# Testing, Learning, and Optimization in High Dimensions

by

Khashayar Gatmiry

B.S., Sharif University of Technology (2019)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Masters of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Department of Electrical Engineering and Computer Science
May 13, 2022

Certified by. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Stefanie Jegelka
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Jonathan Kelner
Professor of Mathematics
Thesis Supervisor

Accepted by . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee of Graduate Students

# Testing, Learning, and Optimization in High Dimensions

by

Khashayar Gatmiry

Submitted to the Department of Electrical Engineering and Computer Science
on May 13, 2022, in partial fulfillment of the
requirements for the degree of
Masters of Science

## Abstract

In this thesis we study two separate problems: (1) What is the sample complexity of testing the class of Determinantal Point Processes? and (2) Introducing a new analysis for optimization and generalization of deep neural networks beyond their linear approximation. For the first problem, we characterize the optimal sample complexity up to logarithmic factors by proposing almost matching upper and lower bounds. For the second problem, we propose a new regime for the parameters and the algorithm of a three layer network model which goes beyond the Neural tangent kernel (NTK) approximation; as a result, we introduce a new data dependent complexity measure which generalizes the NTK complexity measure introduced by [Arora et al., 2019a]. We show that despite nonconvexity, a variant of Stochastic gradient descent (SGD) converges to a good solution for which we prove a novel generalization bound that is proportional to our complexity measure.

Thesis Supervisor: Stefanie Jegelka
Title: Associate Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Jonathan Kelner
Title: Professor of Mathematics

# Contents

# Chapter 1

# Introduction

In this modern era of data science, we have observed a dramatic increase in the computation power and resources; consequently, more and more sophisticated models are being used among various inference tasks. Moreover, given the inherent complexity of lots of datasets today, one often requires considering a rich enough family of models; Particularly, this family should be able to describe the underlying data distribution to an acceptable level. In fact, the right choice of modeling is critical in the success of the prediction task later on; it is worthy to mention that picking a good model is usually tied up with having a good domain knowledge in the related field. As an example of a modeling choice, the family of Determinantal point processes has been empoyed successfully in machine learning tasks in which one requires to model diversity and repulsion in the sampled subsets of a ground set. We will elaborate more on this parameteric family shortly.

On the other hand, this rise of complexity in statistical models is in spite of the fact that often times we can only have access to a limited number of samples from the underlying generative process; while in many cases, the parameter count of the model, i.e. the number of "degrees of freedom" exceeds the number of samples. This phenomenon is known as overparameterization, and is regarded as one of the major reasons for the success of deep learning. As a result, acquiring a more refined view of the sample complexity in various inference tasks has become vital. In this regard, the computational overhead of the related algorithm and its tradeoff with the statistical aspects is of both theoretical and practical interest, as ultimately we would have to run the algorithm on a computer.

In this thesis, we study the complexity of inference tasks in two important settings: (1) Testing the family of Determinantal Point processes, and (2) Learning a predictor using overparameterized neural networks. In the following, we first focus on the topic of distribution testing, then move on to optimization and generalization for deep learning.

## 1.1 Testing Determinantal point processes

The general framework of distribution testing is, that given a class of distributions $\mathcal{C}$ and observing samples from an underlying distribution $q$, one asks whether $q$ is in $\mathcal{C}$ or it is $\epsilon$-far from it in some distance $d$ between probability measures. The question of interest here is how many samples one needs to observe to be able to solve this decision problem with at least a constant chance of success. Distribution testing has obtained decent amount of attention in the community recently [Paninski, 2008, Batu et al., 2001, Acharya et al., 2015, Diakonikolas et al., 2018, Chan et al., 2014, Diakonikolas and Kane, 2016, Batu et al., 2004, Aliakbarpour et al., 2019]. In this thesis, we ask this question for the class of Determinantal Point Processes which we introduce next.

Determinantal point processes (DPP) are an important parametric family of distributions over subsets of a finite or infinite ground set that have recently been popularized in the machine learning community due to their ability to model diversity and repulsion [Macchi, 1975, Hough et al., 2006, Kulesza and Taskar, 2012, Li et al., 2016b, Kulesza and Taskar, 2012]. A DPP over subsets of a ground set of cardinality $n$ can be characterized by an $n$ by $n$ matrix $K$, where the marginal probabilities can be computed as

$$\mathbb{P}(A \subseteq \mathcal{J}) = \det(K_A).$$

Above, $K_A$ is the principal submatrix of $K$ corresponding to subset $A$. What makes

DPPs particularly attractive for various ML tasks is their ability to model what is called negative dependence between the elements of the ground set, i.e. given that a fixed element is in the sampled subset, the chance of another element being in the subset decreases.

In particular, given the vast popularity of deploying DPPs in different machine learning tasks these days, it is important to study mechanisms by which we can test the hypothesis that a given dataset has been generated by an underlying DPP distribution, or is at least close to one. In this thesis, we settle the question of sample complexity in testing DPPs and prove that it almost scales as $\sqrt{N}/\epsilon^2$, where $N$ is the size of the support and is exponentially large in the cardinality of the ground set, and $\epsilon$ is the target accuracy that we require in $\ell_1$ distance. Next, we move on to the second topic, i.e. optimization and generalization in neural networks.

## 1.2 Optimization and Generalization in Deep Learning

Over the past decades, deep learning has been quite successful in many learning and prediction tasks [Krizhevsky et al., 2012, Silver et al., 2016, Hinton et al., 2012]; this has triggered a vast interest in trying to mathematically understand the surprising generalization behavior of neural nets [Neyshabur et al., 2015, Bartlett et al., 2017, Dziugaite and Roy, 2017]. Although many approaches have been proposed ignoring the computational properties of the optimization algorithm, it has become clear that in order to understand the generalization phenomenon in deep learning, the training algorithm indeed plays an important role and introduces some kind of implicit regularization [Chizat and Bach, 2018, Mei et al., 2018].

In this regard, many researchers have stepped in to study the landscape of the loss [Nguyen and Hein, 2017, Soltanolkotabi et al., 2018, Kawaguchi, 2016b]. Since the introduction of the concept of Neural Tangent Kernel (NTK) in the work of [Jacot et al., 2018], there has been a line of work in understanding the generalization of neural networks in an algorithmic way by relating the learned network to the RKHS space of the NTK. Later on, Arora et al. [2019a] showed a non-asymptotic analysis by introducing a data dependent complexity measure that captures the generalization behavior of two layer neural networks, all through the lens of NTK. Moreover, there has been an interest in the community in going beyond the NTK, which is the result of a linear approximation, and taking alternative, potentially more powerful viewpoints toward understanding the optimization and generalization properties [Allen-Zhu et al., 2019a, Sirignano and Spiliopoulos, 2020, Javanmard et al., 2020].

In this thesis, we introduce a different point of view on how one can go beyond the NTK analysis by considering a specialized variant of Stochastic gradient descent (SGD) for training a deliberately chosen three-layer network architecture. As a result of going to this new regime that we introduce, we show a generalization of the data-dependent complexity introduced by authors in [Arora et al., 2019a], which also makes it robust when having noise in the labels. To achieve this, we study the nonconvex landscape of the regularized loss function with the $\ell_2$ norm of the weights and show a novel convergence phenomenon. Our techniques root in the fact that SGD can in general escape saddle points of the objective [Ge et al., 2015a].

The work in this thesis regarding the DPP testing problem has been published in the following:

[1] Gatmiry, Khashayar, Maryam Aliakbarpour, and Stefanie Jegelka. "Testing Determinantal Point Processes." Advances in Neural Information Processing Systems 33 (2020): 12779-12791.

[2] Gatmiry, Khashayar, Stefanie Jegelka, and Jonathan Kelner. "Optimization and Adaptive Generalization of Three layer Neural Networks." International Conference on Learning Representations. 2021

# Chapter 2

# Testing DPPs

### Abstract

Determinantal point processes (DPPs) are popular probabilistic models of diversity. In this thesis, we investigate DPPs from a new perspective: property testing of distributions. Given sample access to an unknown distribution $q$ over the subsets of a ground set, we aim to distinguish whether $q$ is a DPP distribution, or $\epsilon$-far from all DPP distributions in $\ell_1$-distance. In this work, we propose the first algorithm for testing DPPs. Furthermore, we establish a matching lower bound on the sample complexity of DPP testing, up to logarithmic factors. This lower bound also implies a new hardness result for the problem of testing the more general class of log-submodular distributions.

## 2.1 Introduction

Determinantal point processes (DPPs) are a rich class of discrete probability distributions that were first studied in the context of quantum physics [Macchi, 1975] and random matrix theory [Dyson, 1962]. Initiated by the seminal work of [Kulesza and Taskar, 2012], DPPs have gained a lot of attention in machine learning, due to their ability to naturally capture notions of diversity and repulsion. Moreover, they are easy to define via a similarity (kernel) matrix, and, as opposed to many other probabilistic models, offer tractable exact algorithms for marginalization, conditioning and sampling [Anari et al., 2016, Hough et al., 2006, Kulesza and Taskar, 2012, Li et al., 2016b]. Therefore, DPPs have been explored in a wide range of applications, including video summarization [Gong et al., 2014b,a], image search [Kulesza and Taskar, 2011b, Affandi et al., 2014], document and timeline summarization Lin and Bilmes [2012], recommendation [Wilhelm et al., 2018], feature selection in bioinformatics [Batmanghelich et al., 2014], modeling neurons [Snoek et al., 2013], and matrix approximation [Dereziński and Mahoney, 2020, Deshpande et al., 2006, Li et al., 2016a].

A *Determinantal Point Process* is a distribution over the subsets of a ground set $[n] = \{1, 2, \dots n\}$, and parameterized by a *marginal kernel* matrix $K \in \mathbb{R}^{n \times n}$ with eigenvalues in $[0, 1]$, whose $(i, j)$th entry expresses the similarity of items $i$ and $j$. Specifically, the marginal probability that a set $A \subseteq [n]$ is observed in a random $\mathcal{J} \sim \mathbf{Pr}_K[.]$ is $\mathbb{P}(A \subseteq \mathcal{J}) = \det(K_A)$, where $K_A$ is the principal submatrix of $K$ indexed by $A$. This implies $\mathbb{P}(\{i, j\} \subseteq \mathcal{J}) = \det(K_{\{i,j\}}) = K_{i,i}K_{j,j} - K_{i,j}^2$ for items $i$ and $j$, which means similar items are less likely to co-occur in $\mathcal{J}$.

Despite the wide theoretical and applied literature on DPPs, one question has not yet been addressed: *Given a sample of subsets, can we test whether it was generated by a DPP?* This question arises, for example, when trying to decide whether a DPP may be a suitable mathematical model for a dataset at hand. To answer this question, we study DPPs from the perspective of *property testing.* Property testing aims to decide whether a given distribution has a property of interest, by observing as few samples as possible. In the past two decades, property testing has received a lot of attention, and questions such as testing uniformity, independence, identity to a known or an unknown given distribution, and monotonicity have been studied in this framework Canonne [2015], Rubinfeld [2012].

More precisely, we ask *How many samples from an unknown distribution are required to distinguish, with high probability, whether it is a DPP or $\epsilon$-far from the class of DPPs in $\ell_1$-distance?* Given the rich mathematical structure of DPPs, one may hope for a tester that is exceptionally efficient. Yet, we show that testing is still not easy, and establish a lower bound of $\Omega(\sqrt{N}/\epsilon^2)$ for the sample size of any valid tester, where $N = 2^n$ is the size of the domain. In fact, this lower bound applies to the broader class of *log-submodular* measures, and may hence be of wider interest given the popularity of submodular set functions in machine learning. Even more generally, the lower bound holds for testing *any* subset of log-submodular distributions that include the uniform measure.

We note that the $\sqrt{N}$ dependence on the domain size is not uncommon in distribution testing, since it is required even for testing simple structures such as uniform distributions [Paninski, 2008]. However, achieving the optimal sample complexity is nontrivial. We provide the first algorithm for testing DPPs; it uses $\tilde{O}(\sqrt{N}/\epsilon^2)$ samples. This algorithm achieves the lower bound and hence settles the complexity of testing DPPs. Moreover, we show how prior knowledge on bounds of the spectrum of $K$ or its entries $K_{ij}$ can improve logarithmic factors in the sample complexity. Our approach relies on *testing via learning*. As a byproduct, our algorithm is the first to provably learn a DPP in $\ell_1$-distance, while previous learning approaches only considered parameter recovery in $K$ [Urschel et al., 2017, Brunel et al., 2017], which does not imply recovery in $\ell_1$-distance.

In short, we make the following contributions:

- We show a lower bound of $\Omega(\sqrt{N}/\epsilon^2)$ for the sample complexity of testing any subset of the class of *log-submodular* measures which includes the uniform measure, implying the same lower bound for testing DPP distributions and strongly Rayleigh [Borcea et al., 2009] measures.

- We provide the first tester for the family of DPP distributions using $\tilde{O}(\sqrt{N}/\epsilon^2)$ samples. The sample complexity is optimal with respect to $\epsilon$ and the domain size $N$, up to logarithmic factors, and does not depend on other parameters. Additional assumptions on $K$ can improve the algorithm's complexity.

- As a byproduct of our algorithm, we give the first algorithm to learn DPP distributions in $\ell_1$ distance.

## 2.2 Related work

**Distribution testing.** *Hypothesis testing* is a classical tool in statistics for inference about the data and model [Neyman and Pearson, 1933, Lehmann and Romano, 2005].

About two decades ago, the framework of *distribution testing* was introduced, to view such statistical problems from a computational perspective [Goldreich and Ron, 2011, Batu et al., 2013]. This framework is a branch of *property testing* [Rubinfeld and Sudan, 1996], and focuses mostly on discrete distributions. Property testing analyzes the non-asymptotic performance of algorithms, i.e., for finite sample sizes. By now, distribution testing has been studied extensively for properties such as uniformity [Paninski, 2008], identity to a known [Batu et al., 2001, Acharya et al., 2015, Diakonikolas et al., 2018] or unknown distribution [Chan et al., 2014, Diakonikolas and Kane, 2016], independence [Batu et al., 2001], monotonicity [Batu et al., 2004, Aliakbarpour et al., 2019], k-modality [Daskalakis et al., 2014], entropy estimation [Batu et al., 2005, Wu and Yang, 2016], and support size estimation [Raskhodnikova et al., 2009, Valiant and Valiant, 2017, Wu et al., 2019]. The surveys [Canonne, 2015, Rubinfeld, 2012] provide further details.

**Testing submodularity and real stability.** Property testing also includes testing properties of functions. As opposed to distribution testing, where observed samples are given, testing functions allows an active query model: given query access to a function $f : \mathcal{X} \to \mathcal{Y}$, the algorithm picks points $x \in \mathcal{X}$ and obtains values $f(x)$. The goal is to determine, with as few queries as possible, whether $f$ has a given property or is $\epsilon$-far from it. Closest to our work in this different model is the question of testing submodularity, in Hamming distance and $\ell_p$-distance Chakrabarty and Huang [2012], Seshadhri and Vondrák [2014], Feldman and Vondrak [2016], Blais and Bommireddi [2016], since any DPP-distribution is log-submodular. In particular, Blais and Bommireddi [2016] show that testing submodularity with respect to any $\ell_p$ norm is feasible with a constant number of queries, independent of the function's domain size. The vast difference between this result and our lower bound for log-submodular distributions lies in the query model – given samples versus active queries – and demonstrates the large impact of the query model. DPPs also belong to the family of *strongly Rayleigh* measures [Borcea et al., 2009], whose generating functions are real stable polynomials. Raghavendra et al. [2017] develop an algorithm for testing real stability of bivariate polynomials, which, if nonnegative, correspond to distributions over two items.

**Learning DPPs.** The problem of learning DPPs has been of great interest in machine learning. Unlike testing, in learning one commonly assumes that the underlying distribution is indeed a DPP, and aims to estimate the marginal kernel $K$. It is well-known that maximum likelihood estimation for DPPs is a highly non-concave optimization problem, conjectured to be NP-hard [Brunel et al., 2017, Kulesza, 2012]. To circumvent this difficulty, previous work imposes additional assumptions, e.g., a parametric family for $K$ [Kulesza and Taskar, 2011b,a, Affandi et al., 2014, Kulesza and Taskar, 2012, Bardenet and AUEB, 2015, Lavancier et al., 2015], or low-rank structure [Gartrell et al., 2016, 2017, Dupuy and Bach, 2018]. A variety of optimization and sampling techniques have been used, e.g., variational methods [Djolonga and Krause, 2014, Gillenwater et al., 2014, Bardenet and AUEB, 2015], MCMC [Affandi et al., 2014], first order methods [Kulesza and Taskar, 2012], and fixed point algorithms [Mariet and Sra, 2015]. Brunel et al. [2017] analyze the asymptotic convergence rate of the Maximum likelihood estimator. To avoid likelihood maximization, Urschel et al. [2017] propose an algorithm based on the method of moments, with statistical guarantees. Its complexity is determined by the *cycle sparsity* property of the DPP. We further discuss the implications of their result in our context in Section 2.4. Using similar techniques, Brunel [2018] considers learning the class of signed DPPs, i.e., DPPs that allow skew-symmetry, $K_{i,j} = \pm K_{j,i}$.

## 2.3 Notation and definitions

Throughout the thesis, we consider discrete probability distributions over subsets of a *ground set* $[n] = \{1, 2, \ldots, n\}$, i.e., over the power set $2^{[n]}$ of size $N := 2^n$. We refer to such distributions via their probability mass function $p : 2^{[n]} \to \mathbb{R}^{\geq 0}$ satisfying $\sum_{S \subseteq [n]} p(S) = 1$. For two distributions $p$ and $q$, we use $\ell_1(q,p) = \frac{1}{2} \sum_{S \subseteq [n]} |q(S) - p(S)|$ to indicate their $\ell_1$ (total variation) distance, and $\chi^2(q,p) = \sum_{S \subseteq [n]} \frac{(q(S)-p(S))^2}{p(S)}$ to indicate their $\chi^2$-distance. Unlike the $\ell_1$-distance, the $\chi^2$-distance is a pseudo-distance, and can be lower bounded as $\chi^2(q,p) \geq 4\ell_1(q,p)^2$ by a simple application of the

Cauchy-Schwarz inequality.

