\| \widehat{\theta}_k^{(t-1)} - \eta \mathcal{G}_k^{(t-1)} - \theta_k \right\|_2 \\ &\leq \left\| \widehat{\theta}_k^{(t-1)} - \eta \nabla L_k \left( \widehat{\theta}_k^{(t-1)} \right) - \theta_k \right\|_2 + \eta \left\| \mathcal{G}_k^{(t-1)} - \nabla L_k \left( \widehat{\theta}_k^{(t-1)} \right) \right\|_2 \\ &\leq \sqrt{1 - \frac{\mu^2}{4\zeta^2}} \left\| \widehat{\theta}_k^{(t-1)} - \theta_k \right\|_2 + \frac{CC_\xi \sqrt{\log M_k} \eta \sigma_k}{\sqrt{W_k}} \\ &\leq \left( 1 - \frac{\mu^2}{8\zeta^2} \right) \left\| \widehat{\theta}_k^{(t-1)} - \theta_k \right\|_2 + \frac{CC_\xi \log M_k \eta \sigma_k}{\sqrt{W_k}}, \end{aligned}$$

where the first inequality holds by the triangle inequality, the second inequality holds by Lemma A.4.4 with  $\eta = \frac{\mu}{2\zeta^2}$  and (5.44), and the last inequality holds by the fact that  $\sqrt{1 - \frac{\mu^2}{4\zeta^2}} \leq \sqrt{\left(1 - \frac{\mu^2}{8\zeta^2}\right)^2}$ .

Recursively applying the above displayed inequality, we get

$$\begin{aligned} &\left\| \widehat{\theta}_k^{(t)} - \theta_k \right\|_2 \\ &\leq \left( 1 - \frac{\mu^2}{8\zeta^2} \right)^t \left\| \widehat{\theta}_k^{(0)} - \theta_k \right\|_2 + \frac{CC_\xi \sqrt{\log M_k} \eta \sigma_k}{\sqrt{W_k}} \left( 1 + \left( 1 - \frac{\mu^2}{8\zeta^2} \right) + \dots + \left( 1 - \frac{\mu^2}{8\zeta^2} \right)^{t-1} \right) \\ &\leq \left( 1 - \frac{\mu^2}{8L^2} \right)^t \left\| \widehat{\theta}_k^{(0)} - \theta_k \right\|_2 + \frac{CC_\xi \sqrt{\log M_k} \eta \sigma_k}{\sqrt{W_k}} \frac{1}{\frac{\mu^2}{8\zeta^2}} \\ &= \left( 1 - \frac{\zeta^2}{8\mu^2} \right)^t \left\| \widehat{\theta}_k^{(0)} - \theta_k \right\|_2 + \frac{CC_\xi \sqrt{\log M_k} \sigma_k}{\mu \sqrt{W_k}}, \end{aligned}$$

where the last inequality holds by plugging in  $\eta = \frac{\mu}{2\zeta^2}$ .

Now we show  $\cap_{t=0}^{T-1} \Omega_t \left( \frac{C\sqrt{\log M_k} \sigma_k}{\sqrt{W_k}}, \xi; b \right)$  occurs with high probability.

Recall that  $\sigma_k^2 = \sup_{\theta} \mathbb{E}_{X \sim \mathcal{D}_k} \left[ \left\| \nabla l(\theta; X) - \nabla L_k(\theta) \right\|_2^2 \right]$  is the variance of the gradient and  $g_i^{(t)} = \sum_{r=1}^{n_i} \nabla l(\widehat{\theta}_k^{(t)}; X_{ir}^{(t)})$  is the sum of the gradient from client  $i$ . Since the data is

freshly drawn from  $\mathcal{D}_k$  and hence independent from  $\widehat{\theta}_k^{(t)}$ , we have

$$\mathbb{E} \left[ \left\| \frac{1}{\sum_{i \in \mathcal{B}_j} n_i} \sum_{i \in \mathcal{B}_j} g_i^{(t)} - \nabla \mathcal{L}_k \left( \widehat{\theta}_k^{(t)} \right) \right\|_2^2 \right] \leq \frac{\sigma_k^2}{W_k}.$$

By Chebyshev's inequality, we have

$$\mathbb{P} \left[ \left\| \frac{1}{\sum_{i \in \mathcal{B}_j} n_i} \sum_{i \in \mathcal{B}_j} g_i^{(t)} - \nabla \mathcal{L}_k \left( \widehat{\theta}_j^{(t)} \right) \right\|_2 > \frac{C \sqrt{\log M_k} \sigma_k}{\sqrt{C_\xi W_k}} \right] \leq \frac{C'}{\log M_k}.$$

Thus, we get

$$\mathbb{P} \left[ \Omega_t \left( \frac{C \sigma_k \sqrt{\log M_k}}{\sqrt{W_k}}, \xi; b \right) \right] \geq \mathbb{P} \left[ F \geq (1 - \xi)b + \alpha_k \widehat{M}_k \right].$$

for some constant  $C$  where  $F \sim \text{Binom}(b, 1 - \frac{C'}{\log M_k})$ .

By Chernoff's bound for the binomial distribution, we have

$$\mathbb{P} \left[ F \geq (1 - \xi)b + \alpha_k \widehat{M}_k \right] \geq 1 - \exp \left( -b \psi \left( \xi - \frac{\alpha_k \widehat{M}_k}{b} \middle\| \frac{C'}{\log M_k} \right) \right)$$

for  $\xi - \alpha_k \widehat{M}_k / b > C' / \log M_k$ , where  $\psi(p' || p) = p' \log \frac{p'}{p} + (1 - p') \log \frac{1 - p'}{1 - p}$ .

Since  $b = \max\{2.5 \alpha_k \widehat{M}_k, 1\}$ , we have  $\alpha_k \widehat{M}_k / b < 2/5$ .

Thus, taking  $\xi - \alpha_k \widehat{M}_k / b = 0.05$  so that  $\xi < 1/2$ , we have

$$\begin{aligned} \psi \left( \xi - \frac{\alpha_k \widehat{M}_k}{b} \middle\| \frac{C'}{\log M_k} \right) &= \psi(0.05 || C' / \log M_k) \\ &= 0.05 \log 0.05 - 0.05 \log \frac{C'}{\log M_k} + 0.95 \log \frac{0.95}{1 - C' / \log M_k} \\ &= C \log \log M_k, \end{aligned}$$

for some constant  $C > 0$ . This completes the proof. □

*Proof of Theorem 9.* For the ease of presentation, denote  $\Delta_k^{(t)} = \mathcal{G}_k^{(t)} - \nabla L_k \left( \widehat{\theta}_k^{(t)} \right)$ .

Similar to the proof of Theorem 8, we condition on  $\cap_{t=0}^{T-1} \Omega_t \left( \frac{C\sqrt{\log M_k \sigma_k}}{\sqrt{W_k}}, \xi; b \right)$  which ensures

$$\sup_{t \in [T-1]} \left\| \Delta_k^{(t)} \right\|_2 \leq \frac{CC_\xi \sqrt{\log M_k \sigma_k}}{\sqrt{W_k}}.$$

For any  $1 \leq t \leq T$ , by the smoothness of  $L_k$ , we have

$$\begin{aligned} & L_k \left( \widehat{\theta}_k^{(t+1)} \right) \\ & \leq L_k \left( \widehat{\theta}_k^{(t)} \right) + \left\langle \nabla L_k \left( \widehat{\theta}_k^{(t)} \right), \widehat{\theta}_k^{(t+1)} - \widehat{\theta}_k^{(t)} \right\rangle + \frac{\zeta}{2} \left\| \widehat{\theta}_k^{(t+1)} - \widehat{\theta}_k^{(t)} \right\|_2^2 \\ & = L_k \left( \widehat{\theta}_k^{(t)} \right) + \left\langle \nabla L_k \left( \widehat{\theta}_k^{(t)} \right), -\eta \mathcal{G}_k^{(t)} \right\rangle + \frac{\zeta \eta^2}{2} \left\| \mathcal{G}_k^{(t)} \right\|_2^2 \\ & \stackrel{(a)}{=} L_k \left( \widehat{\theta}_k^{(t)} \right) - \eta \left\| \nabla L_k \left( \widehat{\theta}_k^{(t)} \right) \right\|_2^2 - \eta \left\langle \nabla L_k \left( \widehat{\theta}_k^{(t)} \right), \Delta_k^{(t)} \right\rangle \\ & \quad + \frac{\zeta \eta^2}{2} \left( \left\| \nabla L_k \left( \widehat{\theta}_k^{(t)} \right) \right\|_2^2 + \left\| \Delta_k^{(t)} \right\|_2^2 + 2 \left\langle \nabla L_k \left( \widehat{\theta}_k^{(t)} \right), \Delta_k^{(t)} \right\rangle \right) \\ & = L_k \left( \widehat{\theta}_k^{(t)} \right) - \left( \eta - \frac{\zeta \eta^2}{2} \right) \left\| \nabla L_k \left( \widehat{\theta}_k^{(t)} \right) \right\|_2^2 + \frac{\zeta \eta^2}{2} \left\| \Delta_k^{(t)} \right\|_2^2 - (\eta - \zeta \eta^2) \left\langle \nabla L_k \left( \widehat{\theta}_k^{(t)} \right), \Delta_k^{(t)} \right\rangle \\ & \stackrel{(b)}{\leq} L_k \left( \widehat{\theta}_k^{(t)} \right) - \frac{1}{2\zeta} \left\| \nabla L_k \left( \widehat{\theta}_k^{(t)} \right) \right\|_2^2 + \frac{CC_\xi^2 \log M_k \sigma_k^2}{2\zeta W_k} \end{aligned}$$

where (a) holds by  $\mathcal{A}_k^{(t)} = \Delta_k^{(t)} + \nabla \mathcal{L}_k \left( \widehat{\theta}_k^{(t)} \right)$  and (b) holds by plugging in  $\eta = \frac{1}{\zeta}$ . Recursively applying the above inequality for  $t = 0, 1, \dots, T-1$ , we have

$$L_k \left( \widehat{\theta}_k^{(T)} \right) \leq L_k \left( \widehat{\theta}_k^{(0)} \right) - \frac{1}{2\zeta} \sum_{t=0}^{T-1} \left\| \nabla L_k \left( \widehat{\theta}_k^{(t)} \right) \right\|_2^2 + \frac{TCC_\xi^2 \log M_k \sigma_k^2}{2\zeta W_k}.$$