**Determinantal Point Processes (DPPs).** A DPP is a discrete probability distribution parameterized by a positive semidefinite kernel matrix $K \in \mathbb{R}^{n \times n}$, with eigenvalues in $[0, 1]$. More precisely, the marginal probability for any set $S \subseteq [n]$ to occur in a sampled set $\mathcal{J}$ is given by the principal submatrix indexed by rows and columns in $S$: $\mathbf{Pr}_{\mathcal{J} \sim K}[S \subseteq \mathcal{J}] = \det(K_S)$. We refer to the probability mass function of the DPP by $\mathbf{Pr}_K[J] = \mathbf{Pr}_{\mathcal{J} \sim K}[\mathcal{J} = J]$. A simple application of the inclusion-exclusion principle reveals an expression in terms of the complement $\bar{J}$ of $J$:

$$\mathbf{Pr}_K[J] = |\det(K - I_{\bar{J}})|. \tag{2.1}$$

**Distribution testing.** We mathematically define a *property* $\mathcal{P}$ to be a set of distributions. A distribution $q$ has the property $\mathcal{P}$ if $q \in \mathcal{P}$. We say two distributions $p$ and $q$ are *$\epsilon$-far* from (*$\epsilon$-close* to) each other, if and only their $\ell_1$-distance is at least (at most) $\epsilon$. Also, $q$ is $\epsilon$-far from $\mathcal{P}$ if and only if it is $\epsilon$-far from any distribution in $\mathcal{P}$. We define the *$\epsilon$-far set* of $\mathcal{P}$ to be the set of all distributions that are $\epsilon$-far from $\mathcal{P}$. We say an algorithm is an *$(\epsilon, \delta)$-tester* for property $\mathcal{P}$ if, upon receiving samples from an unknown distribution $q$, the following is true with probability at least $1 - \delta$:

- If $q$ has the property $\mathcal{P}$, then the algorithm outputs accept.
- If $q$ is $\epsilon$-far from $\mathcal{P}$, then the algorithm outputs reject.

We refer to $\epsilon$ and $\delta$ as *proximity parameter* and *confidence parameter*, respectively. Note that if we have an $(\epsilon, \delta)$-tester for a property with a confidence parameter $\delta < 0.5$, then we can achieve an $(\epsilon, \delta')$-tester for an arbitrarily small $\delta'$ by multiplying the sample size by an extra factor of $\Theta(\log(\delta/\delta'))$. This *amplification* technique Dubhashi and Panconesi [1998] is a direct implication of the Chernoff bound when we run the initial tester $\Theta(\log(\delta/\delta'))$ times and take the majority output as the answer.

## 2.4 Main results

We begin by summarizing our main results, and explain more proof details in Sections 2.5 and 2.6.

**Upper bound.** Our first result is the first upper bound on the sample complexity of testing DPPs.

**Theorem 1** (Upper Bound). *Given samples from an unknown distribution $q$ over $2^{[n]}$, there exists a deterministic $(\epsilon, 0.01)$-tester for determining whether $q$ is a DPP or it is $\epsilon$-far from all DPP distributions. The tester uses*

$$O(C_{N,\epsilon}\sqrt{N}/\epsilon^2) \tag{2.2}$$

*samples with logarithmic factors $C_{N,\epsilon} = \log^2(N)(\log(N) + \log(1/\epsilon))$.*

Importantly, the sample complexity of our upper bound grows as $\tilde{O}(\sqrt{N}/\epsilon^2)$, which is optimal up to a logarithmic factor (Theorem 2). With additional assumptions on the spectrum and entries of $K$, expressed as $(\alpha, \zeta)$-*normal* DPPs, we obtain a refined analysis.

**Definition 1.** *For $\zeta \in [0, 0.5]$ and $\alpha \in [0, 1]$, a DPP with marginal kernel $K$ is* $(\alpha, \zeta)$-normal *if:*

1. *the eigenvalues of $K$ are in the range $[\zeta, 1-\zeta]$; and*
2. *for $i, j \in [n] : K_{i,j} \neq 0 \Rightarrow |K_{i,j}| \geq \alpha$.*

The notion of $\alpha$-normal DPPs was also used in Urschel et al. [2017]. Since $K$ has eigenvalues in $[0, 1]$, its entries $K_{i,j}$ are at most one. Hence, we always assume $0 \leq \zeta \leq 0.5$ and $0 \leq \alpha \leq 1$.

**Lemma 1.** *For $(\alpha, \zeta)$-normal DPPs, with knowledge of $\alpha$ and $\zeta$, the factor in Theorem 1 becomes $C'_{N,\epsilon,\zeta,\alpha} = \log^2(N)(1 + \log(1/\zeta) + \min\{\log(1/\epsilon), \log(1/\alpha)\})$.*

Even more, if at least one of $\epsilon$ or $\alpha$ is not too small, i.e., if $\epsilon = \tilde{\Omega}(\zeta^{-2}N^{-1/4})$ or $\alpha = \tilde{\Omega}(\zeta^{-1}N^{-1/4})$ hold, then $C'_{N,\epsilon,\zeta,\alpha}$ reduces to $\log^2(N)$. With a minor change in the algorithm, the bound in Lemma 1 also holds for the problem of testing whether $q$ is an $(\alpha, \zeta)$-normal DPP, or $\epsilon$-far only from just the class of $(\alpha, \zeta)$-normal DPPs, instead of all DPPs (Section 2.13).

Our approach tests DPP distributions via *learning*: At a high-level, we learn a DPP model from the data as if the data were generated from a DPP distribution. Then, we use a new batch of data and test whether the DPP we learnt seems to have generated the new batch of the data. More accurately, given samples from $q$, we pretend $q$ is a DPP with kernel $K^*$, and use a proper learning algorithm to estimate a kernel matrix $\hat{K}$.

But, Urschel et al. [2017] derive a lower bound on the complexity of learning $K^*$ which, in the worst case, may lead to a sub-optimal sample complexity for testing. To reduce the sample complexity of learning, we do not work with a single accurate estimate $\hat{K}$, but construct a set $\mathcal{M}$ of candidate DPPs as potential estimates for $q$. We show that, with only $\Theta(\sqrt{N}/\epsilon^2)$ samples, we can obtain a set $\mathcal{M}$ such that, with high probability, we can determine if $q$ is a DPP by testing if $q$ is close to any DPP in $\mathcal{M}$. We prove that $\Theta(\log(|\mathcal{M}|)\sqrt{N}/\epsilon^2)$ samples suffice for this algorithm to succeed with high probability.

Small-scale experiments in Section 2.17 validate the algorithm empirically.

**Lower Bound.** Our second main result is an information-theoretic lower bound, which shows that the sample complexity of our tester in Theorem 1 is optimal up to logarithmic factors.

**Theorem 2** (Lower Bound). *Given $\epsilon \leq 0.0005$ and $n \geq 22$, any $(\epsilon, 0.01)$-tester needs at least $\Omega(\sqrt{N}/\epsilon^2)$ samples to distinguish if $q$ is a DPP or it is $\epsilon$-far from the class of DPPs.*

*Given $\alpha \in [0, 0.5]$, the same bound holds for distinguishing if $q$ is an $(\alpha, \zeta)$-normal DPP or it is $\epsilon$-far from the class of DPPs (or $\epsilon$-far from the class of $(\alpha, \zeta)$-normal DPPs).*

In fact, we prove a more general result (Theorem 4): testing whether $q$ is in any subclass $\Upsilon$ of the family of log-submodular distributions that includes the uniform distribution requires $\Omega(\sqrt{N}/\epsilon^2)$ samples. DPPs are such a subclass [Kulesza and Taskar, 2012]. A distribution $f$ over $2^{[n]}$ is *log-submodular* if for every $S \subset S' \subseteq [n]$ and $i \notin S'$, it holds that $\log(f(S' \cup \{i\})) - \log(f(S')) \leq \log(f(S \cup \{i\})) - \log(f(S))$. Given the interest in log-submodular distributions [Djolonga and Krause, 2014, Tschiatschek et al., 2016, Djolonga et al., 2018, Gotovos et al., 2015, 2018], this result may be of wider interest. Moreover, our lower bound applies to another important subclass $\Upsilon$, *strongly Rayleigh measures* [Borcea et al., 2009], which underlie recent progress in algorithms and mathematics [Gharan et al., 2011, Frieze et al., 2014, Spielman and Srivastava, 2011, Anari and Gharan, 2015], and sampling in machine learning [Anari et al., 2016, Li et al., 2017, 2016b].

Our lower bound stands in stark contrast to the *constant* sample complexity of testing whether a given function is submodular Blais and Bommireddi [2016], implying a wide complexity gap between access to given samples and access to an evaluation oracle (see Section 2.2). We prove our lower bounds by a reduction from a randomized instance of uniformity testing.

## 2.5 An Algorithm for Testing DPPs

We first construct an algorithm for testing the smaller class of $(\alpha, \zeta)$-normal DPPs, and then show how to extend this result to all DPPs via a coupling argument.

Our testing algorithm relies on learning: given samples from $q$, we estimate a kernel $\hat{K}$ from the data, and then test whether the estimated DPP has generated the observed samples. The magnitude of any entry $\hat{K}_{i,j}$ can be estimated from the marginals for $S = \{i, j\}$ and $i, j$, since $\mathbf{Pr}_K[S] = K_{i,i}K_{j,j} - K_{i,j}^2 = \mathbf{Pr}_K[i]\mathbf{Pr}_K[j] - K_{i,j}^2$. But, determining the signs is more challenging. Urschel et al. [2017] estimate signs via higher order moments that are harder to estimate, but it is not clear whether the resulting $\hat{K}$ yields a sufficiently accurate estimate of the distribution to obtain an optimal sample complexity for testing. Hence, instead, we construct a set $\mathcal{M}$ of candidate DPPs such that, with high probability, there is a $\tilde{p} \in \mathcal{M}$ that is close to $q$ if and only if $q$ is a DPP. Our tester, Algorithm 1, tests closeness to $\mathcal{M}$ by individually

**Algorithm 1** DPP-Tester

```
procedure DPP-Tester($\epsilon$, $\delta$, sample access to $q$)
    $\mathcal{M} \leftarrow$ construct the set of DPP distributions as described in Theorem 3.
    for each $p$ in $\mathcal{M}$ do
        Use robust $\chi^2 - \ell_1$ testing to check if $\chi^2(q,p) \leq \epsilon^2/500$, or $\ell_1(q,p) \geq \epsilon$.
        if the tester outputs accept then
            Return accept.
    Return reject
```

testing closeness of each candidate in $\mathcal{M}$.

**Constructing $\mathcal{M}$.** The DPPs in $\mathcal{M}$ arise from variations of an estimate for $K^*$, obtained with $\Theta(\sqrt{N}/\epsilon^2)$ samples. Via the above strategy, we first estimate the magnitude $|K^*_{ij}|$ of each matrix entry. Separating the case $K^*_{ij} = 0$, one can compute confidence intervals for this estimation around $+|\widehat{K}_{ij}|$ and $-|\widehat{K}_{ij}|$. We then pick candidate entries from these confidence intervals, such that at least one is close to the true $K^*_{i,j}$. The candidate matrices $K$ are obtained by all possible combinations of candidate entries Since these are not necessarily valid marginal kernels, we project them onto the positive semidefinite matrices with eigenvalues in $[0, 1]$. Then, $\mathcal{M}$ is the set of all DPPs parameterized by these projected candidate matrices $\Pi(K)$. Its cardinality is given in Theorem 3 and, as an explicit function of $N$ and $\epsilon$, in Section 2.15.

If $q$ is a DPP with kernel $K^*$, then, by construction, our candidates contain a $\tilde{K}$ close to $K^*$. The projection of $\tilde{K}$ remains close to $K^*$ in Frobenius distance. We show that this closeness of the matrices implies closeness of the corresponding distributions $q$ and $\tilde{p} = \mathbf{Pr}_{\Pi(\tilde{K})}[.]$ in $\ell_1$-distance: $\ell_1(q, \tilde{p}) = O(\epsilon)$. Conversely, if $q$ is $\epsilon$-far from being a DPP, then it is, by definition, $\epsilon$-far from $\mathcal{M}$, which is a subset of all DPPs.

**Testing $\mathcal{M}$.** To test whether $q$ is close to $\mathcal{M}$, a first idea is to do robust $\ell_1$ identity testing, i.e., for every $p \in \mathcal{M}$, test whether $\ell_1(q,p) \geq \epsilon$ or $\ell_1(q,p) = O(\epsilon)$. But, $\mathcal{M}$ can contain the uniform distribution, and it is known that robust $\ell_1$ uniformity testing needs $\Omega(N/\log N)$ samples [Valiant and Valiant, 2017], as opposed to the optimal dependence $\sqrt{N}$.

Hence, instead, we use a combination of $\chi^2$ and $\ell_1$ distances for testing, and test

$\chi^2(q,p) = O(\epsilon^2)$ versus $\ell_1(q,p) \geq \epsilon$. This is possible with fewer samples [Acharya et al., 2015]. To apply this robust $\chi^2$-$\ell_1$ identity testing (described in Section 2.5.1), we must prove that, with high probability, there is a $\tilde{p}$ in $\mathcal{M}$ with $\chi^2(q,\tilde{p}) = O(\epsilon^2)$ if and only if $q$ is a DPP. Theorem 3, proved in Section 2.8, asserts this result if $q$ is an $(\alpha,\zeta)$-normal DPP. This is stronger than its $\ell_1$ correspondent, since $4\ell_1^2(q,\tilde{p}) \leq \chi^2(q,\tilde{p})$.

To prove Theorem 3, we need to control the distance between the atom probabilities of $q$ and $\tilde{p}$. We analyze these atom probabilities, which are given by determinants, via a lower bound on the smallest singular values $\sigma_n$ of the family of matrices $\{K - I_{\bar{J}} : J \subseteq [n]\}$.

**Lemma 2.** *If the kernel matrix $K$ has all eigenvalues in $[\zeta, 1-\zeta]$, then, for every $J \subseteq [n]$:*

$$\sigma_n(K - I_{\bar{J}}) \geq \zeta(1-\zeta)/\sqrt{2}.$$

Lemma 2 is proved in Section 2.9. In Theorem 3, we observe $m = \lceil (\ln(1/\delta) + 1)\sqrt{N}/\epsilon^2 \rceil$ samples from $q$, and use the parameter $\varsigma := \lceil 200 n^2 \zeta^{-1} \min\{2\xi/\alpha, \sqrt{\xi/\epsilon}\} \rceil$, with $\xi := N^{-\frac{1}{4}}\sqrt{\log(n)+1}$.

**Theorem 3.** *Let $q$ be an $(\alpha,\zeta)$-normal DPP distribution with marginal kernel $K^*$. Given the parameters defined above, suppose we have $m$ samples from $q$. Then, one can generate a set $\mathcal{M}$ of DPP distributions of cardinality $|\mathcal{M}| = (2\varsigma + 1)^{n^2}$, with $\varsigma$ defined as above, such that, with probability at least $1-\delta$, there is a distribution $\tilde{p} \in \mathcal{M}$ with $\chi^2(q,\tilde{p}) \leq \epsilon^2/500$.*

### 2.5.1 Correctness of the Testing Algorithm for $(\alpha,\zeta)$-normal DPPs

Next, we show that with high probability, our resulting testing algorithm succeeds with high probability. This finishes the proof of Lemma 1. For simplicity, we set the confidence parameter in Algorithm 1 to $\delta = 0.01$. In this case, DPP-Tester aims to output accept if $q$ is a $(\alpha,\zeta)$-normal DPP, and reject if $q$ is $\epsilon$-far from all DPPs, in both cases with probability at least 0.99.

To finish the proof for the adaptive sample complexity, we need to argue that our DPP-Tester succeeds with high probability, i.e., that with high probability all of the identity tests, with each $p \in \mathcal{M}$, succeed. The algorithm uses robust $\chi^2$-$\ell_1$ identity testing [Acharya et al., 2015], to test $\chi^2(q,p) \leq \epsilon^2/500$ versus $\ell_1(q,p) \geq \epsilon$. In our framework, the $\chi^2$-$\ell_1$ identity tester works as follows. It uses a Poissonization trick that simplifies the analysis. Given the average sample size $m$, the $\chi^2$-$\ell_1$ tester first samples $m' \sim \text{Poisson}(m)$, then obtains $m'$ samples from $q$. For each $p \in \mathcal{M}$, it then computes the statistic

$$Z^{(m)} = \sum_{J \subseteq [n]:\, p(J) \geq \epsilon/50N} \frac{(N(J) - mp(J))^2 - N(J)}{mp(J)}, \tag{2.3}$$

where $N(J)$ is the number of samples that are equal to set $J$, and compares $Z^{(m)}$ with the threshold $C = m\epsilon^2/10$.

Acharya et al. [2015] show that for $m = \Theta(\sqrt{N}/\epsilon^2)$, $Z^{(m)}$ concentrates around its mean, which is strictly below $C$ if $p$ satisfies $\chi^2(q,p) \leq \epsilon^2/500$, and strictly above $C$ if $\ell_1(q,p) \geq \epsilon$. Let $\mathcal{E}_1$ be the event that all these robust tests, for every $p \in \mathcal{M}$, simultaneously answer correctly. To make sure that $\mathcal{E}_1$ happens with high probability, we use amplification (Section 2.3): while we use the same set of samples to test against every $p \in \mathcal{M}$, we multiply the sample size by $\Theta(\log(|\mathcal{M}|))$ to be confident that each test answers correctly with probability at least $1 - O(|\mathcal{M}|^{-1})$. A union bound then implies that $\mathcal{E}_1$ happens with arbitrarily large constant probability.

Theorem 3 states that, indeed, with $\Theta(\sqrt{N}/\epsilon^2)$ samples, if $q$ is an $(\alpha, \zeta)$-normal DPP, then $\mathcal{M}$ contains a distribution $\tilde{p}$ such that $\chi^2(q, \tilde{p}) \leq \epsilon^2/500$, with high probability. We call this event $\mathcal{E}_2$. DPP-Tester succeeds in the case $\mathcal{E}_1 \cap \mathcal{E}_2$: If $q$ is an $(\alpha, \zeta)$-normal DPP, then at least one $\chi^2$-$\ell_1$ test accepts $\tilde{p}$ and consequently the algorithm accepts $q$ as a DPP. Conversely, if $q$ is $\epsilon$-far from all DPPs, then $\ell_1(q,p) \geq \epsilon$ for every $p \in \mathcal{M}$, so all the $\chi^2$-$\ell_1$ tests reject simultaneously and DPP-Tester rejects $q$ as well. With a union bound on the events $\mathcal{E}_1^c$ and $\mathcal{E}_2^c$, it follows that $\mathcal{E}_1 \cap \mathcal{E}_2$ happens with arbitrarily large constant probability too, independent of whether $q$ is a DPP or not.

Adding the sample complexities for generating $\mathcal{M}$ and for the $\chi^2$-$\ell_1$ tests and observing $\log(|\mathcal{M}|) = O(1 + \log(1/\zeta) + \min\{\log(1/\epsilon), \log(1/\alpha)\})$ completes the proof of Lemma 1.

### 2.5.2 Extension to general DPPs

Next, we generalize our testing result from $(\alpha, \zeta)$-normal DPPs to general DPPs to prove the general sample complexity in Theorem 1. The key idea is that, if some eigenvalue of $K^*$ is very close to zero or one, we couple the process of sampling from $K^*$ with sampling from another kernel $\Pi_z(K^*)$ whose eigenvalues are bounded away from zero and one, i.e., parameterizing a $(0, z)$-normal DPP. This coupling enables us to test $(0, z)$-normal DPPs instead, by tolerating an extra failure probability, and transfer the above analysis for $(\alpha, \zeta)$-normal DPPs. We state our coupling argument in the following Lemma, proved in Section 2.11.

**Lemma 3.** *For a value $z \in [0, 1]$, we denote the projection of a marginal kernel $K$ onto the convex set $\{A \in S_n^+ |\ zI \leq A \leq (1 - z)I\}$ by $\Pi_z(K)$, where $S_n^+$ is the set of positive semidefinite matrices. For $z = \delta/2mn$, consider the following stochastic processes:*

1. *derive $m$ i.i.d samples $\{\mathcal{J}_K^{(t)}\}_{t=1}^m$ from $\mathbf{Pr}_K[.]$;*
2. *derive $m$ i.i.d samples $\{\mathcal{J}_{\Pi_z(K)}^{(t)}\}_{t=1}^m$ from $\mathbf{Pr}_{\Pi_z(K)}[.]$.*

*There exists a coupling between (1) and (2) such that*

$$\mathbf{Pr}_{\textit{coupling}}\Big[\{\mathcal{J}_K^{(t)}\}_{t=1}^m = \{\mathcal{J}_{\Pi_z(K)}^{(t)}\}_{t=1}^m\Big] \geq 1 - \delta.$$

We can use this coupling argument as follows. Suppose the constant $c_1$ is such that using $c_1 C_{N,\epsilon,\alpha,\zeta}\sqrt{N}/\epsilon^2$ samples suffice for `DPP-Tester` to output the correct answer for testing $(\alpha, \zeta)$-normal DPPs, with probability at least 0.995. Such a constant exists as we just proved. Now, we show that with $m^* = c_2 C_{N,\epsilon}\sqrt{N}/\epsilon^2$ samples for large enough constant $c_2$, we obtain a tester for the set of all DPPs. To this end, we use the parameter setting of our algorithm for $(0, \bar{z})$ normal DPPs, where $\bar{z} = 0.005/(2m^* n)$ is a function of $c_2$, $\epsilon$, and $N$. One can readily see that $c_2$ can be picked large enough, such that $m^* \geq c_1 C_{N,\epsilon,0,\bar{z}} \sqrt{N}/\epsilon^2$, with $c_2$ being just a function of $c_1$. This way, by the definition of $c_1$, the algorithm can test for $(0, \bar{z})$-normal DPPs with success probability 0.995. So, if $q$ is a $(0, \bar{z})$-normal DPP, or if it is $\epsilon$-far from all DPPs, then the algorithm outputs correctly with probability at least 0.995.

It remains to check what happens when $q$ is a DPP with kernel $K^*$, but not $(0, \bar{z})$-normal. Indeed, DPP-Tester successfully decides this case too: due to our coupling, the product distributions $\mathbf{Pr}^{(m^*)}_{K^*}[.]$ and $\mathbf{Pr}^{(m^*)}_{\Pi_{\bar{z}}(K^*)}[.]$ over the space of data sets have $\ell_1$-distance at most 0.005, so we have
$\mathbf{Pr}^{(m^*)}_{K^*}[\text{Acceptance Region}] \geq \mathbf{Pr}^{(m^*)}_{\Pi_{\bar{z}}(K^*)}[\text{Acceptance Region}] - 0.005 \geq 0.995 - 0.005 = 0.99$, where the last inequality follows from the fact that $\Pi_{\bar{z}}(K^*)$ is an $(0, \bar{z})$-normal DPP. Hence, for such $c_2$, DPP-Tester succeeds with $c_2 C_{N,\epsilon} \sqrt{N}/\epsilon^2$ samples to test all DPPs with probability 0.99, which completes the proof of Theorem 1.

**Learning DPPs.** Our tester implicitly provides a method to learn a DPP $q$ in $\ell_1$-distance: the $\chi^2 - \ell_1$ tester can only accept candidate DPPs $p \in \mathcal{M}$ for which we either have $\chi^2(q,p) \leq \epsilon^2/500$ or $\ell_1(q,p) < \epsilon$. Since $\ell_1(q,p) \leq 1/2\sqrt{\chi^2(q,p)} < \epsilon$, any such $p$ is a DPP with distance $\ell_1(q,p) \leq \epsilon$ to the underlying distribution $q$. If $q$ is a DPP, we will find such a $p$ with high probability.

## 2.6 Lower bound

Next, we establish the lower bound in Theorem 2 for testing DPPs, which implies that the sample complexity of DPP-Tester is tight up to logarithmic factors. In fact, our lower bound is more general: it applies to the problem of testing any subset $\Upsilon$ of the larger class of log-submodular distributions, whenever $\Upsilon$ includes the uniform measure:

**Theorem 4.** *Let $\Upsilon$ be any subset of log-submodular distributions that contains the uniform measure. For $\epsilon \leq 0.0005$ and $n \geq 22$, given sample access to a distribution $q$ over subsets of $[n]$, any $(\epsilon, 0.01)$-tester that checks whether $q \in \Upsilon$ or $q$ is $\epsilon$-far from all* *log-submodular distributions requires $\Omega(\sqrt{N}/\epsilon^2)$ samples.*

One may also wish to test if $q$ is $\epsilon$-far only from the distributions in $\Upsilon$. A tester for this question, however, would correctly return reject for any $q$ that is $\epsilon$-far from the set of all log-submodular distributions, and can hence distinguish the cases in Theorem 4 too. Hence, the lower bound extends to this question.

Theorem 2 is simply a consequence of Theorem 4: we may set $\Upsilon$ to be the set of all DPPs, or all $(\alpha, \zeta)$-normal DPPs. Both classes include the uniform distribution over $2^{[n]}$, which is an $(\alpha, \zeta)$-normal DPP with marginal kernel $I/2$, where $I$ is the $n \times n$ identity matrix. By the same argument, the lower bound applies to distinguishing $(\alpha, \zeta)$-normal DPPs from the $\epsilon$-far set of all DPPs for $\alpha \in [0, 0.5]$.

**Proof of Theorem 4.** To prove Theorem 4, we construct a hard uniformity testing problem that can be decided by our desired tester for $\Upsilon$. In particular, we construct a family $\mathcal{F}$, such that it is hard to distinguish between the uniform measure and a randomly selected distribution $h$ from $\mathcal{F}$. While the uniform measure is in $\Upsilon$, we will show that $h$ is far from the set of log-submodular distributions with high probability. Hence, a tester for $\Upsilon$ can, with high probability, correctly decide between $\mathcal{F}$ and the uniform measure.

We obtain $\mathcal{F}$ by randomly perturbing the atom probabilities of the uniform measure over $2^{[n]}$ by $\pm\epsilon'/N$, with $\epsilon' = c \cdot \epsilon$ for a sufficiently large constant $c$ (specified in the later sections). More concretely, for every vector $r \in \{\pm 1\}^N$ whose entries are indexed by the subsets $S \subseteq [n]$, we define the distribution $h_r \in \mathcal{F}$ as

$$\forall S \subseteq [n]: \quad h_r(S) \propto \bar{h}_r(S) = \frac{1 + r_S \epsilon'}{N},$$

where $\bar{h}_r$ is the corresponding unnormalized measure.

We assume that $h_r$ is selected from $\mathcal{F}$ uniformly at random, i.e., each entry $r_S$ is a Rademacher random variable independent from the others. In particular, it is known that distinguishing such a random $h_r$ from the uniform distribution requires $\Omega(\sqrt{N}/\epsilon'^2)$ samples [Diakonikolas and Kane, 2016, Paninski, 2008].

To reduce this uniformity testing problem to our testing problem for $\Upsilon$ and obtain the lower bound $\Omega(\sqrt{N}/\epsilon'^2) = \Omega(\sqrt{N}/\epsilon^2)$ for the sample complexity of our problem, it remains to prove that $h_r$ is $\epsilon$-far from the class of log-submodular distributions with high probability. Hence, Lemma 4 finishes the proof.

**Lemma 4.** *For $\epsilon \leq 0.0005$, $n \geq 22$ and $c = 1024$, a distribution $h_r$ drawn uniformly from $\mathcal{F}$ is $\epsilon$-far from all log-submodular distributions with probability at least* 0.99*.*

**Proof sketch for Lemma 4.** We fix an arbitrary log-submodular distribution $f$ and first show that (1) the $\ell_1$-distance $\ell_1(f, \bar{h}_r)$ between $f$ and the unnormalized measure $\bar{h}_r$ is large with high probability, independent of $f$ (we define the $\ell_1$-distance of general measures the same as for probability measures). Then, (2) we show that if $\ell_1(f, \bar{h}_r)$ is large, $\ell_1(f, h_r)$ is also large.

To address (1), we define a family $\mathcal{S}_r$ of subsets that, as we prove, satisfies:

(P1) With high probability, $\mathcal{S}_r$ has cardinality at least $N/64$.

(P2) For every $S \in \mathcal{S}_r$, there is a contribution of at least $\epsilon'/8N$ to $\ell_1(f, \bar{h}_r)$ from the term $V_S$ defined as

$$\begin{aligned} V_S :=& \tfrac{1}{2}|\bar{h}_r(S) - f(S)| + \tfrac{1}{2}|\bar{h}_r(S \cup \{1\}) - f(S \cup \{1\})| + \\ & \tfrac{1}{2}|\bar{h}_r(S \cup \{2\}) - f(S \cup \{2\})| + \tfrac{1}{2}|\bar{h}_r(S \cup \{1,2\}) - f(S \cup \{1,2\})|. \end{aligned}$$

Together, the above properties imply that $\ell_1(\bar{h}_r, f) \geq (N/64) \times (\epsilon'/8N) = \epsilon'/512$.

We define the important family $\mathcal{S}_r$ as

$$\mathcal{S}_r := \{S \subseteq [n] \setminus \{1,2\} \,|\, r_{(S \cup \{1,2\})} = 1,\ r_{(S \cup \{2\})} = -1,\ r_{(S \cup \{1\})} = -1\}.$$

Property (P1) follows from a Chernoff bound for the random variables $\mathbb{1}\{S \in \mathcal{S}_r\}$, $\forall S \subseteq [n] \setminus \{1,2\}$, where $\mathbb{1}\{.\}$ is the indicator function. For proving Property P2, we distinguish two cases, depending on the ratio $f((S \cup \{1,2\})/f(S \cup \{2\})$. One of these cases relies on the definition of log-submodularity.

Finally, to show that (2) a large $\ell_1(f, \bar{h}_r)$ implies a large $\ell_1(f, h_r)$, we control the normalization constant $\sum_{S \subseteq [n]} \bar{h}_r(S)$. The full proof may be found in Section 2.10.

## 2.7 Discussion

In this thesis, we initiate the study of distribution testing for DPPs. Our lower bound of $\Omega(\sqrt{N}/\epsilon^2)$, where $N$ is the domain size $2^n$, shows that, despite the rich mathematical structure of DPPs, testing whether $q$ is a DPP or $\epsilon$-far from it has a sample complexity similar to uniformity testing. This bound extends to related distributions that have gained interest in machine learning, namely log-submodular distributions and strongly Rayleigh measures. Our algorithm DPP-Tester demonstrates that this lower bound is tight for DPPs, via an almost matching upper bound of $\tilde{O}(\sqrt{N}/\epsilon^2)$.

One may wonder what changes when using the moment-based learning algorithm from [Urschel et al., 2017] instead of the learner from Section 2.5, which yields optimal testing sample complexity. With [Urschel et al., 2017], we obtain a single estimate $\hat{K}^{\text{new}}$ for $K^*$, apply a single robust $\chi^2$-$\ell_1$ test against $\mathbf{Pr}_{\hat{K}^{\text{new}}}[.]$, and return its output. The resulting algorithm DPP-Tester2 shows a statistical-computational tradeoff: since it performs only one test, it gains in running time, but its sample complexity could be larger: Theorem 5, proved in Section 2.14, states upper bounds that are is no longer logarithmic in $\alpha$ and $\zeta$, and larger than $O(\sqrt{N}/\epsilon^2)$.

**Theorem 5.** *To test against the class of $(\alpha, \zeta)$-normal DPPs, DPP-Tester2 needs $O\Big(n^4 \log(n)/\epsilon^2\alpha^2\zeta^2 + \ell(4/\alpha)^{2\ell}\log(n) + \sqrt{N}/\epsilon^2\Big)$ samples, and runs in time $O(Nn^3 + n^6 + mn^2)$, where $m$ is the number of input samples and $\ell$ is the cycle sparsity[1] of the graph corresponding to the non-zero entries of $K^*$.*

Assuming a constant cycle sparsity may improve the sample complexity, but our lower bound still applies even with assumptions on cycle sparsity.

While the results in this thesis focus on sample complexity for general DPPs, it is an interesting avenue of future work to study whether additional structural assumptions, or a widening to strongly log-concave measures [Anari et al., 2018, 2019], can lead to further statistical and computational benefits or tradeoffs.

[1] The cycle sparsity of a graph is the smallest $\ell'$ such that the cycles with length at most $\ell'$ constitute a basis for the cycle space of the graph.

## 2.8 Proof of the Learning Guarantee

In this section, we prove Theorem 3. First, recall the definition of $(\alpha, \zeta)$-normal DPPs (Definition 1) below.

**Definition 1.** *For $\zeta \in [0, 0.5]$ and $\alpha \in [0, 1]$, a DPP with marginal kernel $K$ is* $(\alpha, \zeta)$-normal *if:*

1. *The eigenvalues of $K$ are in the range $[\zeta, 1-\zeta]$; and*

2. *For $i, j \in [n] : K_{i,j} \neq 0 \Rightarrow |K_{i,j}| \geq \alpha$.*

We assume that $n$ is the size of the ground set with $N = 2^n$. We set $m = \lceil (\ln(1/\delta) + 1)\sqrt{N}/\epsilon^2 \rceil$ to be the number of samples, and use the parameter $\varsigma := \lceil 200n^2\zeta^{-1} \min\{2\xi/\alpha, \sqrt{\xi/\epsilon}\} \rceil$, with $\xi := N^{-\frac{1}{4}}\sqrt{\log(n) + 1}$. Below, we restate Theorem 3 for convenience.

**Theorem 3.** *Let $q$ be an $(\alpha, \zeta)$-normal DPP distribution with marginal kernel $K^*$. Given the parameters defined above, suppose we have $m$ samples from $q$. Then, one can generate a set $\mathcal{M}$ of DPP distributions with cardinality $|\mathcal{M}| = (2\varsigma + 1)^{n^2}$, such that, with probability at least $1 - \delta$, there is a distribution $\tilde{p} \in \mathcal{M}$ with $\chi^2(q, \tilde{p}) \leq \epsilon^2/500$.*

*Proof of Theorem 3.* To prove Theorem 3, first we estimate each entry of the marginal kernel $K^*$ and generate the set $\mathcal{M}$ of our candidate DPPs, which contains a DPP $\tilde{p} \in \mathcal{M}$ whose marginal kernel is close to $K^*$ in the Frobenius distance. Then, we show that that the closeness between the marginal kernels of $\tilde{p}$ and $q$ implies the desired upper bound in $\chi^2$-distance and $\ell_1$-distance of the two distributions. We start by introducing the initial estimate $\hat{K}$ which is obtained by estimating the entries of $K^*$ from our samples.