Since  $L_k \left( \widehat{\theta}_k^{(T)} \right) \geq L_k \left( \theta_k \right)$ , we have

$$\frac{1}{2\zeta} \sum_{t=0}^{T-1} \left\| \nabla L_k \left( \widehat{\theta}_k^{(t)} \right) \right\|_2^2 \leq L_k \left( \widehat{\theta}_k^{(0)} \right) - L_k \left( \theta_k \right) + \frac{TCC_\xi^2 \log M_k \sigma_k^2}{2\zeta W_k}.$$

Multiplying both hand sides above by  $2\zeta/T$ , we complete the proof of Theorem 9.  $\square$

### 5.3.2 Proof of Lemma 4.4.2

The proof follows the idea sketched in Section 4.4.2. For the ease of presentation, denote  $R_k = |\mathcal{C}_k \cap \widehat{\mathcal{C}}_{\pi(k)}|$  as the number of correctly classified clients in cluster  $k$  and  $H_k = \sum_{i \in \mathcal{C}_k \cap \widehat{\mathcal{C}}_{\pi(k)}} n_i$ . Note that both  $R_k$  and  $H_k$  are random where the randomness comes from  $\widehat{\mathcal{C}}_{\pi(k)}$ .

Recall the definition

$$W_k = \min_{j \in [b]} \sum_{i \in \mathcal{C}_k \cap \mathcal{B}_j} n_i,$$

where  $\{\mathcal{B}_j\}_{j \in [b]}$  denotes the random partition of  $\widehat{\mathcal{C}}_{\pi(k)}$  at the first step of Algorithm 5.

To bound  $W_k$ , we control  $\sum_{i \in \mathcal{C}_k \cap \mathcal{B}_j} n_i$  for all  $j$ . Fix any  $j \in [b]$  and condition on  $\widehat{\mathcal{C}}_{\pi(k)}$ . Note that

$$\begin{aligned} \mathbb{E} \left[ \sum_{i \in \mathcal{C}_k \cap \mathcal{B}_j} n_i \mid \widehat{\mathcal{C}}_{\pi(k)} \right] &= \mathbb{E} \left[ \sum_{i \in \mathcal{C}_k \cap \widehat{\mathcal{C}}_{\pi(k)}} n_i \mathbf{1}_{\{i \in \mathcal{B}_j\}} \mid \widehat{\mathcal{C}}_{\pi(k)} \right] \\ &= \frac{1}{b} \sum_{i \in \mathcal{C}_k \cap \widehat{\mathcal{C}}_{\pi(k)}} n_i = \frac{H_k}{b}, \quad \forall j \in [b], \end{aligned}$$

where the second equality holds as the probability of client  $i$  in batch  $\mathcal{B}_j$  is  $1/b$ .

Apply Lemma A.1.5 with  $t = \frac{H_k}{2R_k}$ , we get

$$\sum_{i \in \mathcal{C}_k \cap \mathcal{B}_j} n_i \geq \frac{H_k}{2b} \tag{5.45}$$

with a probability at least

$$1 - \exp \left( -\frac{CH_k}{bR_k(n_{\max} - n_{\min})^2} \right) \tag{5.46}$$

for some universal constant  $C$ .

Next, we bound  $R_k$ ,  $b$  and  $H_k$ , which depend on  $\widehat{\mathcal{C}}_{\pi(k)}$ . By the definition of  $R_k$ , we have

$$R_k = |\mathcal{C}_k \cap \widehat{\mathcal{C}}_{\pi(k)}| \leq |\mathcal{C}_k| = M_k. \tag{5.47}$$

Recall from Theorem 5 that  $\alpha_k$  is the upper bound of the misclassification rate of  $\widehat{\mathcal{C}}_{\pi(k)}$ . Thus, we have  $|\widehat{\mathcal{C}}_{\pi(k)}| \leq \frac{M_k}{1-\alpha_k}$ . As a result,

$$b = \max \left\{ 2.5\alpha_k |\widehat{\mathcal{C}}_{\pi(k)}|, 1 \right\} \leq \max \left\{ \frac{2.5\alpha_k M_k}{1-\alpha_k}, 1 \right\}. \quad (5.48)$$

Note that

$$H_k = \sum_{i \in \mathcal{C}_k \cap \widehat{\mathcal{C}}_{\pi(k)}} n_i = \sum_{i \in \mathcal{C}_k} n_i \mathbf{1}_{\{\widehat{z}_i = z_i\}}.$$

From (5.40), we have

$$\mathbb{E} [\mathbf{1}_{\{\widehat{z}_i = z_i\}}] = \mathbb{P} [\widehat{z}_i = z_i] \geq 1 - \frac{32K\tau_k^2}{\Gamma_k^2 n_i}.$$

Thus,

$$\mathbb{E} [H_k] \geq \sum_{i \in \mathcal{C}_k} n_i \left( 1 - \frac{32K\tau_k^2}{\Gamma_k^2 n_i} \right) = N_k - \frac{32KM_k\tau_k^2}{\Gamma_k^2} \geq CN_k, \quad (5.49)$$

for some universal constant  $1 > C > 0$ , where the last inequality holds under Assumption 4.

To show  $H_k$  cannot be much smaller than  $\mathbb{E} [H_k]$ , we utilize the independence of  $\{\mathbf{1}_{\{\widehat{z}_i = z_i\}}\}$  and apply Lemma A.1.4 to get

$$\mathbb{P} [H_k < c\mathbb{E} [H_k]] = \mathbb{P} \left[ \frac{H_k}{n_{\max}} < c\mathbb{E} \left[ \frac{H_k}{n_{\max}} \right] \right] = \exp(-\Omega(\mathbb{E} [H_k] / n_{\max})), \quad (5.50)$$

where  $c > 0$  is some universal constant.

Plugging (5.49) into (5.50), we have with probability at least  $1 - \exp(-\Omega(N_k/n_{\max}))$ ,

$$H_k \geq CN_k, \quad (5.51)$$

for some universal constant  $C > 0$ .

Plugging (5.47), (5.48) and (5.51) back into (5.45) and (5.46), we have

$$\sum_{i \in \mathcal{C}_k \cap \mathcal{B}_j} n_i \geq \frac{H_k}{2b} \geq C \frac{H_k}{\max\{\alpha_k M_k, 1\}},$$

with probability at least

$$\begin{aligned}
& 1 - \exp(-\Omega(N_k/n_{\max})) - \exp\left(-\Omega\left(\frac{H_k^2}{R_k b(n_{\max} - n_{\min})^2}\right)\right) \\
&= 1 - \exp(-\Omega(N_k/n_{\max})) - \exp\left(-\Omega\left(\frac{N_k^2}{M_k \max\{2.5\alpha_k M_k/(1 - \alpha_k), 1\}(n_{\max} - n_{\min})^2}\right)\right) \\
&\stackrel{(a)}{=} 1 - \exp(-\Omega(\log M_k))
\end{aligned}$$

where (a) holds under Assumption 5.

The proof is completed by taking the union bounds over  $j$  and using the fact that  $b \leq M_k$ .

## 5.4 Proof of Lemma 4.4.1

Recall that  $\{\tilde{\mathcal{C}}_j\}_{j \in [K]}$  are the output clusters.

Here, we claim that for any cluster  $\tilde{\mathcal{C}}_j$ , if  $|\tilde{\mathcal{C}}_j| > \epsilon m$ , then we must have  $\mathcal{C}_k \subset \tilde{\mathcal{C}}_i$  for some  $k \in [K]$ .

Since each node is only assigned once, meaning that  $\tilde{\mathcal{C}}_j$  are disjoint, we complete the proof.

Now we prove the claim. For the ease of presentation, we use  $\widehat{\mathbf{X}}(v)$  to denote the rows of  $\widehat{\mathbf{X}}$  corresponding to node  $v$ .

Recall that  $\tilde{\mathcal{C}}_j$  is initiated from some node  $v$  picked at **Node assignment** step.

**Case 1:**  $v \in \mathcal{C}_k \cap \mathcal{A}_g$  For any nodes  $u$  and  $u'$  in  $\mathcal{C}_k \cap \mathcal{A}_g$ ,  $\widehat{\mathbf{X}}(u)$  and  $\widehat{\mathbf{X}}(u')$  differ by no more than  $\epsilon m$  coordinates, and hence  $u$  and  $u'$  must be neighbors.

Since  $v \in \mathcal{C}_k \cap \mathcal{A}_g$ , all nodes in  $\mathcal{C}_k \cap \mathcal{A}_g$  must share at least  $\rho m - 2$  neighbors with  $v$ . As a result, all nodes of cluster  $k$  must be included at **Node assignment** step.

Moreover, good nodes from different clusters cannot be neighbors since their rows differ by at least  $\rho m$  coordinates. Thus, nodes from  $\mathcal{C}_j$  for  $j \neq k$  will not be included.

**Case 2:  $v$  is Byzantine** Since  $|\tilde{\mathcal{C}}_j| > \epsilon m$ , at least one good node is included in  $\tilde{\mathcal{C}}_j$ . Denote such good node as  $u$  and assume  $u \in \mathcal{C}_k$ .

Since  $u$  and  $v$  share more than  $\epsilon m + 1$  neighbors, at least one of their common neighbor is in  $\mathcal{A}_g$ . Denote the collection of good clients that are common neighbors of  $u$  and  $v$  as  $\mathcal{V}$ . Since all nodes in  $\mathcal{V}$  are neighbors of  $u$ , we know  $\mathcal{V} \subset \mathcal{C}_k \cap \mathcal{A}_g$ .

Now we show all the neighbors of  $v$  in  $\mathcal{A}_g$  must belong to  $\mathcal{V}$ .

Recall that one of the neighbors of  $v$  is in  $\mathcal{V} \subset \mathcal{C}_k$ . Thus,  $\hat{\mathbf{X}}(v)$  differs by no more than  $\epsilon m$  coordinates with some nodes in cluster  $\mathcal{C}_k \cap \mathcal{A}_g$ . As a result,  $\hat{\mathbf{X}}(v)$  must differ by at least  $\rho m - \epsilon m$  coordinates with  $\hat{\mathbf{X}}(w)$  for any  $w \in \mathcal{C}_r \cap \mathcal{A}_g$  for  $r \neq k$ . Thus, any  $w \in \mathcal{C}_r \cap \mathcal{A}_g$ ,  $r \neq k$  cannot be neighbor of  $v$ . In other word, all neighbors of  $v$  in  $\mathcal{A}_g$  must be in  $\mathcal{V}$ .

Since  $v$  has degree more than  $2\epsilon m + 1$ , at least  $\epsilon m + 2$  neighbors of  $v$  are in  $\mathcal{A}_g$ . Hence,  $|\mathcal{V}| \geq \epsilon m + 2$ .

Therefore, for any  $w' \in \mathcal{C}_k$ , at least  $\epsilon m + 1$  nodes in  $\mathcal{V}$  are common neighbors of  $w'$  and  $v$ . As a result,  $w'$  must be included in  $\tilde{\mathcal{C}}_i$  at **Node assignment** step.

Since all neighbors of  $v$  in  $\mathcal{A}_g$  must belong to  $\mathcal{V} \subset \mathcal{C}_k$ , any node  $u' \in \mathcal{C}_r$ ,  $r \neq k$  cannot share more than  $\epsilon m - 1$  neighbors with  $v$ . Therefore,  $u'$  cannot be included in  $\tilde{\mathcal{C}}_i$ .

This completes the proof.

## 5.5 Proof of Intermediate Results in Section 5.1

### 5.5.1 Proof of Proposition 5.1.5

Throughout this section, for any matrix  $\mathbf{M}$ , we use

$$\|\mathbf{M}\|_\infty = \sup_{i,j} |\mathbf{M}(i,j)|.$$

We first present one corollary of Lemma 5.1.3 which will be used in the proof of Proposition 5.1.5.

**Corollary 3.** [CL15, Lemma 6.9] Under the same condition in Lemma 5.1.3, for any  $i, j \in [K]$ , we have

$$x_j^\top \left( \widetilde{\mathbf{W}} + \text{diag} \{ \xi \} \right) x_i \leq q \sqrt{m_j m_i}, \quad (5.52)$$

$$\alpha_i(s) + e_s^\top \mathbf{Z}_i^\top \mathbf{1}_{m_i} \leq (\gamma + \xi_s) x_i(s) + \lambda m_i + \lambda \|x_i\|_1, \forall s \in [q] \quad (5.53)$$

$$0 \leq \alpha_j \leq (q + m_j - 1) \lambda \mathbf{1}_q, \quad (5.54)$$

where  $q = \epsilon m$ .

*Proof of Corollary 3.* We first prove (5.52). Multiplying both hand sides of (5.11) by  $x_i^\top$  and noticing (5.13), we have

$$\begin{aligned} x_i^\top \left( \widetilde{\mathbf{W}} + \text{diag} \{ \xi \} \right) x_i &= - \left( \widetilde{\mathbf{Z}}_i x_i \right)^\top \mathbf{1}_{m_i} \\ &\stackrel{(a)}{\leq} (1 - \lambda) q \|x_i\|_\infty \mathbf{1}_{m_i}^\top \mathbf{1}_{m_i} \leq q m_i \end{aligned}$$

where (a) holds by the fact that each entry of  $\widetilde{\mathbf{Z}}_i$  must be larger than  $\lambda - 1$  and the last inequality holds by  $\|x_i\|_\infty \leq 1$ .

By Cauchy-Schwartz inequality, we have

$$x_j^\top \left( \widetilde{\mathbf{W}} + \text{diag} \{ \xi \} \right) x_i \leq \sqrt{x_i^\top \left( \widetilde{\mathbf{W}} + \text{diag} \{ \xi \} \right) x_i} \sqrt{x_j^\top \left( \widetilde{\mathbf{W}} + \text{diag} \{ \xi \} \right) x_j} = q \sqrt{m_i m_j}.$$

Now we prove (5.53). By plugging in the definition of  $\widetilde{\mathbf{Z}}_i$  and  $\widetilde{\mathbf{W}}$  from (5.8) into (5.11), we have

$$(\gamma \mathbf{I}_q + \lambda \mathbf{J}_q - \mathbf{W} + \text{diag} \{ \xi \}) x_i + \lambda m_i \mathbf{1}_q = \alpha_i + \mathbf{Z}_i^\top \mathbf{1}_q.$$

Taking the  $s$ -th coordinate on both hands side, we have

$$(\gamma + \xi_s) x_i(s) + \lambda \|x_i\|_1 - [\mathbf{W} x_i](s) + \lambda m_i = \alpha_i(s) + e_s^\top \mathbf{Z}_i^\top \mathbf{1}_{m_i}.$$

Since each entry of  $\mathbf{W}$  and each coordinate of  $x_i$  are non-negative,  $[\mathbf{W}x_i](s) \geq 0$ . As a result, we obtain

$$\alpha_i(s) + e_s^\top \mathbf{Z}_i^\top \mathbf{1}_{m_i} \leq (\gamma + \xi_s) x_i(s) + \lambda \|x_i\|_1 + \lambda m_i.$$

Lastly, we prove (5.54). When  $\alpha_i(s) = 0$ , the inequality clearly holds.

When  $\alpha_i(s) > 0$ , by (5.13), we have  $x_j(s) = 0$ . Since  $\mathbf{Z}_i$  is non-negative, following (5.53), we have

$$\begin{aligned} \alpha_i(s) &\leq \lambda m_i + \lambda \sum_{j=1}^q x_i(j) \\ &\leq (q + m_i - 1)\lambda. \end{aligned}$$

□

Now we prove the proposition. Fix any  $i < j$ .

Note that from (5.22) and (5.23), we have  $\mathbf{1}_{m_i}^\top g_{ij} = \mathbf{1}_{m_i}^\top \Phi_{ij} \mathbf{1}_{m_j} = g_{ji}^\top \mathbf{1}_{m_j}$ . Thus, by (5.11), we get

$$v_{ij} = \mathbf{1}_{m_i}^\top g_{ij} = g_{ji}^\top \mathbf{1}_{m_j} = \lambda m_i m_j - \mathbf{1}_{m_i}^\top \mathbf{K}_{ij} \mathbf{1}_{m_j} - x_i^\top \left( \widetilde{\mathbf{W}} + \text{diag} \{ \xi \} \right) x_j. \quad (5.55)$$

Plugging (5.22), (5.23) and (5.55) into (5.24), we get

$$\begin{aligned}
\Phi_{ij} &= \frac{1}{m_j} \left( \tilde{\mathbf{Z}}_i x_j - \frac{1}{m_i} \mathbf{1}_{m_i} \alpha_i^\top x_j + \lambda m_j \mathbf{1}_{m_i} - \mathbf{K}_{ij} \mathbf{1}_{m_j} \right) \mathbf{1}_{m_j}^\top \\
&\quad + \frac{1}{m_i} \mathbf{1}_{m_i} \left( \tilde{\mathbf{Z}}_j x_i - \frac{1}{m_j} \mathbf{1}_{m_j} \alpha_j^\top x_i + \lambda m_i \mathbf{1}_{m_j} - \mathbf{K}_{ij}^\top \mathbf{1}_{m_i} \right)^\top \\
&\quad - \frac{1}{m_i m_j} \left( \lambda m_i m_j - \mathbf{1}_{m_i}^\top \mathbf{K}_{ij} \mathbf{1}_{m_j} - x_i^\top \left( \tilde{\mathbf{W}} + \text{diag} \{ \xi \} \right) x_j \right) \mathbf{J}_{m_i, m_j} \\
&= \underbrace{\left( \lambda + \frac{1}{m_i m_j} \mathbf{1}_{m_i}^\top \mathbf{K}_{ij} \mathbf{1}_{m_j} \right) \mathbf{J}_{m_i, m_j}}_{\text{(I)}} + \underbrace{\frac{1}{m_i} \mathbf{1}_{m_i} x_i^\top \tilde{\mathbf{Z}}_j^\top + \frac{1}{m_j} \tilde{\mathbf{Z}}_i x_j \mathbf{1}_{m_j}^\top}_{\text{(II)}} \\
&\quad - \underbrace{\left( \frac{1}{m_i} \mathbf{J}_{m_i} \mathbf{K}_{ij} + \frac{1}{m_j} \mathbf{K}_{ij} \mathbf{J}_{m_j} \right)}_{\text{(III)}} \\
&\quad - \underbrace{\frac{1}{m_i m_j} \left( x_i^\top \left( \tilde{\mathbf{W}} + \text{diag} \{ \xi \} \right) x_j + \mathbf{1}_{m_i}^\top \tilde{\mathbf{Z}}_i x_j + x_i^\top \tilde{\mathbf{Z}}_j^\top \mathbf{1}_{m_j} \right) \mathbf{J}_{m_i, m_j}}_{\text{(IV)}}. \tag{5.56}
\end{aligned}$$

By (5.8), we have

$$\lambda - 1 \leq \tilde{\mathbf{Z}} = \lambda \mathbf{J} - \mathbf{Z} \leq 1. \tag{5.57}$$

Hence,  $\tilde{\mathbf{Z}}_i$  and  $\tilde{\mathbf{Z}}_j$  can have negative entries, meaning that each entry of (II) and (IV) can be negative.