**Estimating entries of $K^*$:** Note that one can write the entries of the matrix

$K^*$ in terms of the marginal probabilities of subsets of size one and two as follows:

$$\mathbf{Pr}_{\mathcal{J}\sim K^*}[i \in \mathcal{J}] = \det\left(\left[K^*_{i,i}\right]\right) = K^*_{i,i}, \tag{2.4}$$

$$\mathbf{Pr}_{\mathcal{J}\sim K^*}[\{i,j\} \subseteq \mathcal{J}] = \det\left(\begin{bmatrix} K^*_{i,i} & K^*_{i,j} \\ K^*_{j,i} & K^*_{j,j} \end{bmatrix}\right) = K^*_{i,i}K^*_{j,j} - {K^*_{i,j}}^2. \tag{2.5}$$

Given the sampled subsets $\{\mathcal{J}^{(t)}\}_{t=1}^m$, we can estimate the above marginal probabilities using the number of appearances of every single element and every pair of elements among $\mathcal{J}^{(1)}, \mathcal{J}^{(2)}, ..., \mathcal{J}^{(m)}$. We use $\mathbb{1}E$ to denote the indicator variable of the event $E$. For each $i \in [n]$, we estimate $K^*_{i,i}$ by the average of the $\mathbb{1}\{\{i\} \subseteq \mathcal{J}^{(t)}\}$'s:

$$\hat{K}_{i,i} \coloneqq \frac{1}{m}\sum_{t=1}^{m} \mathbb{1}\{\{i\} \subseteq \mathcal{J}^{(t)}\}\,.$$

We also denote the averages of the $\mathbb{1}\{\{i,j\} \subseteq \mathcal{J}^{(t)}\}$'s by $\hat{u}_{i,j}$.

$$\hat{u}_{i,j} \coloneqq \frac{1}{m}\sum_{t=1}^{m} \mathbb{1}\{\{i,j\} \subseteq \mathcal{J}^{(t)}\}\,.$$

Using the estimates $\hat{u}_{i,j}$, $\hat{K}_{i,i}$, and $\hat{K}_{j,j}$, we can also estimate ${K^*_{i,j}}^2$ by the term $\hat{K}_{i,i}\hat{K}_{j,j} - \hat{u}_{i,j}$, based on Equation (2.5). To derive confidence intervals for our estimates, we use the Hoeffding bound and a union bound, which implies that with probability at least $1-\delta$:

$$\forall i \in [n]: \quad \hat{K}_{i,i} \in \left[\mathbf{Pr}_{\mathcal{J}\sim K^*}[i \subseteq \mathcal{J}] - \xi\epsilon\ ,\ \mathbf{Pr}_{\mathcal{J}\sim K^*}[i \subseteq \mathcal{J}] + \xi\epsilon\right], \tag{2.6}$$

$$\forall \{i,j\} \subseteq [n], i \neq j: \quad \hat{u}_{i,j} \in \left[\mathbf{Pr}_{\mathcal{J}\sim K^*}[\{i,j\} \subseteq \mathcal{J}] - \xi\epsilon\ ,\ \mathbf{Pr}_{\mathcal{J}\sim K^*}[\{i,j\} \subseteq \mathcal{J}] + \xi\epsilon\right]\,, \tag{2.7}$$

where $\xi \coloneqq N^{-\frac{1}{4}}\sqrt{\log(n)+1}$. Note that Equation (2.5) does not reveal any information about the sign of $K^*_{i,j}$. However, we can estimate its magnitude $|K^*_{i,j}|$. Thus, we consider the following two estimates for $K^*_{i,j}$:

$$\forall \{i,j\} \subseteq [n], i \neq j : \quad \begin{aligned} \hat{K}^{(+)}_{i,j} &:= \sqrt{\max\{\hat{K}_{i,i}\hat{K}_{j,j} - \hat{u}_{i,j}\ ,\ 0\}}\,, \\ \hat{K}^{(-)}_{i,j} &:= -\sqrt{\max\{\hat{K}_{i,i}\hat{K}_{j,j} - \hat{u}_{i,j}\ ,\ 0\}}\,. \end{aligned} \tag{2.8}$$

Now, let $\hat{K}_{i,j}$ be whichever of $\hat{K}^{(+)}_{i,j}$ or $\hat{K}^{(-)}_{i,j}$ that has the same sign as $K^*_{i,j}$. Then, according to Equations (2.6), (2.7), and (2.8), we achieve:

$$\begin{aligned} \left|\hat{K}^2_{i,j} - {K^*}^2_{i,j}\right| &\leq \left|K^*_{i,i}K^*_{j,j} - \hat{K}_{i,i}\hat{K}_{j,j}\right| + |\mathbf{Pr}_{\mathcal{J}\sim K^*}[\{i,j\} \subseteq \mathcal{J}] - \hat{u}_{i,j}| \\ &\leq \max\{|(K^*_{i,i} + \xi\epsilon)(K^*_{j,j} + \xi\epsilon) - K^*_{i,i}K^*_{j,j}|, |(K^*_{i,i} - \xi\epsilon)(K^*_{j,j} - \xi\epsilon) - K^*_{i,i}K^*_{j,j}|\} + \xi\epsilon \\ &\leq 3\xi\epsilon + (\xi\epsilon)^2 \leq 4\xi\epsilon, \end{aligned}$$

where we used $\xi\epsilon \leq 1$ and that $\forall i,j \in [n] : |K^*_{i,j}| \leq 1$. Moreover, using the fact that $\hat{K}_{i,j}$ and $K^*_{i,j}$ have the same sign,

$$|\hat{K}_{i,j} - K^*_{i,j}|^2 \leq |\hat{K}_{i,j} - K^*_{i,j}||\hat{K}_{i,j} + K^*_{i,j}| = |\hat{K}^2_{i,j} - {K^*}^2_{i,j}| \leq 4\xi\epsilon,$$

which gives

$$|\hat{K}_{i,j} - K^*_{i,j}| \leq 2\sqrt{\xi\epsilon}. \tag{2.9}$$

On the other hand, we have the lower bound $\alpha$ on the absolute value of the non-zero entries of $K^*$ from the $\alpha$-normality condition (1), so for non-zero $K^*_{i,j}$ we have:

$$|\hat{K}_{i,j} - K^*_{i,j}| \leq \frac{4\xi\epsilon}{|\hat{K}_{i,j} + K^*_{i,j}|} = \frac{4\xi\epsilon}{|\hat{K}_{i,j}| + |K^*_{i,j}|} \leq \frac{4\xi\epsilon}{\alpha}. \tag{2.10}$$

Combining Equation (2.10) and Equation (2.9), we obtain:

$$|\hat{K}_{i,j} - K^*_{i,j}| \leq 2\epsilon \min\left\{\frac{2\xi}{\alpha}, \sqrt{\frac{\xi}{\epsilon}}\right\}. \tag{2.11}$$

Note that by dropping the $\alpha$-normality condition, we still have the bound $|\hat{K}_{i,j} -$

$K^*_{i,j}| \leq 2\sqrt{\xi\epsilon}$. Hence, the upper bound in Equation (2.11) holds even by setting $\alpha = 0$, which is equivalent to having no $\alpha$-normality for $K^*$.

**Generating candidate matrices and DPPs for $\mathcal{M}$:** Our goal is to eventually bound the $\chi^2$-distance between $q$ and our estimated distribution. To achieve this goal (as we see shortly), it is enough that one estimates each entry of $K^*$ up to an additive error of

$$\wp := \frac{\epsilon\zeta}{100n^2}. \tag{2.12}$$

In some natural parameter regimes, i.e. when $\epsilon = \tilde{\Omega}(\zeta^{-2}N^{-\frac{1}{4}})$ or $\alpha = \tilde{\Omega}(\zeta^{-1}N^{-\frac{1}{4}})$, $\wp$ is larger than the upper bound that we already have in Equation (2.11) and so we can return the distribution of $\hat{K}$ as our estimate for $q$. However, if this is not the case, we need more candidates to make sure at least one of them is close to $K^*_{i,j}$. Note that $K^*_{i,j}$ is already in the range $\left[\hat{K}_{i,j} - 2\epsilon \min\left\{2\xi/\alpha, \sqrt{\xi/\epsilon}\right\}, \hat{K}_{i,j} + 2\epsilon \min\left\{2\xi/\alpha, \sqrt{\xi/\epsilon}\right\}\right]$ with high probability. Therefore, we divide this range into $\varsigma := \lceil 2\epsilon \min\left\{2\xi/\alpha, \sqrt{\xi/\epsilon}\right\}/\wp\rceil = \lceil 200n^2\zeta^{-1}\min\{2\xi/\alpha, \sqrt{\xi/\epsilon}\}\rceil$ intervals of equal length. This way, it is guaranteed that the true $K^*_{i,j}$ is $\wp$-close to one of the midpoints of these intervals (except when $K^*_{i,j}$ is zero which we handle separately). As discussed, this partitioning (is called *bracketing* technique in the literature of learning theory) allows the algorithm to achieve the optimal sample complexity.

Now, we claim that there are $2\varsigma + 1$ candidates for $K^*_{i,j}$. This number comes from the fact that we do not know whether $\hat{K}_{i,j}$ is equal to $\hat{K}^{(+)}_{i,j}$ or $\hat{K}^{(-)}_{i,j}$ a priori. Thus, each option provides $\varsigma$ candidates. Also, we have to consider the case $K^*_{i,j} = 0$ separately because the lower bound $\alpha$ only holds for non-zero entries $K^*_{i,j}$. By considering all the combinations of candidates for each entry, we obtain a set $M$ of matrices. Since each entry has a $\wp$-close candidate, there exists a matrix $\tilde{K} \in M$ such that all of its entries are $\wp$-close to the true kernel matrix $K^*$. Therefore, this matrix is $(n\wp)$-close to $K^*$ in the Frobenius distance. As we discussed in section 2.5, we project each $K \in M$ onto the set of valid marginal kernels and consider the set of candidate distributions $\mathcal{M} := \{\mathbf{Pr}_{\Pi(K)}[.] | K \in M\}$. The projection, $\Pi(K)$, is with respect to the Frobenius distance between matrices, and it is easy to see that computing it is equivalent to rounding up the eigenvalues of $K$ that are negative to zero, and rounding down the ones that are greater than one to one. Now for the DPP distribution $\tilde{p} = \mathbf{Pr}_{\Pi(\tilde{K})}[.] \in \mathcal{M}$, we prove the following claims:

(C1) The kernels $\Pi(\tilde{K})$ and $K^*$ are close in operator norm:

$$\|\Pi(\tilde{K}) - K^*\|_2 \leq \frac{\epsilon\zeta}{100n}.$$

(C2) The singular values of $\Pi(\tilde{K})$ are in the range $[99\zeta/100, 1 - 99\zeta/100]$.

For the first claim (C1), it is enough to write

$$\|\Pi(\tilde{K}) - K^*\|_2 \leq \|\Pi(\tilde{K}) - K^*\|_F = \|\Pi(\tilde{K}) - \Pi(K^*)\|_F \leq \|\tilde{K} - K^*\|_F \leq n\wp = \frac{\epsilon\zeta}{100n}. \tag{2.13}$$

where $\|.\|_2$ and $\|.\|_F$ refer to matrix operator norm and Frobenius norm respectively. The first inequality holds because the spectral norm is bounded by the Frobenius norm, the first equality follows from the fact that $K^*$ is a valid marginal kernel, and the second inequality is because of the contraction property of projection.

Next, we prove the second claim (C2). Using the variational characterization of the Operator norm and noting the fact that $\Pi(\tilde{K}) - K^*$ is symmetric (thus its singular values are the absolute values of its eigenvalues), we have

$$\|\Pi(\tilde{K}) - K^*\|_2 = \max_{v, \|v\|_2 = 1} |v^T(\Pi(\tilde{K}) - K^*)v|.$$

Combining this with Equation (2.13) then implies the following for every normalized vector $\|v\|_2 = 1$:

$$-\frac{\epsilon\zeta}{100n} \leq v^T(\Pi(\tilde{K}) - K^*)v \leq \frac{\epsilon\zeta}{100n}. \tag{2.14}$$

Since $\mathbf{Pr}_{K^*}[.]$ is $\zeta$-normal due to our assumption, we also have

$$\zeta \leq v^T K^* v \leq 1 - \zeta. \tag{2.15}$$

Combining Inequalities (2.14) and (2.15) yields

$$v^T \Pi(\tilde{K}) v \geq \zeta - \frac{\epsilon\zeta}{100n} \geq \zeta - \frac{\zeta}{100} = \frac{99\zeta}{100},$$

and similarly

$$v^T \Pi(\tilde{K}) v \leq 1 - \zeta + \frac{\epsilon\zeta}{100n} \leq 1 - \frac{99\zeta}{100},$$

for any arbitrary normalized vector $v$. Finally, using the variational characterization of the smallest and largest eigenvalues, we obtain that all eigenvalues of $\Pi(\tilde{K})$ are in the range $[99\zeta/100, 1 - 99\zeta/100]$. Note that the singular values of $\Pi(\tilde{K})$ are the absolute values of its eigenvalues, simply because $\Pi(\tilde{K})$ is symmetric, which completes the proof of the second claim (C2). We use these claims (C1), (C2) in the next part.

**Closeness in parameter space implies closeness of the distributions:** In this part of the proof, we show that closeness between $K^*$ and $\Pi(\tilde{K})$ in operator norm ensures the closeness of the distributions $q$ and $\tilde{p}$ with respect to the $\chi^2$-distance and $\ell_1$-distance. This result is based on the following Lemma, whose proof we defer to the end of this section.

**Lemma 5.** *For arbitrary symmetric matrices $B$ and $E$, we have*

$$\Big| |\det(B+E)| - |\det(B)| \Big| \leq |\det(B)| \frac{n\|E\|_2}{\sigma_n(B)} \left( \frac{\|E\|_2}{\sigma_n(B)} + 1 \right)^{n-1},$$

*where $\sigma_n(B)$ is the smallest singular value of $B$.*

Now consider an arbitrary set $J \subseteq [n]$ and its complement $\bar{J}$. Recall that Equation (2.1) gives:

$$\tilde{p}(J) = |\det(\Pi(\tilde{K}) - I_{\bar{J}})|\,,\; q(J) = |\det(K^* - I_{\bar{J}})|.$$

Therefore, setting $B \coloneqq \Pi(\tilde{K}) - I_{\bar{J}}$ and $E \coloneqq K^* - \Pi(\tilde{K})$ in Lemma 5, we can upper bound $|q(J) - \tilde{p}(J)|$ as

$$|q(J) - \tilde{p}(J)| \leq \tilde{p}(J) \frac{n\|E\|_2}{\sigma_n(B)} \left( \frac{\|E\|_2}{\sigma_n(B)} + 1 \right)^{n-1} . \tag{2.16}$$

Furthermore, from the second claim (C2) of the previous part, the singular values of $\Pi(\tilde{K})$ are in the range $[99\zeta/100, 1 - 99\zeta/100]$, which means the kernel matrix $\Pi(\tilde{K})$ satisfies the condition of Lemma 2. Therefore, from Lemma 2, the smallest singular value of $B$ is lower bounded as $\sigma_n(B) \geq \frac{99\zeta/100(1-99\zeta/100)}{\sqrt{2}} \geq \frac{99\zeta}{200\sqrt{2}}$, where we used $1 - 99\zeta/100 > 1/2$. Combining this with the first claim (C1) of the previous part implies

$$\frac{\|E\|_2}{\sigma_n(B)} \leq \frac{2\sqrt{2}\epsilon}{99n}.$$

Hence, Equation (2.16) gives:

$$|q(J) - \tilde{p}(J)| \leq \tilde{p}(J) \frac{2\sqrt{2}\epsilon}{99} \left( \frac{2\sqrt{2}\epsilon}{99n} + 1 \right)^{n-1} \leq \frac{\epsilon}{25} \tilde{p}(J), \tag{2.17}$$

where the last inequality follows from

$$\left( \frac{2\sqrt{2}\epsilon}{99n} + 1 \right)^{n-1} < \left( \frac{2\sqrt{2}}{99n} + 1 \right)^{n-1} < \frac{99}{50\sqrt{2}} \quad \forall n \in \mathbb{N}.$$

Note that $J \subseteq [n]$ is arbitrary, so Equation (2.17) finally yields the desired bound on the $\ell_1$-distance and $\chi^2$-distance between $q$ and $\tilde{p}$:

$$\ell_1(q, \tilde{p}) = \frac{1}{2} \sum_{J \subseteq [n]} |q(J) - \tilde{p}(J)| \leq \sum_{J \subseteq [n]} \frac{\epsilon}{50} \tilde{p}(J) = \frac{\epsilon}{50},$$
$$\chi^2(q, \tilde{p}) = \sum_{J \subseteq [n]} \frac{(q(J) - \tilde{p}(J))^2}{\tilde{p}(J)} < \sum_{J \subseteq [n]} \frac{\epsilon^2}{500} \tilde{p}(J) = \frac{\epsilon^2}{500}.$$

□

*Proof of Lemma 5.* Let $\sigma_1 \geq \cdots \geq \sigma_n$ be the singular values of $B$. For every $0 \leq k \leq n$, we denote $s_k$ the $k$th elementary symmetric function on the singular values of $B$, i.e.

$$s_0 = 1,\ \forall\, 1 \leq k \leq n:\ s_k = \sum_{1 \leq i_1 < \ldots < i_k \leq n} \sigma_{i_1} \ldots \sigma_{i_k},$$

Note that since $B$ is symmetric, the singular values are the absolute values of the eigenvalues, which implies the relation $|\det(B)| = \sigma_1 \cdots \sigma_n$.

Now Corollary 2.7 of [Ipsen and Rehman, 2008] states the following determinant's perturbation inequality:

$$\Big| \det(B+E) - \det(B) \Big| \leq \sum_{i=1}^{n} s_{n-i} \|E\|_2^i.$$

From this, we can derive

$$\begin{aligned} \Big| |\det(B+E)| - |\det(B)| \Big| \leq \Big| \det(B+E) - \det(B) \Big| &\leq \sum_{i=1}^{n} s_{n-i} \|E\|_2^i \\ &= |\det(B)| \sum_{i=1}^{n} \frac{s_{n-i}}{\sigma_1 \ldots \sigma_n} \|E\|_2^i, \end{aligned}$$

where in the last equality, we multiplied and divided the sum by $|\det(B)|$. Moving forward, we bound $s_{n-i}$ by $\binom{n}{i} \sigma_1 \cdots \sigma_{n-i}$:

$$\begin{aligned} \Big| |\det(B+E)| - |\det(B)| \Big| &\leq |\det(B)| \sum_{i=1}^{n} \binom{n}{i} \frac{\sigma_1 \ldots \sigma_{n-i}}{\sigma_1 \ldots \sigma_n} \|E\|_2^i \\ &= |\det(B)| \sum_{i=1}^{n} \binom{n}{i} \frac{1}{\sigma_{n-i+1} \ldots \sigma_n} \|E\|_2^i \\ &\leq |\det(B)| \sum_{i=1}^{n} \binom{n}{i} \left( \frac{\|E\|_2}{\sigma_n} \right)^i \\ &\leq |\det(B)|\, n \sum_{i=1}^{n} \binom{n-1}{i-1} \left( \frac{\|E\|_2}{\sigma_n} \right)^i \\ &= |\det(B)| \frac{n\, \|E\|_2}{\sigma_n} \sum_{i=0}^{n-1} \binom{n-1}{i} \left( \frac{\|E\|_2}{\sigma_n} \right)^i \\ &= |\det(B)| \frac{n \|E\|_2}{\sigma_n} \left( \frac{\|E\|_2}{\sigma_n} + 1 \right)^{n-1}. \end{aligned}$$

□

## 2.9 Uniform Lower Bound on the Smallest Singular Value of $K - I_{\bar{J}}$

In this section, we prove Lemma 2: given a marginal kernel $K$ whose eigenvalues are in the range $[\zeta, 1-\zeta]$, we prove the uniform lower bound $\zeta(1-\zeta)/\sqrt{2}$ on the singular values of the family of matrices $\{K - I_{\bar{J}}\}_{J \subseteq [n]}$. This Lemma is used in the proof of Theorem 3 and enables us to control the distances between the atom probabilities of $\mathbf{Pr}_K[.]$ and $\mathbf{Pr}_{\Pi(\tilde{K})}[.]$.