Therefore, to show  $\Phi_{ij} \geq 0$ , we show for any  $s, t$ , the  $(s, t)$ -th entry of (I) is larger than the sum of the absolute values of the  $(s, t)$ -th entry of (II), (III) and (IV).

For (I), since  $\mathbf{K}_{ij} \geq 0$ , we get

$$\lambda + \frac{1}{m_i m_j} \mathbf{1}_{m_i}^\top \mathbf{K}_{ij} \mathbf{1}_{m_j} \geq \lambda. \tag{5.58}$$

Now we analyze (II). By (5.57), we have

$$\left\| \tilde{\mathbf{Z}}_i \right\|_\infty \leq 1. \tag{5.59}$$

Hence, we have

$$\left\| \tilde{\mathbf{Z}}_p x_{p'} \right\|_{\infty} \leq \left\| \tilde{\mathbf{Z}}_p \right\|_{\infty} \|x_{p'}\|_1 = \|x_{p'}\|_1, \quad (p, p') \in \{(j, i), (i, j)\}. \quad (5.60)$$

By the second constraint of (5.10), we have  $\sup_s \{x_i^2(s) + x_j^2(s)\} \leq 1$ . This implies  $\|x_i + x_j\|_{\infty} \leq \sqrt{2}$ .

As a result,

$$\begin{aligned} \left\| \frac{1}{m_i} \mathbf{1}_{m_i} x_i^{\top} \tilde{\mathbf{Z}}_j^{\top} + \frac{1}{m_j} \tilde{\mathbf{Z}}_i x_j \mathbf{1}_{m_j}^{\top} \right\|_{\infty} &\leq \left\| \frac{1}{m_i} \tilde{\mathbf{Z}}_j x_i \right\|_{\infty} + \left\| \frac{1}{m_j} \tilde{\mathbf{Z}}_i x_j \right\|_{\infty} \\ &\leq \frac{1}{m_i} \|x_i\|_1 + \frac{1}{m_j} \|x_j\|_1 \\ &\stackrel{(a)}{\leq} \frac{\sum_{r=1}^q (x_i(r) + x_j(r))}{\min\{m_i, m_j\}} \\ &\leq \frac{q \|x_i + x_j\|_{\infty}}{\min\{m_i, m_j\}} \leq \frac{\sqrt{2}q}{\min\{m_i, m_j\}}. \end{aligned} \quad (5.61)$$

where (a) holds by the fact that  $x_i \geq 0, \forall i$ .

Next, we bound (III). By (5.26) which gives  $\|\mathbf{K}_{ij} \mathbf{1}_{m_j}\|_{\infty} \leq a_{\text{out}} m_j$ , we have

$$\left\| \frac{1}{m_i} \mathbf{J}_{m_i} \mathbf{K}_{ij} + \frac{1}{m_j} \mathbf{K}_{ij} \mathbf{J}_{m_j} \right\|_{\infty} \leq 2a_{\text{out}}. \quad (5.62)$$

Lastly, we bound (IV). Following (5.60) and (5.52), we have

$$\begin{aligned} &\left| \frac{1}{m_i m_j} \left( x_i^{\top} \left( \tilde{\mathbf{W}} + \text{diag}\{\xi\} \right) x_j + \mathbf{1}_{m_i}^{\top} \tilde{\mathbf{Z}}_i x_j + x_i^{\top} \tilde{\mathbf{Z}}_j^{\top} \mathbf{1}_{m_j} \right) \right| \\ &\leq \frac{q\sqrt{m_i m_j} + \sqrt{2}(m_i + m_j)q}{m_i m_j} \leq \frac{(1 + \sqrt{2})q}{\min\{m_i, m_j\}}. \end{aligned} \quad (5.63)$$

Combining (5.58),(5.61),(5.62) and (5.63), we get  $\Phi_{ij} > 0$  for all  $i \neq j$  as long as

$$\lambda > \frac{(1 + 2\sqrt{2})q}{\min_{k \in [K]} m_k} + 2a_{\text{out}}. \quad (5.64)$$

**Remark 5.5.1.** Under SBM, in order for  $\Phi \geq 0$ , [CL15] requires

$$\lambda > \frac{C_2 q}{\min_k m_k} + a_{\text{out}}$$

for some constant  $C_2$  where  $c_1$  there represents the probability of an edge formed between one node in cluster  $i$  and one node in cluster  $j$ .

In comparison to (5.64), we see an extra  $c_1$  in our study. We now explain why this is inevitable. Under SBM, each entry of  $\mathbf{K}_{ij}$  is formed with i.i.d. Bernoulli distribution. Therefore, we expect the connectivity of individual client to be close to the average connectivity. This implies that  $\frac{1}{m_i m_j} \mathbf{1}^\top \mathbf{K}_{ij} \mathbf{1}$  is close to  $a_{\text{out}}$ . As a result, one can show with high probability, all entries of  $(I)$  are lower bounded by  $\lambda + a_{\text{out}} - \delta'$  for some small  $\delta'$ . However, in our study without any distribution assumption, we do not have this guarantee as there can be a large discrepancy of connectivity between clients. For example, we can have  $\mathbf{K}_{ij} = 1$  when  $i = 1, j \leq a_{\text{out}}$  or  $j = 1, i \leq a_{\text{out}}$  and 0 otherwise. As a result,

$$\frac{1}{m_i m_j} \mathbf{1}^\top \mathbf{K}_{ij} \mathbf{1} < \frac{(m_i + m_j) a_{\text{out}}}{m_i m_j} < \frac{2a_{\text{out}}}{\min_i m_i},$$

which shows that  $\frac{1}{m_i m_j} \mathbf{1}^\top \mathbf{K}_{ij} \mathbf{1}$  can be much smaller than  $a_{\text{out}}$ .

## 5.5.2 Proof of Proposition 5.1.6

For the ease of presentation, we use  $I$  denote the collection of good clients.

**Diagonal dominance for  $i \in I$ :** Fix arbitrary  $k \in [K]$  and assume client  $i$  belongs to cluster  $k$ , i.e.,  $m_{k-1} < i \leq m_k$ . We now show  $\widehat{\mathbf{S}}(i, i) > \sum_{j \neq i} |\widehat{\mathbf{S}}(i, j)|$ .

Firstly, we derive an lower bound of  $\widehat{\mathbf{S}}(i, i) = w^2 \widetilde{\mathbf{S}}(i, i)$ .

By (5.27) and (5.14), we have

$$\widetilde{\mathbf{S}}_{kk} = \mathbf{J}_{m_k} + \text{diag} \left\{ \mathbf{K}_{kk} \mathbf{1}_{m_k} - \widetilde{\mathbf{Z}}_k x_k \right\} - \mathbf{K}_{kk} - \lambda m_k \mathbf{I}_k - \delta \mathbf{I}_k \quad (5.65)$$

For any  $0 < s \leq m_k$ , by (5.59),

$$\left[ \mathbf{K}_{kk} \mathbf{1} - \tilde{\mathbf{Z}}_k x_k \right] (s) \geq [\mathbf{K}_{kk} \mathbf{1}] (s) - \lambda q. \quad (5.66)$$

Plugging (5.66) into (5.65), we have

$$\widehat{\mathbf{S}}(i, i) = w^2 \tilde{\mathbf{S}}(i, i) \geq w^2 (1 + d(i) - \lambda q - \lambda m_k - \delta), \quad (5.67)$$

where  $d_k \triangleq \mathbf{K}_{kk} \mathbf{1} \in \mathbb{R}^{m_k}$  and  $d \triangleq (d_1^\top, \dots, d_k^\top)^\top \in \mathbb{R}^{m-q}$  is the within-cluster degree for all good clients.

Next we derive an upper bound on the sum of the absolute values of the off-diagonal entries  $\sum_{j \neq i} |\widehat{\mathbf{S}}(i, j)|$ . Note that

$$\sum_{j \neq i} |\widehat{\mathbf{S}}(i, j)| = \sum_{m_{k-1} < j \leq m_k, j \neq i} |\widehat{\mathbf{S}}(i, j)| + \sum_{j \leq m_{k-1} \text{ or } m_k < j \leq m-q} |\widehat{\mathbf{S}}(i, j)| + \sum_{j > m-q} |\widehat{\mathbf{S}}(i, j)|. \quad (5.68)$$

Intuitively, the first term on the right hand side corresponds to the off-diagonal elements within  $\widehat{\mathbf{S}}_{kk}$ . The second term corresponds to the entries in  $\widehat{\mathbf{S}}_{kt}$  for  $t \neq k$ . The last term corresponds to the entries associated with Byzantine clients. We now bound each term of (5.68) in order.