*Proof of Lemma 2.* Let $\lambda_1 \geq ... \geq \lambda_n$ be the eigenvalues of $K$ and $v_1, ..., v_n$ be an orthonormal set of their corresponding eigenvectors. We fix a subset $J \subseteq [n]$ and lower bound the smallest singular value of $K - I_{\bar{J}}$ based on its variational characterization:

$$\sigma_n(K - I_{\bar{J}}) = \min_{\|v\|_2=1} \sqrt{v^T(K - I_{\bar{J}})^2 v}. \tag{2.18}$$

Given a normalized vector $v$: $\|v\|_2 = 1$, we represent $v$ in the basis $\{v_i\}_{i=1}^n$ as $v = \sum_{i=1}^n \alpha_i v_i$. Because $\{v_i\}_{i=1}^n$ is orthonormal, we have

$$1 = \|v\|^2 = \sum_{i=1}^n \alpha_i^2 \|v_i\|^2 = \sum_{i=1}^n \alpha_i^2.$$

Now we can express $v^T(K - I_{\bar{J}})^2 v$ as:

$$\begin{aligned} v^T(K - I_{\bar{J}})^2 v &= \left(\sum_{i=1}^n \alpha_i v_i\right)^T (K - I_{\bar{J}})^2 \left(\sum_{i=1}^n \alpha_i v_i\right) \\ &= \sum_{1\leq i,j\leq n} \alpha_i \alpha_j v_i^T (K - I_{\bar{J}})^2 v_j \\ &= \sum_{1\leq i,j\leq n} \alpha_i \alpha_j v_i^T K^2 v_j + \sum_{1\leq i,j\leq n} \alpha_i \alpha_j \big(v_i^T I_{\bar{J}}^2 v_j - v_i^T K I_{\bar{J}} v_j - v_i^T I_{\bar{J}} K v_j\big). \end{aligned}$$

Observe that ${v_i}^T K^2 v_i = \lambda_i^2 \|v_i\|^2 = \lambda_i^2$ and ${v_i}^T K^2 v_j = \lambda_i \lambda_j {v_i}^T v_j = 0$ for $i \neq j$.

We define some additional notation here: For any subset $J \subseteq [n]$, let $(v_i)_J$ be the restriction of $v_i$ into support $J$. We also denote the inner product of the vectors $v_i$ and $v_j$ restricted to $J$ by $\langle v_i, v_j \rangle_J$. Using these notations, we can further simplify the terms ${v_i}^T I_{\bar{J}}^2 v_j$, ${v_i}^T K I_{\bar{J}} v_j$ and ${v_i}^T I_{\bar{J}} K v_j$ to $\langle v_i, v_j \rangle_{\bar{J}}$, $\lambda_i \langle v_i, v_j \rangle_{\bar{J}}$, and $\lambda_j \langle v_i, v_j \rangle_{\bar{J}}$ respectively. Substituting them above results in

$$\begin{aligned} v^T (K - I_{\bar{J}})^2 v =& \sum_{i=1}^{n} \alpha_i^2 \lambda_i^2 + \sum_{1 \le i,j \le n} (1 - \lambda_i - \lambda_j) \alpha_i \alpha_j \langle v_i, v_j \rangle_{\bar{J}} \\ =& \sum_{i=1}^{n} \alpha_i^2 \lambda_i^2 - \sum_{1 \le i,j \le n} \alpha_i \alpha_j \lambda_i \lambda_j \langle v_i, v_j \rangle_{\bar{J}} + \sum_{1 \le i,j \le n} \alpha_i \alpha_j (1 - \lambda_i)(1 - \lambda_j) \langle v_i, v_j \rangle_{\bar{J}} \end{aligned}$$

where the last equality simply follows from the Equation $(1 - \lambda_i)(1 - \lambda_j) = 1 - \lambda_i - \lambda_j + \lambda_i \lambda_j$. Now substituting $\langle v_i, v_j \rangle_{\bar{J}}$ by $\langle v_i, v_j \rangle - \langle v_i, v_j \rangle_J$ in the second term above, we obtain

$$\begin{aligned} & v^T (K - I_{\bar{J}})^2 v \\ =& \sum_{i=1}^{n} \alpha_i^2 \lambda_i^2 - \sum_{1 \le i,j \le n} \alpha_i \alpha_j \lambda_i \lambda_j \langle v_i, v_j \rangle \\ +& \sum_{1 \le i,j \le n} \alpha_i \alpha_j \lambda_i \lambda_j \langle v_i, v_j \rangle_J + \sum_{1 \le i,j \le n} \alpha_i \alpha_j (1 - \lambda_i)(1 - \lambda_j) \langle v_i, v_j \rangle_{\bar{J}} \\ =& \sum_{i=1}^{n} \alpha_i^2 \lambda_i^2 - \sum_{i=1}^{n} \alpha_i^2 \lambda_i^2 + \sum_{1 \le i,j \le n} \alpha_i \alpha_j \lambda_i \lambda_j \langle v_i, v_j \rangle_J + \sum_{1 \le i,j \le n} \alpha_i \alpha_j (1 - \lambda_i)(1 - \lambda_j) \langle v_i, v_j \rangle_{\bar{J}} \\ =& \Big\| \sum_{i=1}^{n} \alpha_i \lambda_i (v_i)_J \Big\|^2 + \Big\| \sum_{i=1}^{n} \alpha_i (1 - \lambda_i)(v_i)_{\bar{J}} \Big\|^2. \end{aligned} \tag{2.19}$$

Hence, it suffices to derive a lower bound on $\Big\| \sum_{i=1}^{n} \alpha_i \lambda_i (v_i)_J \Big\|^2 + \Big\| \sum_{i=1}^{n} \alpha_i (1 - \lambda_i)(v_i)_{\bar{J}} \Big\|^2$ independent from $J$. To this end, we define the column vectors $w_1 = \Big( \alpha_i \lambda_i \Big)_{i=1}^{n}$, $w_2 = \Big( \alpha_i (1 - \lambda_i) \Big)_{i=1}^{n}$. Furthermore, define $R \coloneqq \Big( v_1 \Big| \dots \Big| v_n \Big)$ as the matrix with $v_i$ as its $i$th column, and let ${v_1'}^T, ..., {v_n'}^T$ be the rows of $R$. Because $\{v_i\}_{i=1}^{n}$ is an orthonormal set, $R$ is a unitary matrix, so $\{v_j'\}_{j=1}^{n}$ is also an orthonormal set. Next, let

$V$ and $V^T$ be the subspaces spanned by the set of vectors $\{v'_j\}_{j\in\bar{J}}$ and $\{v'_j\}_{j\in J}$ respectively. Because $\{v'_j\}_{j=1}^n$ is an orthonormal set, the subspaces $V$ and $V^\perp$ are orthogonal to each other. Let $\nu_1 = \sum_{j\in\bar{J}}({v'_j}^T w_1)v'_j$ and ${\nu_1}^\perp = \sum_{j\in J}({v'_j}^T w_1)v'_j$ be the projections of $w_1$ onto $V$ and $V^\perp$ respectively. Similarly, define $\nu_2 = \sum_{j\in\bar{J}}({v'_j}^T w_2)v'_j$ and ${\nu_2}^\perp = \sum_{j\in J}({v'_j}^T w_2)v'_j$ as the projections of $w_2$ onto $V$ and $V^\perp$. Now by decomposing $w_1$ on $V$ and $V^\perp$, we can write

$$w_1 = \nu_1 + {\nu_1}^\perp.$$

Similarly, we have

$$w_2 = \nu_2 + {\nu_2}^\perp.$$

Moreover, from the orthonormality of $v'_1, ..., v'_n$, we obtain

$$\begin{aligned}\|{\nu_1}^\perp\|^2 =& \Big\| \sum_{j\in J}({v'_j}^T w_1)v'_j\Big\|^2 = \sum_{j\in J}({v'_j}^T w_1)^2 = \sum_{j\in J}(\sum_{i=1}^n R_{j,i}(w_1)_i)^2 \\ =& \sum_{j\in J}(\sum_{i=1}^n (v_i)_j(w_1)_i)^2 = \Big\|\sum_{i=1}^n \alpha_i\lambda_i(v_i)_J\Big\|^2.\end{aligned}$$

Similarly, one obtains

$$\|\nu_2\|^2 = \Big\|\sum_{j\notin J}({v'_j}^T w_2)v'_j\Big\|^2 = \Big\|\sum_{i=1}^n \alpha_i(1-\lambda_i)(v_i)_{\bar{J}}\Big\|^2.$$

Combining the last two equations with Equation (2.19), we obtain

$$v^T(K - I_{\bar{J}})^2 v = \|\nu_2\|^2 + \|{\nu_1}^\perp\|^2. \tag{2.20}$$

Now, it suffices to bound $\|\nu_2\|^2 + \|{\nu_1}^\perp\|^2$. Note that

$$\begin{aligned}\|w_1\|^2 &= \sum_{i=1}^n \alpha_i^2\lambda_i^2 \le \sum \alpha_i^2 = 1, \\ \|w_2\|^2 &= \sum_{i=1}^n \alpha_i^2(1-\lambda_i)^2 \le \sum \alpha_i^2 = 1.\end{aligned}$$

which implies $\|\nu_1\|, \|\nu_2\|, \|{\nu_1}^\perp\|, \|{\nu_2}^\perp\| \leq 1$. Moreover, the condition $\zeta \leq \lambda_i \leq 1-\zeta$ implies $\lambda_i(1-\lambda_i) \geq \zeta(1-\zeta)$. Therefore, on one hand, we get

$$\langle w_1, w_2 \rangle = \sum_{i=1}^{n} \lambda_i(1-\lambda_i)\alpha_i^2 \geq \zeta(1-\zeta)\sum_{i=1}^{n} \alpha_i^2 = \zeta(1-\zeta). \tag{2.21}$$

On the other hand,

$$\begin{aligned} \langle w_1, w_2 \rangle &= \langle \nu_1 + {\nu_1}^\perp, \nu_2 + {\nu_2}^\perp \rangle = \langle \nu_1, \nu_2 \rangle + \langle {\nu_1}^\perp, {\nu_2}^\perp \rangle \\ &\leq \|\nu_1\|\|\nu_2\| + \|{\nu_1}^\perp\|\|{\nu_2}^\perp\| \leq \|\nu_2\| + \|{\nu_1}^\perp\| \\ &\leq \sqrt{2(\|\nu_2\|^2 + \|{\nu_1}^\perp\|^2)} = \sqrt{2v^T(K - I_{\bar{J}})^2 v}. \end{aligned} \tag{2.22}$$

where the last equality follows from Equation (2.20). Combining Equations (2.21) and (2.22), we conclude $v^T(K - I_{\bar{J}})^2 v \geq \zeta^2(1-\zeta)^2/2$. Recall that $v$ is an arbitrary normalized vector, and $J$ is an arbitrary subset of $[n]$, so the variational characterization of $\sigma_n$ in Equation (2.18) yields the desired lower bound $\sigma_n(K - I_{\bar{J}}) \geq \zeta(1-\zeta)/\sqrt{2}$ for every $J \subseteq [n]$. □

## 2.10 Lower Bound for Testing Log-submodular Distributions

In this section, we rigorously prove Lemma 4, which in turn completes the proof of Theorem 4. We assume that $\epsilon'$, $\mathcal{F}$, $h_r$ and $\bar{h}_r$ are defined as in Section 2.6.

*Detailed Proof of Lemma 4.* Given $\epsilon' \leq \frac{2}{3}$ and a log-submodular distribution $f$, we first show that the $\ell_1$-distance between $f$ and the unnormalized measure $\bar{h}_r$ is large with high probability independent of $f$ (we define the $\ell_1$-distance of general measures the same as for probability measures.) To this end, we define the following family of subsets based on $h_r$, that is random:

$$\mathcal{S}_r := \{S \subseteq [n] \setminus \{1,2\} \mid r_{(S\cup\{1,2\})} = 1,\ r_{(S\cup\{2\})} = -1,\ r_{(S\cup\{1\})} = -1\}. \tag{2.23}$$

We prove that $\mathcal{S}_r$ has the following properties:

(P1) With high probability, the cardinality of $\mathcal{S}_r$ is at least $N/64$.

(P2) For every $S \in \mathcal{S}_r$, there is a contribution of at least $\epsilon'/8N$ to the $\ell_1$-distance between $\bar{h}_r$ and $f$ from the term $V_S$ defined as

$$\begin{aligned}V_S :=& \frac{1}{2}|\bar{h}_r(S) - f(S)| + \frac{1}{2}|\bar{h}_r(S \cup \{1\}) - f(S \cup \{1\})| + \\ & \frac{1}{2}|\bar{h}_r(S \cup \{2\}) - f(S \cup \{2\})| + \frac{1}{2}|\bar{h}_r(S \cup \{1,2\}) - f(S \cup \{1,2\})|.\end{aligned}$$

Note that based on these two properties, one can simply derive

$$\ell_1(\bar{h}_r, f) \geq \frac{N}{64} \times \frac{\epsilon'}{8N} = \frac{\epsilon'}{512} \tag{2.24}$$

with high probability.

To show that the event $\mathcal{Q}_1 := \{|\mathcal{S}_r| \geq N/64\}$ happens with high probability for the first property (P1), we use a Chernoff bound for the random variables $\mathbb{1}\{S \in \mathcal{S}_r\}$, $\forall S \subseteq [n] \setminus \{1,2\}$, where $\mathbb{1}\{.\}$ is the indicator function. Clearly, for each $S \subseteq [n] \setminus \{1,2\}$, we have $\mathbb{E}[\mathbb{1}\{S \in \mathcal{S}_r\}] = \mathbf{Pr}[S \in \mathcal{S}_r] = 1/8$, and $\mathbb{E}[|\mathcal{S}_r|] = N/32$. Therefore,

$$\mathbf{Pr}[\mathcal{Q}_1^c] = \mathbf{Pr}\left[\sum_{S \in [n] \setminus \{1,2\}} \mathbb{1}\{S \in \mathcal{S}_r\} < \left(1 - \frac{1}{2}\right)\mathbb{E}[|\mathcal{S}_r|]\right] \leq \exp\left(-0.5\frac{N}{32}(\frac{1}{2})^2\right) = \exp\left(-\frac{N}{256}\right).$$

We conclude for $n \geq n_1 = 11$, $\mathcal{Q}_1$ happens with probability at least 0.995.

We now prove the second property (P2). Fix a set $S \in \mathcal{S}_r$ and define the constant $\rho := \frac{1+\epsilon'}{1-3\epsilon'/4}$. To prove $V_S \geq \frac{\epsilon'}{8N}$, we consider two cases:

**Case 1:** $\frac{f(S \cup \{1,2\})}{f(S \cup \{2\})} \leq \rho$

Here, we formalize a helper inequality in the following Lemma, and prove it at the end of this section.

**Lemma 6.** *For $a, b \geq 0$, the condition $\frac{a}{b} \leq \rho$ implies $|1 + \epsilon' - a| + |1 - \epsilon' - b| \geq \frac{\epsilon'}{4}$.*

Now from $S \in \mathcal{S}_r$, we get $\bar{h}_r(S \cup \{1,2\}) = \frac{1+\epsilon'}{N}$ and $\bar{h}_r(S \cup \{2\}) = \frac{1-\epsilon'}{N}$. Hence,

$$\begin{aligned} V_S &\geq \frac{1}{2}|\bar{h}_r(S \cup \{1,2\}) - f(S \cup \{1,2\})| + \frac{1}{2}|\bar{h}_r(S \cup \{2\}) - f(S \cup \{2\})| \\ &= \frac{1}{2}\Big|\frac{1+\epsilon'}{N} - f(S \cup \{1,2\})\Big| + \frac{1}{2}\Big|\frac{1-\epsilon'}{N} - f(S \cup \{2\})\Big| \geq \frac{\epsilon'}{8N}, \end{aligned}$$

where the last inequality follows from Lemma 6, by setting $a = Nf(S \cup \{1,2\})$, $b = Nf(S \cup \{2\})$.