For the first term, note that by (5.65),

$$|\tilde{\mathbf{S}}(i, j)| = \begin{cases} 0 & \text{if } \mathbf{K}(i, j) = 1 \\ 1 & \text{if } \mathbf{K}(i, j) = 0 \end{cases}$$

for any  $m_{k-1} < j \leq m_k, i \neq j$  where  $\mathbf{K}$  is defined in (5.7).

Since there are  $d(i)$  number of 1's in  $\{\mathbf{K}(i, j)\}_{j=m_{k-1}+1}^{m_k}$ , we get

$$\sum_{m_{k-1} < j \leq m_k, j \neq i} |\widehat{\mathbf{S}}(i, j)| = w^2 \sum_{m_{k-1} < j \leq m_k, j \neq i} |\tilde{\mathbf{S}}(i, j)| = w^2 (m_k - d(i) - 1). \quad (5.69)$$

For the second term, by (5.27) and (5.14), we have

$$\tilde{\mathbf{S}}_{kt} = -\mathbf{K}_{kt}, \forall t \neq k.$$

By the definition of  $a_{\text{out}}$  in (4.1), we have

$$\sum_{j \leq m_{k-1}, m_k < j \leq m-q} \left| \widehat{\mathbf{S}}(i, j) \right| = w^2 \sum_{j \leq m_{k-1}, m_k < j \leq m-q} \left| \widehat{\mathbf{S}}(i, j) \right| \leq w^2 a_{\text{out}}(m - q - m_k). \quad (5.70)$$

Now we bound the last term of (5.68). Note that  $\left\{ \widehat{\mathbf{S}}(i, j) \right\}_{j > m-q}$  corresponds to some row of  $\mathbf{T}_k = \tilde{\mathbf{Z}}_k - \frac{1}{m_k} \mathbf{1}_{m_k} \alpha_k^\top \in \mathbb{R}^{m_k \times q}$ . Since each entry of  $\tilde{\mathbf{Z}}_k$  is either  $\lambda$  or  $\lambda - 1$ , following (5.54), we know

$$\|\mathbf{T}_k\|_\infty \leq \max \left\{ \lambda, 1 - \lambda + \frac{(q + m_k - 1)\lambda}{m_k} \right\} \leq 1 + \frac{\epsilon}{\rho} \lambda,$$

where the last inequality holds by  $q/m_k < \epsilon/\rho$ .

As a result,

$$\sum_{j > m-q} \left| \widehat{\mathbf{S}}(i, j) \right| \leq w \sum_{j > m-q} \left| \widehat{\mathbf{S}}(i, j) \right| \leq wq \|\mathbf{T}_k\|_\infty \leq w(1 + \frac{\epsilon}{\rho} \lambda)q.$$

Combining the above inequality with (5.67), (5.69), (5.70) and (5.68), we obtain the following sufficient condition for the  $i$ -th row of  $\widehat{\mathbf{S}}$  to be diagonally dominant where  $m_{k-1} < i \leq m_k$ :

$$w(d(i) + 1 - \lambda q - \lambda m_k - \delta) > w(m_k - d(i) - 1) + w a_{\text{out}}(m - q - m_k) + \left(1 + \frac{\epsilon}{\rho} \lambda\right) q.$$

Since  $d(i) > a_{\text{in}} m_k$  by (4.1), we get

$$(2a_{\text{in}} - 1) m_k > \lambda q + \lambda m_k + a_{\text{out}}(m - q - m_k) + \frac{(1 + \epsilon \lambda / \rho) q}{w},$$

where the inequality holds by  $\delta < 2/m < 2$ .

Dividing both hand sides by  $m$ , we obtain

$$(2a_{\text{in}} - 1) \rho_k > \lambda \epsilon + \lambda \rho_k + a_{\text{out}} (1 - \epsilon - \rho_k) + \frac{(1 + \epsilon \lambda / \rho) \epsilon}{w}, \quad (5.71)$$

where  $\rho_k = m_k/m$  and  $\epsilon = q/m$ .

**Remark 5.5.2.** In [CL15], the authors utilize the distribution assumption and obtain a more relaxed condition than (5.71). In particular, in view of (5.65), they show

$$\tilde{\mathbf{S}}_{kk} \succeq \text{diag} \left\{ \mathbf{K}_{kk} \mathbf{1}_{m_k} - \tilde{\mathbf{Z}}_k x_k \right\} - (\lambda m_k - \Omega + C_3 \sqrt{m_k a_{\text{in}}}) \mathbf{I}.$$

Similar to our  $\widehat{\mathbf{S}}$ , they need to show

$$\begin{pmatrix} w \mathbf{I}_{m-q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix} \mathbf{S}' \begin{pmatrix} w \mathbf{I}_{m-q} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m \end{pmatrix} \succ \mathbf{0},$$

where

$$\begin{aligned} \mathbf{S}'_{tt} &= \text{diag} \left\{ \mathbf{K}_{tt} \mathbf{1}_{m_t} - \tilde{\mathbf{Z}}_t x_t \right\} - (\lambda t - \Omega + C_3 \sqrt{m_t a_{\text{in}}}) \mathbf{I}, t \in [K] \\ \mathbf{S}'_{st} &= \tilde{\mathbf{S}}_{st}, \forall s, t, s \neq t \end{aligned}$$

Since  $\mathbf{S}'_{kk}$  is a diagonal matrix, one have  $\sum_{m_{k-1} < j \leq m_k, j \neq i} |\mathbf{S}'(i, j)| = 0$ . This is significantly smaller than the corresponding counterpart in (5.69) of our study.

**Diagonal dominance when  $i \in I^c$ :** Assume  $i = n - q + p$  for some  $1 \leq p \leq q$ .

From (5.14), we get

$$\widehat{\mathbf{S}}(i, i) = \widetilde{\mathbf{W}}(p, p) + \xi_p \geq \gamma + \xi_p,$$

where the last inequality holds by  $\widetilde{\mathbf{W}} = \gamma \mathbf{I} + \lambda \mathbf{J}_q - \mathbf{W}$  and  $\mathbf{W} \in \{0, 1\}^{q \times q}$ .

For the off-diagonal terms in row  $i$ , we use a similar decomposition as (5.68):

$$\sum_{j \neq i} |\widehat{\mathbf{S}}(i, j)| = \sum_{j > m-q, j \neq i} |\widehat{\mathbf{S}}(i, j)| + \sum_{j \leq m-q} |\widehat{\mathbf{S}}(i, j)|. \quad (5.72)$$

To bound  $\sum_{j>m-q, j\neq i} \left| \widehat{\mathbf{S}}(i, j) \right|$ , we bound each  $\left| \widehat{\mathbf{S}}(i, j) \right|$  for  $j > m - q, j \neq i$ .

In particular, we have

$$\left| \widehat{\mathbf{S}}(i, j) \right| \leq w \max \{ \lambda - 1, \lambda \} \leq w, \forall j > m - q, j \neq i \quad (5.73)$$

where the first inequality holds since each off-diagonal entry of  $\widetilde{\mathbf{W}}$  is either  $\lambda - 1$  or  $\lambda$  by (5.14) and (5.8). It remains to bound  $\sum_{j\leq m-q} \left| \widehat{\mathbf{S}}(i, j) \right|$  which corresponds to the  $p$ -th row of  $\left\{ \widetilde{\mathbf{Z}}_k^\top - \frac{1}{m_k} \alpha_k \mathbf{1}_{m_k}^\top \right\}_{k\in[K]}$ .

By (5.53), we get

$$\sum_{r=1}^{m_k} \left| \left[ \widetilde{\mathbf{Z}}_k^\top - \frac{1}{m_k} \alpha_k \mathbf{1}_{m_k}^\top \right] (p, r) \right| \leq m_k \lambda + (\gamma - \lambda + \xi_p) x_k(p) + \lambda m_k + \lambda \|x_k\|_1, \forall k \in [K].$$

Therefore, we have

$$\begin{aligned} \sum_{j\leq m-q} \left| \widehat{\mathbf{S}}(i, j) \right| &\leq w \sum_{k=1}^K [m_k \lambda + (\gamma - \lambda + \xi_p) x_k(p) + \lambda m_k + \lambda \|x_k\|_1] \\ &\leq w \left[ 2(m - q)\lambda + \sqrt{K}\gamma + \sqrt{K}\xi_p + \sqrt{K}(q - 1)\lambda \right] \end{aligned} \quad (5.74)$$

where the last inequality holds by Cauchy-Schwartz inequality and the fact that  $\sum_{k=1}^K [x_k(p)]^2 \leq 1$  from Lemma 5.1.3.

Plugging (5.73) and (5.74) into (5.72), we have

$$\sum_{j\neq i} \left| \widehat{\mathbf{S}}(i, j) \right| \leq q - 1 + w \left[ 2(m - q)\lambda + \sqrt{K}\gamma + \sqrt{K}\xi_p + \sqrt{K}(q - 1)\lambda \right].$$

Therefore, to ensure the diagonally dominance of the last  $q$  rows of  $\widehat{\mathbf{S}}$ , we need

$$\gamma + \xi_p > q - 1 + w \left( 2(m - q)\lambda + \sqrt{K}\gamma + \sqrt{K}\xi_p + \sqrt{K}(q - 1)\lambda \right)$$

which is satisfied if

$$\gamma > \frac{q-1}{1-\sqrt{K}w} + \frac{w}{1-\sqrt{K}w} \left(2(m-q) + \sqrt{K}(q-1)\right) \lambda.$$

# Chapter 6

## Conclusion and Future Directions

In this thesis, we delve into the theoretical perspective of employing GD methods in modern machine learning problems.