**Case 2:** $\frac{f(S\cup\{1,2\})}{f(S\cup\{2\})} > \rho$

In this case, the log-submodularity property allows us to write

$$\log(f(S \cup \{1\})) - \log(f(S)) \geq \log(f(S \cup \{1,2\})) - \log(f(S \cup \{2\})) > \log(\rho),$$

or equivalently

$$\frac{f(S \cup \{1\})}{f(S)} > \rho = \frac{1+\epsilon'}{1-3\epsilon'/4}. \tag{2.25}$$

Note that from $S \in \mathcal{S}_r$, we have $\bar{h}_r(S \cup \{1\}) = \frac{1-\epsilon'}{N}$. If $f(S \cup \{1\})$ is larger than $\frac{1-3\epsilon'/4}{N}$, then

$$V_S \geq \frac{1}{2}|\bar{h}_r(S \cup \{1\}) - f(S \cup \{1\})| > \frac{1}{2}\Big(\frac{1-3\epsilon'/4}{N} - \frac{1-\epsilon'}{N}\Big) = \frac{\epsilon'}{8N}$$

and we are done. Otherwise, we have $f(S \cup \{1\}) \leq \frac{1-3\epsilon'/4}{N}$. Combining this with Equation (2.25) gives:

$$f(S) \leq \rho^{-1} f(S \cup \{1\}) \leq \frac{1-3\epsilon'/4}{1+\epsilon'} \times \frac{1-3\epsilon'/4}{N} \leq \frac{1-\epsilon'}{N} - \frac{\epsilon'}{4N},$$

where the last inequality follows from the condition $\epsilon' \leq \frac{2}{3}$. Finally, we obtain

$$V_S \geq \frac{1}{2}|\bar{h}_r(S) - f(S)| \geq \frac{1}{2}\Big(\frac{1-\epsilon'}{N} - (\frac{1-\epsilon'}{N} - \frac{\epsilon'}{4N})\Big) = \frac{\epsilon'}{8N},$$

which completes the proof for the second property (P2). Therefore, under the occurrence of $\mathcal{Q}_1$, we conclude from Equation (2.24) that $\ell_1(\bar{h}_r, f) \geq \frac{\epsilon'}{512}$. To show the $\ell_1$-distance between $h_r$ and $f$ is also large, we control the normalization constant

$L_r := \sum_{S\subseteq[n]} \bar{h}_r(S)$. Define the event $\mathcal{Q}_2 := \{1 - \frac{4\epsilon'}{\sqrt{N}} \leq L_r \leq 1 + \frac{4\epsilon'}{\sqrt{N}}\}$ . A simple Hoeffding bound for the random variables $\frac{1+r_S\epsilon'}{N}$, $\forall S \subseteq [n]$, implies that $\mathcal{Q}_2$ happens with probability at least 0.995. Now under the occurrence of $\mathcal{Q}_1 \cap \mathcal{Q}_2$ and assuming $n \geq n_2 = 22$, we can write:

$$\begin{aligned}
2\ell_1(h_r, f) &= \sum_{S\subseteq[n]} |h_r(S) - f(S)| = \sum_{S\subseteq[n]} \Big|\frac{\bar{h}_r(S)}{L_r} - f(S)\Big| \\
&\geq \sum_{S\subseteq[n]} |\bar{h}_r(S) - f(S)| - \sum_{S\subseteq[n]} \bar{h}_r(S)\Big|\frac{1-L_r}{L_r}\Big| \\
&\geq \frac{\epsilon'}{256} - \frac{4\epsilon'}{L_r\sqrt{N}} \sum_{S\subseteq[n]} \bar{h}_r(S) \geq \epsilon'(\frac{1}{256} - \frac{4}{\sqrt{N}}) \geq \epsilon'(\frac{1}{256} - \frac{1}{512}) = \frac{c\epsilon}{512}.
\end{aligned}$$

A union bound for the events $Q_1^c$ and $Q_2^c$ implies that $\mathcal{Q}_1 \cap \mathcal{Q}_2$ happens with probability at least 0.99. Note that $\mathcal{Q}_1$ and $\mathcal{Q}_2$ does not depend on $f$. Setting $c = 1024$, we conclude that with probability at least 0.99, $\ell_1(h_r, f) \geq \epsilon$ for any log-submodular distribution $f$, given that $\epsilon = \epsilon'/c \leq \frac{2}{3\times 1024}$ and $n \geq \max\{n_1, n_2\} = 22$, which completes the proof of Lemma 4. □

*Proof of Lemma 6.* Here, we prove Lemma 6, which we used above. First note that if $b \geq a$, then clearly $|b - (1-\epsilon')| + |a - (1+\epsilon')| \geq 2\epsilon' > \frac{\epsilon'}{4}$. So we assume $b < a$.

Now define $t := a - (1+\epsilon')$, so that $a = 1 + \epsilon' + t$. Then, we can write

$$\begin{aligned}
|b - (1-\epsilon')| + |a - (1+\epsilon')| &= |\frac{b}{a}(1+\epsilon'+t) - (1-\epsilon')| + |t| \\
&\geq |\frac{b}{a}(1+\epsilon') - (1-\epsilon')| - |\frac{b}{a}t| + |t| \\
&= |\frac{b}{a}(1+\epsilon') - (1-\epsilon')| + (1-\frac{b}{a})|t|.
\end{aligned}$$

The condition $\frac{a}{b} \leq \rho$ implies $\frac{b}{a}(1+\epsilon') \geq 1 - \frac{3\epsilon'}{4}$. Therefore

$$|b - (1-\epsilon')| + |a - (1+\epsilon')| \geq \frac{\epsilon'}{4} + (1-\frac{b}{a})|t| \geq \frac{\epsilon'}{4}.$$

where the last inequality follows from the fact that $1 - \frac{b}{a} > 0$. □

## 2.11 Coupling DPPs

In this section, we fully introduce and prove the coupling argument of Lemma 3. Given a value $0 < z \leq 0.5$ and a DPP whose marginal kernel has eigenvalues that are outside the range $[z, 1-z]$, the goal is to couple it with another DPP, which has a marginal kernel with all eigenvalues in $[z, 1-z]$, such that the data sets generated from these two DPPs are equal with high probability.

*Proof of Lemma 3.* Let $V$ be an orthonormal set of the eigenvectors of $K$. For each $v \in V$, let $\lambda_v$ be its corresponding eigenvalue. To introduce our coupling, we need to define the class of *elementary DPPs* [Kulesza and Taskar, 2012]. A DPP is called *elementary* if the eigenvalues of its marginal kernel are either zero or one. For each subset $V' \subseteq V$ of the eigenvectors of $K$, we consider the elementary DPP $\mathbf{Pr}_{K^{V'}}[.]$ with marginal kernel $K^{V'} := \sum_{v \in V'} vv^T$. It is well-known that any DPP can be viewed as a mixture of its corresponding elementary DPPs [Kulesza and Taskar, 2012], i.e.

$$\mathbf{Pr}_K[.] = \sum_{V' \subseteq V} \Big( \Pi_{v \in V'} \lambda_v \Pi_{v \notin V'} (1 - \lambda_v) \Big) \mathbf{Pr}_{K^{V'}}[.]. \tag{2.26}$$

Using this mixture formulation, we can sample a set from $\mathbf{Pr}_K[.]$ as follows: For each eigenvector $v \in V$, we sub-sample $v$ with probability $\lambda_v$ to obtain the random subset $V'$ of $V$, then we sample $\mathcal{J}_K$ from the elementary DPP with marginal kernel $K^{V'}$. We call this sampling scheme "elementary sampling:"

- (1) For each $v \in V$, sample $y_v \sim \text{Bernoulli}(\lambda_v)$, add $v \in V'$ if $y_v = 1$.

- (2) sample $\mathcal{J}_K \sim \mathbf{Pr}_{K^{V'}}[.]$

According to the mixture formulation in Equation (2.26), the elementary sampling scheme samples $\mathcal{J}_K$ according to $\mathbf{Pr}_K[.]$.

One can readily see that the projected matrix $\Pi_z(K)$ has the same eigenvectors as

$K$ but with corresponding eigenvalues $\{\bar{\lambda}_v\}_{v \in V}$, where

$$\bar{\lambda}_v = \begin{cases} \lambda_v & \text{if } \lambda_v \in [z, 1-z] \\ z & \text{if } \lambda_v < z \\ 1-z & \text{if } \lambda_v > 1-z \end{cases} \tag{2.27}$$

This fact follows from applying the 2-*Weilandt-Hoffman* inequality [Tao, 2012] for the projection operator $\Pi_z(.)$. We can similarly sample $\mathcal{J}_{\Pi_z(K)} \sim \mathbf{Pr}_{\Pi_z(K)}[.]$ with the above elementary sampling scheme. Next, we define a coupling between $\mathcal{J}_K$ and $\mathcal{J}_{\Pi_z(K)}$ as follows:

- (1) For each $v \in V$, sample $x_v \sim \text{Uniform}[0, 1]$. Then add $v$ to $V_1'$ if $x_v \in [0, \lambda_v]$, and add $v$ to $V_2'$ if $x_v \in [0, \bar{\lambda}_v]$.

- (2) if $V_1' = V_2'$, then sample $\mathcal{J} \sim \mathbf{Pr}_{K^{V_1'}}[.]$ and set $\mathcal{J}_K = \mathcal{J}_{\Pi_z(K)} = \mathcal{J}$. Otherwise, independently sample $\mathcal{J}_K \sim \mathbf{Pr}_{K^{V_1'}}[.]$, $\mathcal{J}_{\Pi_z(K)} \sim \mathbf{Pr}_{K^{V_2'}}[.]$.

By looking at the marginal distributions of the sets $\mathcal{J}_K$ and $\mathcal{J}_{\Pi_z(K)}$ sampled above, we observe that $\mathcal{J}_K \sim \mathbf{Pr}_K[.]$, $\mathcal{J}_{\Pi_z(K)} \sim \mathbf{Pr}_{\Pi_z(K)}[.]$, i.e. the marginals of the coupling are as one would expect. Furthermore, if the sampled sets $V_1'$ and $V_2'$ in the first step of the sampling are equal, then $\mathcal{J}_K = \mathcal{J}_{\Pi_z(K)}$. Therefore, to lower bound $\mathbf{Pr}_{\text{coupling}}\big[\mathcal{J}_K = \mathcal{J}_{\Pi_z(K)}\big]$, it is enough to upper bound $\mathbf{Pr}_{\text{coupling}}[\mathcal{W}]$ for the event $\mathcal{W} := \{V_1' \neq V_2'\}$. But we can expand $\mathcal{W}$ as

$$\mathcal{W} = \bigcup_{v \in V} \Big( \{v \in V_1', v \notin V_2'\} \cup \{v \in V_2', v \notin V_1'\} \Big).$$

Note that for each $v \in V$, $\{v \in V_1', v \notin V_2'\} \cup \{v \in V_2', v \notin V_1'\}$ happens with probability $|\lambda_v - \bar{\lambda}_v|$. From Equation (2.27), we observe that $|\lambda_v - \bar{\lambda}_v| \leq z$ for every $v \in V$. Therefore, using a union bound, we obtain

$$\mathbf{Pr}_{\text{coupling}}[\mathcal{W}] \leq nz.$$

Using the definition $z = \delta/2mn$, we conclude that

$$\mathbf{Pr}_{\text{coupling}}\big[\mathcal{J}_K = \mathcal{J}_{\Pi_z(K)}\big] \geq 1 - \mathbf{Pr}_{\text{coupling}}[\mathcal{W}] \geq 1 - nz = 1 - \frac{\delta}{2m}. \tag{2.28}$$

Using this coupling to generate the samples $\{\mathcal{J}_K^{(t)}\}_{t=1}^m$ and $\{\mathcal{J}_{\Pi_z(K)}^{(t)}\}_{t=1}^m$, we can write

$$\begin{aligned}\mathbf{Pr}_{\text{coupling}}\Big[\{\mathcal{J}_K^{(t)}\}_{t=1}^m = \{\mathcal{J}_{\Pi_z(K)}^{(t)}\}_{t=1}^m\Big] &= \Big(\mathbf{Pr}_{\text{coupling}}\big[\mathcal{J}_K = \mathcal{J}_{\Pi_z(K)}\big]\Big)^m \\ &\geq \Big(1 - \frac{\delta}{2m}\Big)^m\end{aligned}$$

For a real number $u$, we have the inequality

$$(1 - \frac{1}{u})^u \leq e^{-1},$$

and for $u \geq 2$, we have

$$(1 - \frac{1}{u})^u \geq e^{-\frac{u}{u-1}} \geq e^{-2}.$$

Applying these inequalities, we finally obtain

$$\mathbf{Pr}_{\text{coupling}}\Big[\{\mathcal{J}_K^{(t)}\}_{t=1}^m = \{\mathcal{J}_{\Pi_z(K)}^{(t)}\}_{t=1}^m\Big] \geq \bigg(\Big(1 - \frac{\delta}{2m}\Big)^{\frac{2m}{\delta}}\bigg)^{\frac{\delta}{2}} \geq e^{-\delta} \geq 1 - \delta.$$

□

## 2.12 A More Detailed Proof of Theorem 1

In this section, we take a more elaborate look at the proof of Theorem 1. The proof is mentioned in Section 2.5.2.

*Detailed proof of Theorem 1.* Lemma 1 tells us there exists a constant $c_1$ such that $c_1 C_{N,\epsilon,\alpha,\zeta}\sqrt{N}/\epsilon^2$ samples suffice for DPP-Tester to successfully test against $(\alpha, \zeta)$-normal DPPs, with probability at least 0.995. For the general problem of testing against any DPP (i.e. without having the normality conditions), we prove that

$m^* = c_2 C_{N,\epsilon}\sqrt{N}/\epsilon^2$ samples suffice to succeed with probability at least 0.99, as long as $c_2 \geq c_1 \max\{23, 2\log(c_1) + 23\}$. To test against all DPPs, we use the parameter setting of DPP-Tester for $(0, \bar{z})$-normal DPPs, where we define $\bar{z} := 0.005/2m^*n$. The key idea is that via the coupling argument of Lemma 3, we can reduce the analysis for testing against all DPPs to the analysis for testing against only $(0, \bar{z})$-normal DPPs. To this end, we use the following Lemma. The derivation of the inequality in Lemma 7 is based on elementary algebraic operations, and we differ its proof to the end of this section.

**Lemma 7.** *For constant $c_2$ picked as large as $c_2 \geq c_1 \max\{23, 2\log(c_1) + 23\}$, we have*

$$m^* \geq C_{N,\epsilon,0,\bar{z}}\sqrt{N}/\epsilon^2. \tag{2.29}$$

Therefore, we pick $c_2 \geq c_1 \max\{23, 2\log(c_1) + 23\}$ to satisfy the inequality $m^* \geq C_{N,\epsilon,0,\bar{z}}\sqrt{N}/\epsilon^2$. This means that given $m^*$ samples, according to the definition of $c_1$, our tester can test against $(0, \bar{z})$-normal DPPs with success probability at least 0.995. Therefore, if the underlying distribution $q$ is an $(0, \bar{z})$-normal DPP, or if it is $\epsilon$-far from all DPPs, then DPP-Tester outputs correctly with probability at least 0.995. It remains to show that the algorithm can also handle a DPP with kernel $K^*$, which is not $(0, \bar{z})$-normal. To see this, note that because of the particular choice of $\bar{z}$, our coupling argument in Lemma 3 implies that the product distributions $\mathbf{Pr}^{(m^*)}_{K^*}[.]$ and $\mathbf{Pr}^{(m^*)}_{\Pi_{\bar{z}}(K^*)}[.]$ over the space of data sets have $\ell_1$-distance at most 0.005. This follows from the fact that for two arbitrary random variables $X$ and $Y$ over the same underlying space, with probability distributions $P_X$ and $P_Y$, we have the following characterization of their $\ell_1$-distance:

$$\ell_1(P_X, P_Y) = \inf_{\text{coupling}(X,Y)} \mathbf{Pr}_{\text{coupling}}[X \neq Y].$$

Therefore, we have $\ell_1\Big(\mathbf{Pr}^{(m^*)}_{K^*}[.], \mathbf{Pr}^{(m^*)}_{\Pi_{\bar{z}}(K^*)}[.]\Big) \leq 0.005$. From this, we can relate the probability of the tester's acceptance region under $\mathbf{Pr}^{(m^*)}_{K^*}[.]$, to the same probability under $\mathbf{Pr}^{(m^*)}_{\Pi_{\bar{z}}(K^*)}[.]$:

$$\mathbf{Pr}^{(m^*)}_{K^*}\left[\text{Acceptance Region}\right] \geq \mathbf{Pr}^{(m^*)}_{\Pi_{\bar{z}}(K^*)}\left[\text{Acceptance Region}\right] - 0.005 \geq 0.995 - 0.005 = 0.99,$$

where the last inequality follows from the fact that $\mathbf{Pr}_{\Pi_{\bar{z}}(K^*)}[.]$ is an $(0, \bar{z})$-normal DPP, according to the definition of $\Pi_{\bar{z}}(K^*)$. Hence, for $c_2 \geq \max\{23, 2\log(c_1) + 23\}$, DPP-Tester, with the particular choice of its parameter $\varsigma$ with respect to $(0, \bar{z})$-normal DPPs, succeeds given $c_2 C_{N,\epsilon}\sqrt{N}/\epsilon^2$ samples to test all DPPs with probability at least 0.99. This completes the proof of Theorem 1. □

*Proof of Lemma 7.* As usual, log(.) denotes the natural logarithm. Inequality (2.29) boils down to

$$c_2 C_{N,\epsilon} \geq c_1 C_{N,\epsilon,0,\bar{z}},$$

or equivalently

$$c_2 \log^2(N)(\log(N) + \log(1/\epsilon)) \geq c_1 \log^2(N)(1 + \log(1/\bar{z}) + \log(1/\epsilon))$$

$$\Leftrightarrow c_2(\log(N) + \log(1/\epsilon)) \geq c_1(1 + \log(1/0.0025) + \log(m^*) + \log(n) + \log(1/\epsilon)). \tag{2.30}$$

Using the inequality $\log(x) \leq x - 1$ for $x > 0$, we get:

$$\begin{aligned}\log(m^*) &= \log(c_2 C_{N,\epsilon}\sqrt{N}/\epsilon^2) \\ &= \log(c_2) + 2\log(\log(N)) + \log(\log(N) + \log(1/\epsilon)) + \frac{1}{2}\log(N) + 2\log(1/\epsilon) \\ &\leq \log(c_2) + 2(\log(N) - 1) + \log(N) + \log(1/\epsilon) - 1 + \frac{1}{2}\log(N) + 2\log(1/\epsilon) \\ &= \log(c_2) - 2 + \frac{7}{2}\log(N) + 3\log(1/\epsilon). \end{aligned} \tag{2.31}$$

Substituting Inequality (2.31) in Inequality (2.30), it is enough to satisfy

$$\frac{c_2}{c_1} \geq \frac{\log(c_2) - 1 + \log(1/0.0025) + 7/2\log(N) + 4\log(1/\epsilon) + \log(n)}{\log(N) + \log(1/\epsilon)} := \varrho.$$

We further upper bound $\varrho$ using the inequalities $\log(n) < \frac{1}{2}\log(N) + 1$ and $\log(N) \geq 0.69$:

$$\begin{aligned}
\varrho &< \frac{\log(c_2) + 6 + 8\log(N) + 4\log(1/\epsilon)}{\log(N) + \log(1/\epsilon)} \\
&= \frac{\log(c_2) + 6}{\log(N) + \log(1/\epsilon)} + \frac{8\log(N) + 4\log(1/\epsilon)}{\log(N) + \log(1/\epsilon)} \\
&\leq 1.5\log(c_2) + 9 + \frac{8(\log(N) + \log(1/\epsilon))}{\log(N) + \log(1/\epsilon)} \\
&= 1.5\log(c_2) + 17.
\end{aligned}$$

Therefore, it is enough to satisfy $c_2/c_1 \geq 1.5\log(c_2) + 17$. But setting $c_2/c_1 = c_3$, this means we should choose $c_3$ large enough so that $c_3 \geq 1.5\log(c_3) + 1.5\log(c_1) + 17$. One can readily check that $c_3 \geq \max\{23, 2\log(c_1) + 23\}$ satisfies this inequality. Consequently, it is enough to pick $c_2$ as large as $c_2 \geq c_1 \max\{23, 2\log(c_1) + 23\}$, which completes the proof of Lemma 7. Note that $c_1 \max\{23, 2\log(c_1) + 23\}$ is almost a linear function of $c_1$. □

## 2.13 Modification of DPP-Tester for distinguishing $(\alpha, \zeta)$-normal DPPs from the $\epsilon$-far set of just the $(\alpha, \zeta)$-normal DPPs

Here, we explain how to manipulate the tester to work when we want to distinguish if $q$ is an $(\alpha, \zeta)$-normal DPP, or $\epsilon$-far only from the class of $(\alpha, \zeta)$-normal DPPs. We suggest that the reader first read the proof of Theorem 3.