In the first half, we focus on the SGD dynamic in training of multi-layer neural networks, where the associated loss function is non-convex and non-smooth. By showing the uniform concentration of the Neural Tangent Kernel (NTK) from all hidden layers of neural networks, we can now capture the contribution of NTK from intermediate layers in characterizing the GD/SGD dynamics. Additionally, in the streaming setting, we show the average prediction error under SGD converges in expectation. Our analysis opens up exciting possibilities for future research directions. For example, extending our study to Markovian data arising in the reinforcement learning is of great interest. We also aim to apply our uniform concentration results to other neural network architectures, such as convolutional neural network (CNN).

In the second half, we explore the use of GD method in the new training paradigm called Federated Learning. To address the challenges posed by data heterogeneity and adversarial attack, we propose a Byzantine-resilient algorithm capable of recovering the cluster structure and estimating model parameters for each cluster. The algorithm first leverages the existence of a few anchor clients to obtain coarse estimate of model parameters for each cluster. Utilize the idea of batching and the geometric median of means, the algorithm then aggregates gradients from local clients to refine the estimates. A notable advantage of our algorithm is its ability to function without any distributional assumptions, a crucial aspect lacking in previous federated learning algorithms. While we currently assume the clients have access to fresh sample at each communication round, and hence the local data on clients are i.i.d. across iterations, one interesting future direction is to extend the this study to the case where local data on each client are repeatedly used. Additionally,

we plan to apply our algorithm to train specific modern machine learning architectures such as deep neural networks. The theoretical understanding of training neural networks in Federated Learning will significantly contribute to facilitating the development of modern AIs in the big-data era.

# Appendix A

## Auxiliary Result

### A.1 Concentration Inequalities

In this section, we provide the concentration inequalities used in this thesis. First of all, we present McDiarmid's inequality.

**Lemma A.1.1.** *[HR19, Theorem 2.3] Let  $X = (X_1, \dots, X_m) \in \mathcal{X}^m$  be an  $n$ -tuple of  $\mathcal{X}$ -valued independent random variables and  $f : \mathcal{X}^m \rightarrow \mathbb{R}$  be a measurable function. Assume the value of  $f(x)$  can change by at most  $c_i > 0$  under an arbitrary change of the  $i$ -th coordinate. Then for any  $t > 0$ ,*

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^m c_i^2}\right).$$

The following lemma is Bernstein inequality which shows the concentration of the sum of i.i.d. sub-exponential random variables.

**Lemma A.1.2.** *[Ver19, Theorem 2.8.1] Let  $X_1, \dots, X_n$  be i.i.d., sub-exponential random variables with sub-exponential norm  $\|X_i\|_{\psi_1} \leq K$ . Then for any  $t > 0$ , we have*

$$\mathbb{P}\left[\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X_1]\right| > t\right] \leq 2 \exp\left(-C \min\left(\frac{nt^2}{K^2}, \frac{nt}{K}\right)\right).$$

Finally, we present Hanson-Wright inequality. For any random variable  $X$ , define  $\|X\|_{\psi_2} \triangleq \inf\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$ .

**Lemma A.1.3.** *(Hanson-Wright inequality) [Ver18, Theorem 6.2.1] Let  $X_1, \dots, X_n$  be a random vector with independent component  $X_i$  which has mean 0, sub-Gaussian parameter*

$\sigma^2$ . Let matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$ . Then for any  $t$ , we have

$$\mathbb{P} \left[ \left| X^\top \mathbf{A} X - \mathbb{E} \left[ X^\top \mathbf{A} X \right] \right| > t \right] \leq 2 \exp \left( -c \min \left( \frac{t^2}{\sigma^4 \|\mathbf{A}\|_F^2}, \frac{t}{\sigma^2 \|\mathbf{A}\|_2} \right) \right). \quad (\text{A.1})$$

**Lemma A.1.4.** (Bound on the tail of a weighted sum of independent Bernoulli random variables) [Rag88, Theorem 2] Let  $a_1, \dots, a_n$  be reals in  $(0, 1]$ . Let  $X_1, \dots, X_n$  be independent Bernoulli random variables with  $\mathbb{E}[X_i] = p_i$ . Denote  $Y = \sum_{i=1}^n a_i X_i$ . Then for any  $\gamma \in (0, 1)$ , we have

$$\mathbb{P}[Y < (1 - \gamma)\mathbb{E}[Y]] < \left[ \frac{e^\gamma}{(1 + \gamma)^{1+\gamma}} \right]^{\mathbb{E}[Y]}.$$

**Lemma A.1.5.** [Ser74, Corollary 1] Consider a finite list of values  $(a_1, a_2, \dots, a_N)$ . Let  $X_1, \dots, X_n$  be a sample of size  $n$  drawn without replacement from  $\{a_i\}_{i=1}^N$ . Then

$$\mathbb{P} \left[ \sum_{i=1}^n X_i - n\mu \geq nt \right] \leq \exp \left( -\frac{2nt^2}{[1 - (n-1)/(N-1)](a_{\max} - a_{\min})^2} \right),$$

where  $\mu = \frac{1}{N} \sum_{i=1}^N a_i$ ,  $a_{\max} = \max_i a_i$  and  $a_{\min} = \min_i a_i$ .

## A.2 VC Dimension

Let  $\mathcal{C}$  be a collection of subsets in  $\mathbb{R}^p$ . For any set  $A$  consisting of finite points in  $\mathbb{R}^p$ , we denote  $\mathcal{C}_A = \{C \cap A : C \in \mathcal{C}\}$ . We say  $\mathcal{C}_A$  shatters set  $A$  if  $|\mathcal{C}_A| = 2^{|A|}$ . Let  $\mathcal{M}_{\mathcal{C}}(n) = \max \{|\mathcal{C}_F| : F \subset \mathbb{R}^p, |F| = n\}$  and  $\mathcal{P}(\mathcal{C}) = \sup \{n : \mathcal{M}_{\mathcal{C}}(n) = 2^n\}$  which is the largest cardinality of a set that can be shattered by  $\mathcal{C}$ .

Consider a Boolean function class  $\mathcal{F}$  on  $\mathbb{R}^p$ . For each  $f \in \mathcal{F}$ , we denote  $D_f = \{x \in \mathbb{R}^p : f(x) = 1\}$ . As a result, the collection  $\mathcal{C}_{\mathcal{F}} \triangleq \{D_f : f \in \mathcal{F}\}$  forms a collection of subsets of  $\mathbb{R}^p$ . Define  $\mathcal{C}_{\mathcal{F}}(A) = \{D_f \cap A : f \in \mathcal{F}\}$ . The VC dimension of  $\mathcal{F}$  is then defined as

$$\text{VC}(\mathcal{F}) \triangleq \mathcal{P}(\mathcal{C}_{\mathcal{F}}) = \sup \left\{ n : \max_A |\mathcal{C}_{\mathcal{F}}(A)| = 2^n : |A| = n, A \subset \mathbb{R}^p \right\}.$$

Now we provide the auxiliary results regarding VC dimension.

The following lemma can be used to obtain the bound of VC dimension of the function class consisting of functions with the form of a product of Boolean functions.

**Lemma A.2.1.** [VDVW09, Theorem 1.1] For Boolean function classes  $\mathcal{H}$  and  $\{\mathcal{F}_i\}_{i=1}^N$ , if for any  $h \in \mathcal{H}$ , there exists functions  $f_1 \in \mathcal{F}_1, \dots, f_N \in \mathcal{F}_N$  such that  $h = \prod_{i=1}^N f_i$ , then we have

$$\text{VC}(\mathcal{H}) \leq \frac{5}{2} \log(4N) \sum_{i=1}^N \text{VC}(\mathcal{F}_i).$$

The next lemma bounds the expectation of the largest deviation of an average of some Boolean function through VC dimension.

**Lemma A.2.2.** [Ver19, Theorem 8.3.23] Let  $\mathcal{F}$  be a class of Boolean functions on a probability space  $(\Omega, \Sigma, \mu)$  with finite VC dimension. Let  $X_1, X_2, \dots, X_n$  be independent random points in  $\Omega$ . Then

$$\mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_X [f(X)] \right| \right] \leq C \sqrt{\frac{\text{VC}(\mathcal{F})}{n}}$$

for some constant  $C$ .

Next, we present Sauer-Shelah Lemma which bounds the cardinality of  $\mathcal{C}_{\mathcal{F}}(A)$  with the VC dimension of  $\mathcal{F}$ .

**Lemma A.2.3.** [Ver19, Theorem 8.3.16] Let  $\mathcal{F}$  be a Boolean function class and  $A = \{a_1, \dots, a_n\}$  be a set of  $n$  points in the space. Then for any  $n \geq \text{VC}(\mathcal{F})$ ,

$$|\{(f(a_1), \dots, f(a_n)) : f \in \mathcal{F}\}| = |\mathcal{C}_{\mathcal{F}}(A)| \leq \left( \frac{en}{\text{VC}(\mathcal{F})} \right)^{\text{VC}(\mathcal{F})}.$$

**Lemma A.2.4.** [HR19, Proposition 7.1] Let  $\mathcal{F} = \{f_{\theta}(y) = \mathbf{1}_{\{(y, \theta) \geq 0\}} : \theta \in \Theta\}$  where  $y \in \mathbb{R}^p$  and  $\Theta$  is some  $q$ -dimensional subspace of  $\mathbb{R}^p$ . Then

$$\text{VC}(\mathcal{F}) = q.$$

Lastly, we provide a lemma that bounds the VC dimension of union of function classes

when the number of function classes is much larger than their VC dimension.

**Lemma A.2.5.** *Suppose  $\mathcal{F} = \cup_{i=1}^N \mathcal{F}_i$  where  $\text{VC}(\mathcal{F}_i) = d$  for all  $i$ , then*

$$\text{VC}(\mathcal{F}) \leq C \max(d \log d, \log N).$$

*Proof of Lemma A.2.5.* Fix arbitrary set  $A = \{y_1, \dots, y_n\}$  of size  $n$ . Since  $\mathcal{F} = \cup_{i=1}^N \mathcal{F}_i$ , we have  $\mathcal{C}_{\mathcal{F}}(A) = \cup_{j=1}^N \mathcal{C}_{\mathcal{F}_j}(A)$ .

Thus, we have

$$|\mathcal{C}_{\mathcal{F}}(A)| \leq \sum_{j=1}^N |\mathcal{C}_{\mathcal{F}_j}(A)| \stackrel{(a)}{\leq} \sum_{j=1}^N n^{\text{VC}(\mathcal{F}_j)} \leq Nn^d.$$

where (a) holds by Lemma A.2.3.

By the definition of VC dimension, if  $n^d N < 2^n$ , then  $\text{VC}(\mathcal{F}) < n$ .

Taking logarithm on both hand sides, we have  $d \log n + \log N < n \log 2$ .

Note that when  $\frac{n}{2} > d \log n$  and  $\frac{n}{2} > \log N$ , i.e.,  $n \geq \max(Cd \log d, 2 \log N)$ , the above inequality clearly holds.

Therefore, we get

$$\text{VC}(\mathcal{F}) \leq C \max(Cd \log d, \log N)$$

for some universal constant  $C$ . □

## A.3 Kernel

Here, we provide some intermediate results regarding kernel operator.

**Lemma A.3.1.** *[STC<sup>+</sup>04, Proposition 3.22] For any positive semi-definite kernel  $\kappa_1$  and  $\kappa_2$ , any function  $\phi$ , we have  $\kappa_3$ ,  $\kappa_4$  and  $\kappa_5$  are positive semi-definite kernels where*

$$\kappa_3(x, y) \triangleq \kappa_1(x, y) + \kappa_2(x, y), \tag{A.2}$$

$$\kappa_4(x, y) \triangleq \kappa_1(x, y)\kappa_2(x, y), \tag{A.3}$$

and

$$\kappa_5(x, y) \triangleq \kappa_1(\phi(x), \phi(y)). \quad (\text{A.4})$$

**Lemma A.3.2.** [STC<sup>+</sup> 04, Theorem 3.13] Suppose  $f(x, y)$  is a kernel function. If for any  $g \in L_2(\mu)$ ,

$$\int \int f(x, y)g(x)g(y)d\mu(x)d\mu(y) \geq 0,$$

then  $f$  is positive semi-definite.

**Lemma A.3.3.** [Mer09] Suppose  $\kappa$  is a positive semi-definite kernel. Then there exists non-negative eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots$  and orthonormal eigenfunctions  $\{\phi_i\}$  such that

$$\kappa(x, y) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(y).$$

**Lemma A.3.4.** For any positive semi-definite kernel operator  $\mathbf{J}$  associated with function  $J$ , we have

$$\|\mathbf{J}\|_2 \leq \|J\|_{\infty}.$$

*Proof of Lemma A.3.4.* By Cauchy-Schwartz inequality, we have

$$\begin{aligned} \|\mathbf{J}\|_2^2 &= \sup_{\|g\|_2=1} \int \left( \int J(x, y)g(y)d\mu(y) \right)^2 d\mu(x) \\ &\leq \sup_{\|g\|_2=1} \int \int J^2(x, y)d\mu(y)d\mu(x) \int g^2(y)d\mu(y) \\ &\leq \|J\|_{\infty}^2 \end{aligned}$$

where  $\|g\|_2 = \sqrt{\int g^2(x)d\mu(x)}$  is the  $L_2$  norm for function  $g$ . □

## A.4 Others

**Lemma A.4.1.** *Let  $\{A_i\}$  and  $\{B_i\}$  be two sequences of events, where  $A_0$  and  $B_0$  are the whole probability spaces. Then we have for any  $n \geq 1$ ,*

$$\mathbb{P}[\cap_{i=1}^n (A_i \cap B_i)] \geq 1 - \sum_{i=1}^n \mathbb{P}[B_i^c | \cap_{k=1}^{i-1} (A_k \cap B_k)] - \sum_{i=1}^n \mathbb{P}[A_i^c]$$

and

$$\mathbb{P}[B_n \cap (\cap_{i=1}^n A_i)] \geq 1 - \sum_{i=1}^n \mathbb{P}[B_i^c | B_{i-1} \cap (\cap_{k=1}^{i-1} A_k)] - \sum_{i=1}^n \mathbb{P}[A_i^c],$$

where  $\cap_{i=1}^0 F_i$  for any event  $F_i$  is understood as the whole probability space.

*Proof of Lemma A.4.1.* Note that for any event  $E$  and  $F$ , we have

$$\mathbb{P}[F] \leq \mathbb{P}[E \cap F] + \mathbb{P}[E^c \cap F] \leq \mathbb{P}[E] + \mathbb{P}[E^c \cap F]. \quad (\text{A.5})$$

Taking  $E = \cap_{i=1}^n (A_i \cap B_i)$  and  $F = \cap_{i=1}^{n-1} (A_i \cap B_i)$ , we have

$$\mathbb{P}[\cap_{i=1}^{n-1} (A_i \cap B_i)] \leq \mathbb{P}[\cap_{i=1}^n (A_i \cap B_i)] + \mathbb{P}[(\cap_{i=1}^n (A_i \cap B_i))^c \cap (\cap_{i=1}^{n-1} (A_i \cap B_i))]. \quad (\text{A.6})$$

Now we bound  $\mathbb{P}[(\cap_{i=1}^n (A_i \cap B_i))^c \cap (\cap_{i=1}^{n-1} (A_i \cap B_i))]$ .

Since  $(\cap_{i=1}^n (A_i \cap B_i))^c = [\cap_{i=1}^{n-1} (A_i \cap B_i)]^c \cup A_n^c \cup B_n^c$ , we have

$$\begin{aligned} \mathbb{P}[(\cap_{i=1}^n (A_i \cap B_i))^c \cap (\cap_{i=1}^{n-1} (A_i \cap B_i))] &\leq \mathbb{P}[(A_n^c \cup B_n^c) \cap (\cap_{i=1}^{n-1} (A_i \cap B_i))] \\ &\leq \mathbb{P}[A_n^c] + \mathbb{P}[B_n^c \cap (\cap_{i=1}^{n-1} (A_i \cap B_i))] \\ &\leq \mathbb{P}[A_n^c] + \mathbb{P}[B_n^c | \cap_{i=1}^{n-1} (A_i \cap B_i)]. \end{aligned}$$

Plugging the aboved displayed equation into (A.6), we have

$$\mathbb{P}[\cap_{i=1}^n (A_i \cap B_i)] \geq \mathbb{P}[\cap_{i=1}^{n-1} (A_i \cap B_i)] - \mathbb{P}[A_n^c] - \mathbb{P}[B_n^c | \cap_{i=1}^{n-1} (A_i \cap B_i)].$$

Recursively replacing  $\mathbb{P} [\cap_{i=1}^{n-1} (A_i \cap B_i)]$  on the right hand side of the above inequality, we obtain the first inequality of Lemma A.4.1.

Similarly, we prove the second inequality of Lemma A.4.1 . From (A.5), taking  $E = B_n \cap (\cap_{i=1}^n A_i)$  and  $F = B_{n-1} \cap (\cap_{i=1}^{n-1} A_i)$ , we have

$$\mathbb{P} [B_{n-1} \cap (\cap_{i=1}^{n-1} A_i)] \leq \mathbb{P} [B_n \cap (\cap_{i=1}^n A_i)] + \mathbb{P} [(B_n \cap (\cap_{i=1}^n A_i))^c \cap (B_{n-1} \cap (\cap_{i=1}^{n-1} A_i))].$$

Since  $(B_n \cap (\cap_{i=1}^n A_i))^c = B_n^c \cup A_n^c \cup (\cap_{i=1}^{n-1} A_i)^c$ , we have

$$\begin{aligned} & \mathbb{P} [(B_n \cap (\cap_{i=1}^n A_i))^c \cap (B_{n-1} \cap (\cap_{i=1}^{n-1} A_i))] \\ & \leq \mathbb{P} [A_n^c] + \mathbb{P} [B_n^c \cap B_{n-1} \cap (\cap_{i=1}^{n-1} A_i)] \\ & \leq \mathbb{P} [A_n^c] + \mathbb{P} [B_n^c | B_{n-1} \cap (\cap_{i=1}^{n-1} A_i)]. \end{aligned}$$

Thus,  $\mathbb{P} [B_n \cap (\cap_{i=1}^n A_i)] \geq \mathbb{P} [B_{n-1} \cap (\cap_{i=1}^{n-1} A_i)] - \mathbb{P} [A_n^c] - \mathbb{P} [B_n^c | B_{n-1} \cap (\cap_{i=1}^{n-1} A_i)]$ .