The only part we change in the algorithm is the way we generate the set of candidate DPPs $\mathcal{M}$; we build the set of candidate marginal kernels $M$ the same way as in the proof of Theorem 3. Given a candidate kernel matrix $K \in M$ and an arbitrary entry $K_{i,j}$, depending on whether $K_{i,j}$ is zero, or picked from the confidence interval around $\hat{K}_{i,j}^{(+)}$ or $\hat{K}_{i,j}^{(-)}$, we define the value $\alpha_{i,j}(K)$ to be zero, $+\alpha$, or $-\alpha$ respectively. Now when we are in the case where the underlying distribution is DPP, according to the way we generate $M$, with high probability there exists a $\tilde{K} \in M$, such that $\tilde{K}_{i,j}$ is $\wp$-close to $K^*_{i,j}$ for every $i, j \in [n]$, and furthermore, $\alpha_{i,j}(\tilde{K})$ is zero if $K^*_{i,j} = 0$, or has the same sign as $K^*_{i,j}$ if $K^*_{i,j} \neq 0$ ($\wp$ is defined in Equation (2.12)). Our goal is to exploit this property of $\alpha_{i,j}(\tilde{K})$'s to redefine $\mathcal{M}$, so that the candidate DPPs in $\mathcal{M}$ are $(\alpha, \zeta)$-normal. To this end, for each matrix $K \in M$, instead of projecting $K$ onto the set of PSD matrices with eigenvalues in $[0, 1]$, we project onto the following convex body with respect to the Frobenius distance, which is a subset of $(\alpha, \zeta)$-normal DPPs:

$$D_K \coloneqq \{A \in S_n^+ | \, \zeta.I \preceq A \preceq (1-\zeta)I, \, \forall i, j \in [n] : \\ A_{i,j}/\alpha_{i,j}(K) \geq 1 \text{ if } \alpha_{i,j}(K) \neq 0, \text{ or } A_{i,j} = 0 \text{ if } \alpha_{i,j}(K) = 0\},$$

and generate $\mathcal{M}$ as

$$\mathcal{M} \coloneqq \{\mathbf{Pr}_{\Pi_{D_K}(K)}[.] | \, K \in M\},$$

where we denote by $\Pi_{D_K}$ the projection map onto $D_K$. Particularly, it is clear that $D_K$ is a subset of $(\alpha, \zeta)$-normal DPPs, and as the intersection of convex sets, $D_K$ is also convex, so projection on $D_K$ is well-defined.

Now when $q$ is a DPP with marginal kernel $K^*$, we know it is $(\alpha, \zeta)$-normal, so for every $i, j \in [n] : |K^*_{i,j}| \geq \alpha$. Combining this with the property that $\alpha_{i,j}(\tilde{K})$ is zero if $K^*_{i,j} = 0$, or it has the same sign as $K^*_{i,j}$ if $K^*_{i,j} \neq 0$, we obtain that $K^* \in D_{\tilde{K}}$. This means $\Pi_{D_{\tilde{K}}}(K^*) = K^*$. Using this relation with the contraction property of projection, we obtain

$$\|\Pi_{D_{\tilde{K}}}(\tilde{K}) - K^*\|_F = \|\Pi_{D_{\tilde{K}}}(\tilde{K}) - \Pi_{D_{\tilde{K}}}(K^*)\|_F \leq \|\tilde{K} - K^*\|_F.$$

Therefore, by substituting the projection $\Pi(K)$ in our algorithm by $\Pi_{D_K}(K)$ for every $K \in M$, the inequality in Equation (2.13) in the proof of Theorem 3 remains to hold, and the rest of the proof for the $\chi^2$-distance bound follows accordingly. On the other hand, with the new projection $\Pi_{D_K}(K)$ instead of $\Pi(K)$, the DPPs that are generated in $\mathcal{M}$ are all $(\alpha, \zeta)$-normal, so if we are in the case that $q$ is $\epsilon$-far from $(\alpha, \zeta)$-normal DPPs, it is also $\epsilon$-far from $\mathcal{M}$. Consequently, our $\chi^2$-$\ell_1$ tests are able to distinguish the two cases as before, and we obtain an $(\epsilon, 0.99)$-tester with sample complexity $\Theta(\sqrt{N}/\epsilon^2)$ for this modified version of our testing problem.

We should note that computing $\Pi_{D_K}(K)$ is trickier than $\Pi(K)$; for $\Pi(K)$, computing the Singular value decomposition (SVD) of $K$ is enough (or we can use iterative algorithms to get an approximate solution faster), but computing $\Pi_{D_K}(K)$ is a general convex problem and is solvable via convex programming approaches.

## 2.14 Analysis of `DPP-Tester2`

In this section, we show the argument in Theorem 5, which is a direct consequence of the sample and time complexities for the moment-based learning algorithm in [Urschel et al., 2017].

*Proof of Theorem 5.* Recall from the proof of Theorem 3 that estimating each entry of $K^*$ up to accuracy $\wp$, defined in Equation (2.12), is enough to prove the desired bound $\chi^2(q, \tilde{p}) \leq \epsilon^2/500$, which in turn enables the final $\chi^2$-$\ell_1$ tester to work correctly.

Now let $\mathcal{D}_n$ be the set of $n \times n$ diagonal matrices with $+1$ or $-1$ on their diagonal. For any $D \in \mathcal{D}_n$, the marginal kernel $DK^*D$ induces the same DPP distribution as $K^*$ does. In other words, $K^*$ is identifiable only up to the multiplication of its rows and columns by $\pm 1$. With this in mind, to get the final guarantee for closeness of the DPP distributions when we use the moment-based learning algorithm, i.e. $\chi^2\Big(q, \mathbf{Pr}_{K^{\text{new}}}[.]\Big) \leq \epsilon^2/500$, it is enough that for some $D \in \mathcal{D}_n$, we estimate the matrix $DK^*D$ entrywise with accuracy $\wp$. In fact, the moment-based learning algorithm gives us such a guarantee; according to [Urschel et al., 2017], in order to compute a $\wp$-accurate estimate of $K^*$ in *pseudo-distance*, the moment-based algorithm requires $O\Big(\Big(\frac{1}{\alpha^2\wp^2} + \ell(\frac{4}{\alpha})^{2\ell}\Big)\log(n)\Big)$ samples, where the pseudo-distance of matrices $K_1$ and $K_2$ is defined as

$$\rho(K_1, K_2) = \min_{D \in \mathcal{D}_n} \Big| DK_1D - K_2 \Big|_\infty = \min_{D \in \mathcal{D}_n} \max_{i,j \in [n]} \big| (DK_1D)_{i,j} - (K_2)_{i,j} \big|.$$

Now substituting $\wp$ from Equation (2.12), the sample complexity of the moment-based algorithm as a subroutine in DPP-Tester2 becomes

$$m = O\Big(\frac{n^4 \log(n)}{\epsilon^2 \alpha^2 \zeta^2} + \ell(\frac{4}{\alpha})^{2\ell} \log(n)\Big), \tag{2.32}$$

where $\ell$ is the cycle sparsity[2] of the graph with vertices $[n]$, whose edges correspond to the non-zero entries of $K^*$.

Adding the complexity of the final $\chi^2$-$\ell_1$ test to the learning complexity in Equation (2.32), the overall sample complexity of DPP-Tester2 is:

$$O\Big(\frac{n^4 \log(n)}{\epsilon^2 \alpha^2 \zeta^2} + \ell(\frac{4}{\alpha})^{2\ell} \log(n) + \frac{\sqrt{N}}{\epsilon^2}\Big).$$

For the time complexity, the run-time of the moment-based algorithm is $O(n^6 + mn^2)$ in the worst-case due to [Urschel et al., 2017], and the run-time of the $\chi^2$-$\ell_1$ test is $O(Nn^3 + m)$, as we have to compute $\mathbf{Pr}_{K^{\text{new}}}[J]$ for each $J \subseteq [n]$, requiring an SVD in time $O(n^3)$. Adding them up results in an overall run time of

$$O(Nn^3 + n^6 + mn^2) = O(\epsilon^4 m^2 n^3 + n^6 + mn^2) = \text{Poly}(m, n)$$

for DPP-Tester2, where the above equality follows from our sample complexity lower bound $m = \Omega(\sqrt{N}/\epsilon^2)$. □

## 2.15 Time complexity of DPP-Tester

In this section, we analyze the time complexity of DPP-Tester.

For each $p \in \mathcal{M}$, to apply the robust $\chi^2$-$\ell_1$ test of Acharya et al. [2015], one has to compute the statistic $Z^{(m)}$ defined in Equation (2.3). To compute $Z^{(m)}$, one should compute $p(J)$ for every $J \subseteq [n]$, which requires a determinant calculation in time $O(n^3)$. Therefore, each robust $\chi^2 - \ell_1$ testing takes time $O(Nn^3)$. There is another $O(m)$ pre-processing time for computing $N(J)$'s. Moreover, computing the projection

[2] The cycle sparsity of a graph is the smallest $\ell'$ such that the cycles with length at most $\ell'$ constitute a basis for the cycle space of the graph.

matrix $\Pi(K)$ for every $K \in M$ requires the Singular value decomposition (SVD) of $K$, which takes time $O(n^3)$. This is because we project with respect to the Frobenius distance, and it follows from the 2-*Weilandt-Hoffman* inequality [Tao, 2012] that computing $\Pi(K)$ can equivalently be done by rounding down the eigenvalues of $K$ that are larger than one to one, and rounding up the eigenvalues that are negative to zero. Computing the initial estimate of the marginal kernel, i.e. $\hat{K}$ in the proof of Theorem 3, also takes time at most $O(\min\{N, m\}n^2)$. Therefore, the overall time complexity becomes

$$O(|\mathcal{M}|Nn^3 + m).$$

To have a time complexity upper bound only in terms of the main variables $n$, $\epsilon$, note that based on what was discussed in section 2.5.2, for the general DPPs without the knowledge of $\zeta$ and $\alpha$, we set the normality parameters in our algorithm as $(\alpha, \zeta) = (0, \bar{z})$, where $\bar{z}$ is $0.005/(2m^*n)$, for $m^* = O(C_{N,\epsilon}\sqrt{N}/\epsilon^2)$. Substituting $C_{N,\epsilon} = \log^2(N)(\log(N) + \log(1/\epsilon))$, we get that $\bar{z}^{-1} = O\Big((n^4 + n^3\log(1/\epsilon))\sqrt{N}/\epsilon^2\Big)$. Substituting $\zeta = \bar{z}$ in Theorem 3, in the definition of $\varsigma$ and ignoring $\alpha$ in the min term, we obtain the following worst-case scenario upper bound on $\varsigma$:

$$\begin{aligned} \varsigma =& O((n^2\zeta^{-1}\sqrt{\xi/\epsilon}) = O\Big(n^2(n^4 + n^3\log(1/\epsilon))\sqrt{N}/\epsilon^2)N^{-\frac{1}{8}}\log^{\frac{1}{4}}(n)\epsilon^{-0.5}\Big) \qquad (2.33)\\ =& O\Big(\epsilon^{-2.5}(n^6 + n^5\log(1/\epsilon))N^{\frac{3}{8}}\log^{\frac{1}{4}}(n)\Big). \qquad (2.34)\end{aligned}$$

Therefore,

$$|\mathcal{M}| = O\Big(\epsilon^{-2.5}(n^6 + n^5\log(1/\epsilon))N^{\frac{3}{8}}\log^{\frac{1}{4}}(n)\Big)^{n^2}.$$

But notice that our matrices are symmetric, hence, we only have to consider different candidates for at most $n(n+1)/2$ entries, which reduces the size of $|\mathcal{M}|$ to

$$|\mathcal{M}| = O\Big(\epsilon^{-2.5}(n^6 + n^5\log(1/\epsilon))N^{\frac{3}{8}}\log^{\frac{1}{4}}(n)\Big)^{n(n+1)/2}.$$

## 2.16 Lower bound on the Sample Complexity of Distinguishing the Uniform distribution from $\mathcal{F}$

In this section, we give a high-level sketch of the approach that Diakonikolas and Kane [2016] use, to argue a lower bound of $\Omega(\sqrt{N}/\epsilon^2)$ on the sample complexity of the problem of testing the uniform distribution against $h_r$, randomly selected from $\mathcal{F}$.

*Proof.* Suppose that we observe samples from the underlying distribution $g$, where $g$ can either be $h_r$ or the uniform distribution. We flip a random coin $X$, and based on that set $g$ to the uniform distribution, or to $h_r$, a distribution randomly selected from $\mathcal{F}$. For every $S \subseteq [n]$, let $N(S)$ be the number of samples that are equal to $S$. We aim to show that given the number of samples satisfy $m = o(\sqrt{N}/\epsilon^2)$, the information in the collection of random variables $\mathcal{A} = \{N(S) | \, S \subseteq [n]\}$ is not enough to guess the value of $X$ strictly better than random guessing, say with success probability greater than 0.51.

To begin, we use the following Lemma without proof, which is exactly Lemma 3.2. in page 19 of [Diakonikolas and Kane, 2016]. This is a classical result in Information theory:

**Lemma 8.** *For random variables $X$ and $\mathcal{A}$, if there exist a function mapping $\mathcal{A}$ to $X$ such that $f(\mathcal{A}) = X$ with probability at least* 0.51*, then we have the following bound on their mutual information:*

$$I(X; \mathcal{A}) \geq 2.10^{-4}.$$

Based on Lemma 8, it is enough to show that $I(X; \mathcal{A}) = o(1)$. To continue, we use the Poissonization trick; instead of directly deriving $m$ samples from $g$, we sample $m'$ from the Poisson distribution with parameter $m$, namely $m' \sim$ Poisson(m), then derive $m'$ samples from $g$. Using this trick, we still have $m' = \Theta(m)$ samples with high probability, so it is enough to bound $I(X, \mathcal{A})$ for $\mathcal{A}$ with respect to the new sampling scheme with Poissonization. Based on properties of the Poisson distribution, the new scheme is equivalent to deriving $N(S) \sim \text{Poisson}(mg(S))$ for each set $S \subseteq [n]$

independent from the others. Furthermore, we showed in the proof of Theorem 4 that $L_r = \Theta(1)$ with high probability, so by using $mL_r$ instead of $m$ samples, the order of sample size does not change. But now, in the case $g = h_r$, $N(S)$ is sampled according to $N(S) \sim \text{Poisson}(mL_r h_r(S)) = \text{Poisson}(m\bar{h}_r(S))$. Thus, one can readily see that again, we can substitute $h_r$ by its unnormalized counterpart $\bar{h}_r$ in our Poisson sampling.

Finally, assuming the sampling scheme $N(S) \sim \text{Poisson}(m\bar{h}_r(S))$, $\forall S \subseteq [n]$, we bound $I(X, \mathcal{A})$. Note that given the value of $X$, the random variables $\{N(S)\}$ are independent, so we have the following bound on the mutual information:

$$I(X; \mathcal{A}) \leq \sum_{S \subseteq [n]} I(X; N(S)). \tag{2.35}$$

It is enough to bound each of the terms $I(X; N(S))$. For that, we bring without proof Lemma 3.3. from [Diakonikolas and Kane, 2016], page 20:

**Lemma 9.** *If $N(S) \sim$ Poisson$(m\bar{h}(S))$ for $X = 0$ and $N(S) \sim$ Poisson$(m/N)$ for $X = 1$, then:*

$$I(X; N(S)) = O(m^2\epsilon^4/N^2).$$

From this Lemma and Equation (2.35), we get $I(X; \mathcal{A}) = o(m^2\epsilon^4/N) = \mathrm{o}(1)$. Combining this with Lemma 8, we conclude that we need at least $\Omega(\sqrt{N}/\epsilon^2)$ samples to non-trivially guess $X$ from the observed samples. This completes the proof of the promised lower bound on the sample complexity of the problem of testing uniform distribution against $\mathcal{F}$. For more details and the proof of Lemmas 8 and 9, we refer the reader to [Diakonikolas and Kane, 2016]. □

## 2.17 Experiments

Finally, we perform small-scale synthetic experiments as a proof of concept.

We generate random DPPs for $n = 4$ by randomly generating kernel matrices

$K$. We draw the eigenvalues of each $K$ uniformly from $[0,1]$, and use eigenvectors of random matrices with entries uniformly sampled from $[0,1]$. To generate a $\Theta(\epsilon)$-far distribution from the class of DPPs, we use our lower bound approach in section 2.6: we add a random perturbation of $\pm\frac{\epsilon}{N}$ to each atom probability of the uniform distribution over $2^n$. Lemma 4 implies that for sufficiently large $n$ and small $\epsilon$, with high probability, we are $\Theta(\epsilon)$ far from the class of DPPs, where the constant in $\Theta(\epsilon)$ is in the range $[1/1024, 1]$. Since we do not know the exact value of this constant, we use the constant $1/2$ to compute the algorithm's acceptance threshold: $C = m(\frac{\epsilon}{2})^2/10$.

We simplified the algorithm slightly in two ways: (1) instead of projecting the candidate matrices, we just ignore the ones that have an eigenvalue outside the range $[0,1]$; (2) Instead of checking multiple candidate entries in the confidence intervals for each $K^*_{ij}$, we only consider the two signed values $+|\widehat{K}_{ij}|$ and $-|\widehat{K}_{ij}|$. The results are obtained by averaging the empirical probabilities over 20 runs.