Recursively applying the above inequality, we obtain the second inequality of Lemma A.4.1. □

**Lemma A.4.2.** *(Cauchy's Interlacing Theorem) [HJ12, Theorem 4.3.28] Let  $\mathbf{A} \in \mathbb{R}^{n \times n}$  be a symmetric matrix and  $\mathbf{B} \in \mathbb{R}^{(n-k) \times (n-k)}$  be  $\mathbf{A}$ 's principal sub-matrix. Then for any  $j = 1, 2, \dots, n - k$ , we have*

$$\lambda_j(\mathbf{A}) \geq \lambda_j(\mathbf{B}) \geq \lambda_{j+k}(\mathbf{A}).$$

**Lemma A.4.3.** *(Gershgorin Theorem) Let  $\mathbf{A} \in \mathbb{R}^{n \times n}$  be a symmetric matrix. Denote  $R_i = \sum_{j \neq i}^n |\mathbf{A}(i, j)|$ . Then the eigenvalues of  $\mathbf{A}$  lies in the unions of the intervals  $[\mathbf{A}(i, i) - R_i, \mathbf{A}(i, i) + R_i], i \in [n]$ .*

The following lemma provides the convergence guarantee of the standard gradient descent algorithm.

**Lemma A.4.4.** *Let  $L$  be any  $\mu$ -strongly convex and  $\zeta$ -smooth function. Denote  $\theta^*$  as its*

unique global minimizer where  $\nabla L(\theta^*) = 0$ . Then we have

$$\|\theta' - \theta^*\|_2 \leq \sqrt{1 - \frac{\mu^2}{4\zeta^2}} \|\theta - \theta^*\|_2,$$

where

$$\theta' = \theta - \frac{\mu}{2\zeta^2} \nabla L(\theta).$$

Proof of the above lemma can be found in [CSX17, Section B].

## Bibliography

- [ADH<sup>+</sup>19] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.
- [AFG<sup>+</sup>23] Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafaël Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. *arXiv preprint arXiv:2302.01772*, 2023.
- [ASR88] Milton Abramowitz, Irene A Stegun, and Robert H Romer. Handbook of mathematical functions with formulas, graphs, and mathematical tables, 1988.
- [AZL19a] Zeyuan Allen-Zhu and Yuanzhi Li. Can sgd learn recurrent neural networks with provable generalization? In *Advances in Neural Information Processing Systems*, pages 10331–10341, 2019.
- [AZL19b] Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? In *Advances in Neural Information Processing Systems*, pages 9017–9028, 2019.
- [AZL20] Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020.
- [AZLS19a] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.
- [AZLS19b] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Advances in neural information processing systems*, pages 6676–6688, 2019.
- [BEMGS17] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.
- [BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- [BK20] Ainesh Bakshi and Pravesh Kothari. Outlier-robust clustering of non-spherical mixtures. *arXiv preprint arXiv:2005.02970*, 2020.
- [CB18] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

- [CBCG04] Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.
- [CC11] Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In *Proceedings of the learning to rank challenge*, pages 1–24. PMLR, 2011.
- [CCGZ20] Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. Mean-field analysis of two-layer neural networks: Non-asymptotic rates and generalization bounds. *arXiv preprint arXiv:2002.04026*, 2020.
- [CG19] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10836–10846, 2019.
- [CI12] María José Cantero and Arieh Iserles. On rapid computation of expansions in ultraspherical polynomials. *SIAM Journal on Numerical Analysis*, 50(1):307–327, 2012.
- [CL15] T Tony Cai and Xiaodong Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. 2015.
- [CS09] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.
- [CSV17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.
- [CSX17] Yudong Chen, Lili Su, and Jiaming Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–25, 2017.
- [DD21] Deepesh Data and Suhas Diggavi. Byzantine-resilient high-dimensional sgd with local iterations on heterogeneous data. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2478–2488. PMLR, 18–24 Jul 2021.
- [DKK<sup>+</sup>22] Ilias Diakonikolas, Daniel M Kane, Sushrut Karmalkar, Ankit Pensia, and Thanasis Pittas. Robust sparse mean estimation via sum of squares. In *Conference on Learning Theory*, pages 4703–4763. PMLR, 2022.
- [DLL<sup>+</sup>19] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019.
- [DWZ<sup>+</sup>18] Simon S Du, Yining Wang, Xiyu Zhai, Sivaraman Balakrishnan, Russ R Salakhutdinov, and Aarti Singh. How many samples are needed to estimate a convolutional neural network? In *Advances in Neural Information Processing Systems*, pages 373–383, 2018.

- [DX13] Feng Dai and Yuan Xu. *Approximation theory and harmonic analysis on spheres and balls*, volume 23. Springer, 2013.
- [DZPS19] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *ICLR 2019*, 2019.
- [FGG<sup>+</sup>22] Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pages 6246–6283. PMLR, 2022.
- [FIM<sup>+</sup>01] Joan Feigenbaum, Yuval Ishai, Tal Malkin, Kobbi Nissim, Martin J Strauss, and Rebecca N Wright. Secure multiparty computation of approximations. In *International Colloquium on Automata, Languages, and Programming*, pages 927–938. Springer, 2001.
- [GCYR20] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.
- [GFJ15] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.
- [GHYR19] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.
- [HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.
- [HLLL19] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, 4(1), 2019.
- [HR19] Bruce Hajek and Maxim Raginsky. Statistical learning theory. *Lecture Notes*, 387, 2019.
- [HTW<sup>+</sup>18] Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, and Qingling Sun. Deep learning for image-based cancer detection and diagnosis- a survey. *Pattern Recognition*, 83:134–149, 2018.
- [ILG07] Elena Ikonomovska, Suzana Loskovska, and Dejan Gjorgjevik. A survey of stream data mining. In *Proceedings of 8th National Conference with International participation, ETAI*, pages 19–21, 2007.
- [Iss18] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

- [JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.
- [KHJ21] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pages 5311–5319. PMLR, 2021.
- [KHJ22] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*. PMLR, 2022.
- [KMA<sup>+</sup>21] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D’Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrede Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends<sup>®</sup> in Machine Learning*, 14(1–2):1–210, 2021.
- [LLL<sup>+</sup>20] Xiaodong Li, Yang Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. When do birds of a feather flock together? k-means, proximity, and conic programming. *Mathematical Programming*, 179:295–341, 2020.
- [LLV21] Chengxi Li, Gang Li, and Pramod K Varshney. Federated learning with soft clustering. *IEEE Internet of Things Journal*, 2021.
- [LWY<sup>+</sup>19] Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.
- [Mer09] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.
- [MLL<sup>+</sup>22] Xiaohang Ma, Lingxia Liao, Zhi Li, Roy Xiaorong Lai, and Miao Zhang. Applying federated learning in software-defined networks: A survey. *Symmetry*, 14(2):195, 2022.

- [MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.
- [MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.
- [MMR<sup>+</sup>17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agueria y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [MMRS20] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.
- [Mut05] Shanmugavelayutham Muthukrishnan. *Data streams: Algorithms and applications*. Now Publishers Inc, 2005.
- [MVW17] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.
- [OMM<sup>+</sup>02] Liadan O’callaghan, Nina Mishra, Adam Meyerson, Sudipto Guha, and Rajeev Motwani. Streaming-data algorithms for high-quality clustering. In *Proceedings 18th International Conference on Data Engineering*, pages 685–694. IEEE, 2002.
- [Pol64] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.
- [Rag88] Prabhakar Raghavan. Probabilistic construction of deterministic algorithms: approximating packing integer programs. *Journal of Computer and System Sciences*, 37(2):130–143, 1988.
- [RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [SCV17] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *arXiv preprint arXiv:1703.04940*, 2017.
- [Ser74] Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48, 1974.
- [SMS20] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

- [SS22] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.
- [STC<sup>+</sup>04] John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.
- [SX19] Lili Su and Jiaming Xu. Securing distributed gradient descent in high dimensional statistical learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(1):1–41, 2019.
- [SXY22] Lili Su, Jiaming Xu, and Pengkun Yang. Global convergence of federated learning for mixed regression. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [SY19] Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective. In *Advances in Neural Information Processing Systems*, pages 2641–2650, 2019.
- [VDVW09] Aad Van Der Vaart and Jon A Wellner. A note on bounds for vc dimensions. *Institute of Mathematical Statistics collections*, 5:103, 2009.
- [Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- [Ver19] Roman Vershynin. *High-dimensional probability*. Cambridge, UK: Cambridge University Press, 2019.
- [Wan10] Yi Wang. Harmonic analysis and isoperimetric inequalities. *Lecture Notes*, 2010.
- [WBZ<sup>+</sup>21] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- [XKG18] Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Generalized byzantine-tolerant sgd. *arXiv preprint arXiv:1802.10116*, 2018.
- [XLS<sup>+</sup>21] Ming Xie, Guodong Long, Tao Shen, Tianyi Zhou, Xianzhi Wang, Jing Jiang, and Chengqi Zhang. Multi-center federated learning. *arXiv preprint arXiv:2108.08647*, 2021.
- [Yah] Yahoo learning to rank challenge(c14). <https://webscope.sandbox.yahoo.com/>.
- [YCKB18] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

- [YLL16] Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.
- [ZCZG20] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

## Biography

Hanjing Zhu is a fifth-year PhD Candidate in Decision Science group at the Fuqua School of Business, Duke University under the supervision of Prof. Jiaming Xu. He receives his M.S. in Statistics at University of Chicago under the supervision of Prof. Weibiao Wu in 2018, and B.S in Economics and Finance at The Hong Kong University of Science and Technology in 2016.

His research interest focuses on machine learning, particularly deep learning and federated learning. He will be joining Amazon SCOT as a research scientist after graduation.