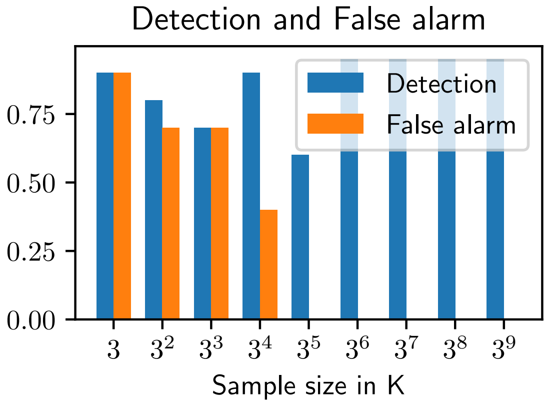

Figure 2-1: Detection and False Alarm rates of the testing algorithm for various numbers of samples and $\epsilon = 0.02$.

Figure 2-1 shows the performance of our tester for various number of samples: detection rate when the underlying distribution is a DPP (blue bars), and False Alarm rate when it is $\Theta(\epsilon)$ far from the class of DPPs (orange bars). For $\epsilon = 0.02$, and the $C$ we picked here, the algorithm correctly accepts most DPPs, but needs more samples to correctly reject non-DPPs.

Our adaptive sample complexity has a weak logarithmic dependence on $\zeta^{-1}$; as a reminder, $\zeta$ measures how close the eigenvalues of $K$ are to zero or one. The coupling

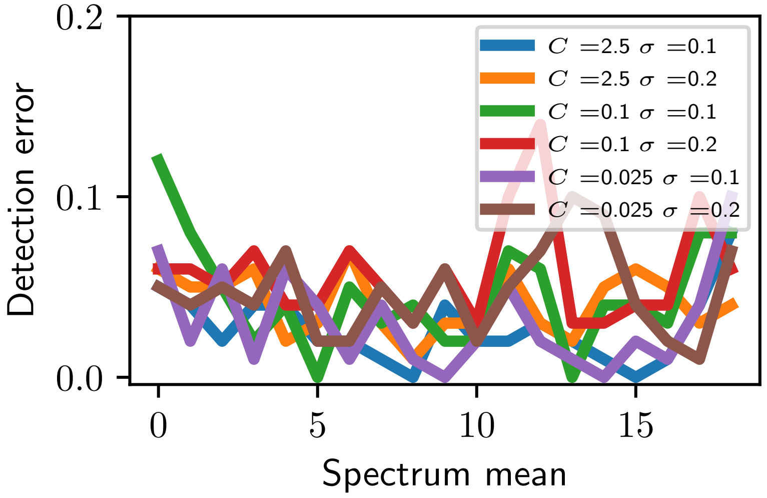

Figure 2-2: Detection errors of the testing algorithm for DPP kernel matrices with eigenvalues sampled from a conditional normal distribution, with different means, variances, and over multiple choices of the algorithm's threshold $C$.

argument in Lemma 3 got rid of this dependence, for $\zeta$ below some threshold. This theory motivates the question how much the accuracy of our tester depends on the spectrum of $K$, in particular, on the distribution of its eigenvalues. To investigate this for $n = 4$, we sample the eigenvalues of $K$ from a normal distribution with mean on one of the equidistant points $0.05, 0.1, \ldots, 0.9, 0.95$ and standard deviations 0.1 or 0.2, conditioned on the interval $[0, 1]$.

Figure 2-2 shows the results for a variety of parameters. The $x$-axis is the mean of the normal distribution, while the $y$-axis is the empirical value of the error probability in Detection (i.e. recovering the underlying DPP), averaged over 100 runs for each setting of the parameters. The sample size is 10000 here. The results suggest that the detection accuracy is only very weakly affected by the mean of the eigenvalues of $K$ and, in particular, the error does not increase a lot at the boundaries.

# Chapter 3

# Optimization and Adaptive Generalization of three layer Neural Networks

### Abstract

While there has been substantial recent work studying generalization of neural networks, the ability of deep networks in automating the process of feature extraction still evades a thorough mathematical understanding. As a step toward this goal, we analyze learning and generalization of a three-layer neural network with ReLU activations in a regime that goes beyond the linear approximation of the network and is hence not captured by the common Neural Tangent Kernel. We show that despite nonconvexity of the empirical loss, a variant of SGD converges in polynomially many iterations to a good solution that generalizes. In particular, our generalization bounds are adaptive: they automatically optimize over a family of kernels that includes the Neural Tangent Kernel to provide the tightest bound.

## 3.1 Introduction

The ability of overparameterized neural networks trained by (stochastic) gradient descent to generalize well on test data [Krizhevsky et al., 2012, Silver et al., 2016, Hinton et al., 2012], even if they perfectly fit the the training data, has intrigued theoretical researchers and led to many approaches for generalization bounds [Neyshabur et al., 2015, Bartlett et al., 2017, Neyshabur et al., 2018, Dziugaite and Roy, 2017, Wei et al., 2019, Golowich et al., 2018, Arora et al., 2018b, Zhou et al., 2018, Konstantinos et al., 2017]. This generalization ability is tied to the optimization procedure, i.e., the trajectory of the training algorithm in a non-convex loss landscape, and the structure of the data.

Hence, several recent works study the training of neural networks. For instance, Safran and Shamir [2018] address the role of overparametrization in avoiding bad local minima, and Zhang et al. [2016] empirically show that overparametrized networks trained by SGD can even perfectly fit to random labels. Within the popular framework of the *Neural Tangent Kernel (NTK)* [Jacot et al., 2018], which uses a linear approximation of the network at initialization, several works analyze the optimization trajectory and show global convergence of (S)GD to a global optimum of the empirical loss [Allen-Zhu et al., 2019b, Li and Liang, 2018, Zou et al., 2018, Du et al., 2018]. Extending the viewpoint to generalization, Arora et al. [2019a,b] exploit the kernel-like behaviour of two-layer networks close to their initialization to prove generalization for the final network, showing that two-layer neural networks generalize as well as Kernel Ridgeless Regression (KRLR) with the NTK. Cao and Gu [2019] show a tighter bound with a *Neural Tangent Random Feature Model.* The kernel approach, however, has two main limitations: First, while KRLR can generalize well in specific high dimensional regimes [Liang et al., 2020], there is theoretical and empirical evidence that it can be inconsistent with noise [Rakhlin and Zhai, 2019]. *Is there an approach for analyzing neural networks that shows they perform at least as well as KRLR, but is also robust to noise?*

Second, importantly, neural networks are known to outperform traditional statistical methods in many regimes as they are able to automate the process of feature extraction from data, as opposed to kernel methods that work with a fixed feature representation. This poses the question of other, adaptive, regimes beyond the linear network approximation. In this realm, Wu et al. [2018] show generalization bounds that, instead of the NTK norm, scale with respect to another functional norm. This norm corresponds to the minimum RKHS norm of the function among a family of kernels, i.e., their method in a sense picks the best kernel in this family. However, this result ignores the computational aspect of the problem. *Are there particular nonlinear regimes beyond NTK for which a gradient-type polynomial-time algorithm, in a way, adaptively chooses a suitable kernel?*

Going beyond the NTK view, a line of work convexifies the optimization problem via an approximation of SGD dynamics with a continuous time gradient flow in the space of probability measures on the hidden units of the network, equipped with the Wasserstein metric [Mei et al., 2018, Chizat and Bach, 2018, Mei et al., 2019, Wei et al., 2018, Sirignano and Spiliopoulos, 2020, Javanmard et al., 2019, Lu et al., 2020]. Taking another perspective, Allen-Zhu et al. [2018] consider a three-layer network model that is not captured by the NTK approximation, and learn an underlying concept class by exploiting saddle-point escape theory for nonconvex SGD [Ge et al., 2015a]. However, evaluating the complexity measure of Allen-Zhu et al. [2018] is rather involved, and only aligns well with functions that are described by a particular network. Whether one can recover the NTK bound (e.g. the NTK norm) from these results is not clear. For the NTK setting, in contrast, Arora et al. [2019a] develop a purely data dependent generalization bound. *Going beyond two layers, is it possible to prove a **data-dependent** complexity measure beyond the NTK regime that recovers the NTK result [Arora et al., 2019a] as a special case?*

In this work, we address the above questions:

- We consider a regime for 3-layer neural networks that is not captured by the NTK approximation and show that, despite nonconvexity, a variant of projected SGD finds a good solution, as measured by the regularized empirical loss, importantly, after polynomially many iterations.

- We introduce a new function norm $\|.\|_\zeta$ as the minimum RKHS norm with respect to a family of kernels $\mathcal{K}$, which is upper bounded by the NTK norm up to constants. We show that for an arbitrary function $f$, the generalization gap of the trained network scales by $\|f\|_\zeta$. This makes our generalization bound adaptive, in the sense that it scales with the best kernel in $\mathcal{K}$. As a byproduct, our bounds are comparable with kernel regression bounds simultaneously with all kernels in $\mathcal{K}$. We hope that our techniques motivate researchers to prove such adaptive generalization bounds for deeper networks, which can potentially result in stronger depth separation.

- We show generalization bounds with a new data-dependent complexity measure that generalizes the NTK-based complexity in [Arora et al., 2019a]. Up to logarithmic factors, our bounds are upper bounded by those NTK-based bounds and hence improve over them (if one substitutes their Lipschitz loss with a smooth one) – see Section 3.6.1 for a simple explicit example. Importantly, our bound can also handle noisy distributions as opposed to [Arora et al., 2019a].

**Further Related work.** While the idea of a learning algorithm that combines multiple kernels has been employed for a while in the community [Sonnenburg et al., 2006, Rakotomamonjy et al., 2007, Duan et al., 2012], our understanding of the connections between deep learning and multiple kernel learning is yet in its infancy. Recently, Dou and Liang [2020] define a time-varying kernel based on the network weights and show that the limit of the gradient flow converges to a suitable dynamic kernel, in the sense that the residual of the link function onto its RKHS could be in a smaller ranked space compared to the orthogonal complement of the RKHS. Ghorbani et al. [2019] analyze the difference between training a two layer ReLU network and its NTK or random feature simplifications, for a mixture of Gaussians input distribution and quadratic target functions. Ignoring the computational hardness imposed by nonconvexity, Bach [2017] prove a dimension dependent generalization bound beyond NTK. In another line of work, Chizat and Bach [2020] study gradient flow on losses with exponential tail and its relation to the max margin solution. Wei et al. [2019] show an interesting separation between the learning power of two layer ReLU networks and their NTK approximation, by showing a sample complexity gap for an artificially constructed distribution.

With a different approach, Allen-Zhu and Li [2020] analyze multi-layer networks with quadratic activations, and prove generalization bounds polynomial in the dimension and precision by assuming an underlying teacher network, which shows a remarkable algorithmic depth separation. The problem of depth separation for neural networks and more generally their expressive power has been investigated by several researchers before [Raghu et al., 2017, Daniely, 2017, Barron, 1994, Funahashi, 1989, Safran and Shamir, 2016, Safran et al., 2019]. The assumption of an underlying teacher network that one seeks to recover is common, too [Li and Yuan, 2017, Zhong et al., 2017, Brutzkus and Globerson, 2017]. Other works focus mainly on the algorithm and use other techniques, such as tensor factorization, to find a global optimum [Tian, 2016, Bakshi et al., 2019, Janzamin et al., 2015, Zhong et al., 2017]. Finally, many authors study the loss landscape under various assumptions [Freeman and Bruna, 2016, Nguyen and Hein, 2017, Soudry and Carmon, 2016, Soltanolkotabi et al., 2018, Ge et al., 2017], some of them consider the simplified case of deep linear networks [Arora et al., 2018a, Saxe et al., 2013, Bartlett et al., 2018, Kawaguchi, 2016a].

## 3.2 Setup and approximation by kernels

We analyze a 3-layer ReLu neural network from inputs $x \in \mathbb{R}^d$ to outputs $y \in \mathbb{R}$ of the form

$$f_{V',W'}(x) = \tfrac{1}{\sqrt{m_2}} a^T \sigma\Big((V^{(0)} + V')W^s \tfrac{1}{\sqrt{m_1}} \sigma((W^{(0)} + W')x)\Big), \tag{3.1}$$

where $a \in \mathbb{R}^{m_2}$ is a vector of random signs, $V^{(0)} \in \mathbb{R}^{m_2 \times m_3}$ and $W^{(0)} \in \mathbb{R}^{m_1 \times d}$ are random weight initializations with i.i.d Gaussian entries $V^{(0)}_{j,k} \sim \mathcal{N}(0, \kappa_2^2), W^{(0)}_{j,k} \sim \mathcal{N}(0, \kappa_1^2)$, and $W^s \in \mathbb{R}^{m_3 \times m_1}$ is a random sign matrix, which is roughly a random projection and change of coordinates into a lower dimensional space. We refer to $W^s \frac{1}{\sqrt{m_1}} \sigma((W^{(0)} + W')x)$ as the first layer and $\frac{1}{\sqrt{m_2}} a^T \sigma((V^{(0)} + V')(.))$ as the second layer. The algorithm trains weight matrices $V'$ and $W'$, and $W^s$, $a$ are fixed. We assume that the outputs are a.s. bounded by a constant, $|y_i| \leq B$, and $\|x_i\| = 1$. As loss $\ell(.,.)$, we use the squared loss. We denote the training (empirical) loss of a function $f$ on our data $\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n$ and the expected loss with respect to the data distribution (population loss) by

$$R_n(f) = \tfrac{1}{n}\sum\nolimits_{i=1}^{n} \ell(f(x_i), y_i), \quad \text{and} \quad R(f) = \mathbb{E}\ell(f(x), y),$$

respectively. Sometimes, we refer to the vector of labels $(y_i)_{i=1}^n$ by $y$. Finally, $\mathcal{H}_K$ is the space of functions with bounded RKHS-norm of kernel $K$, and the notation $\tilde{O}$ hides log factors.

### 3.2.1 Kernel approximations, decomposition and adaptivity

Kernel approximations of neural networks play an important role in our analysis. First, a common approximation is the NTK. The Neural Tangent kernel for a 2-layer ReLu network is

$$H^{\infty}(x_1, x_2) = \langle x_1, x_2 \rangle \cdot F_2\Big(\langle x_1, x_2\rangle / (\|x_1\|\|x_2\|)\Big), \qquad \text{for } \; F_2(x) = \tfrac{1}{4} + \arcsin(x)/(2\pi). \tag{3.2}$$

To introduce adaptivity, a key part of our analysis is to approximate the second layer in the 3-layer network by a product kernel $K^{\infty} \odot G$ that decomposes into a "fixed" part $K^{\infty}$ and an "adaptive" part $G$. To define these kernels, for every $i \in [n]$, let $\phi^{(0)}(x_i)$ be the output of the first layer of the network at initialization, $\phi^{(0)}(x_i) = \frac{1}{\sqrt{m_1}} W^s \sigma(W^{(0)} x_i)$, and $\phi^{(0)}(x_i) + \phi'(x_i)$ be that output for weights $W^{(0)} + W'$. The adaptive kernel $G$ captures the dot product between the learned weights:

$$G(x_i, x_j) = \langle \phi'(x_i), \phi'(x_j) \rangle. \tag{3.3}$$

This form of $G$ motivates the complexity measure we define in the next section, if one thinks of the entries of $\phi'$ as bounded NTK-norm functions of the input. Next, we consider the second layer, where the part $K^{\infty}$ arises from roughly stable activations. To formalize this stability, let $\mathrm{Sgn}(Vx)$ be the diagonal matrix whose diagonal contains the coordinate-wise signs of the vector $Vx$. If we assume that $\mathrm{Sgn}\big((V^{(0)} + V')(\phi^{(0)}(x_i) + \phi'(x_i))\big) \approx \mathrm{Sgn}\big(V^{(0)}\phi^{(0)}(x_i)\big)$ – we prove a rigorous statement in Section 3.6.12 – then

$$f_{W',V'}(x_i) = \tfrac{1}{\sqrt{m_2}} a^T (V^{(0)} + V')(\phi^{(0)}(x_i) + \phi'(x_i)) \tag{3.4}$$

$$\approx \Big\langle V^{(0)} + V', \tfrac{1}{\sqrt{m_2}} a^T \mathrm{Sgn}\Big(V^{(0)}\phi^{(0)}(x_i)\Big)\Big(\phi^{(0)}(x_i) + \phi'(x_i)\Big)^T \Big\rangle. \tag{3.5}$$

Focusing on the adaptive part $\phi'(x)$ of the first layer, we write

$$\Big\langle V^{(0)} + V', \tfrac{1}{\sqrt{m_2}} a^T \mathrm{Sgn}\big(V^{(0)}\phi^{(0)}(x_i)\big)\phi'(x_i)^T \Big\rangle := \Big\langle V^{(0)} + V', \Upsilon(x_i) \Big\rangle \tag{3.6}$$

and can then view the second layer as a function in the RKHS of the product kernel $\langle \Upsilon(x_i), \Upsilon(x_j)\rangle =: \tilde{K}^\infty(x_i, x_j) G(x_i, x_j)$. This defines the new kernel $\tilde{K}^\infty$, which we simplify into the kernel $K^\infty$ that is independent of initialization ($K^\infty$ is defined in Equation (3.8)). To do so, we first observe that $\tilde{K}^\infty$ concentrates around

$$\mathbb{E}_{w\sim\mathcal{N}(0,\kappa_2^2 I)} \mathbf{1}\{w^T\phi^{(0)}(x_i)\}\mathbf{1}\{w^T\phi^{(0)}(x_j)\} = F_2\Big(\frac{\langle \phi^{(0)}(x_i), \phi^{(0)}(x_j)\rangle}{\|\phi^{(0)}(x_i)\|\|\phi^{(0)}(x_j)\|}\Big).$$

Moreover, the Gaussian initialization and assumption $\|x_i\| = 1$ imply that $\langle \phi^{(0)}(x_i), \phi^{(0)}(x_j)\rangle$ concentrates around $m_3 F_3(\langle x_i, x_j\rangle)$, for $F_3 : [-1, 1] \to [0, \frac{1}{2}]$ defined as

$$F_3(x) = \tfrac{1}{2\pi}\sqrt{1 - x^2} + \tfrac{1}{4}x + \tfrac{1}{2\pi} x \arcsin(x), \tag{3.7}$$

$$\text{so} \quad \tilde{K}^\infty(x_i, x_j) \approx F_2(2F_3(\langle x_i, x_j\rangle)) =: K^\infty(x_i, x_j). \tag{3.8}$$

For general $x_1, x_2$ not necessarily unit norm, we define $K^\infty(x_1, x_2) = K^\infty(x_1/\|x_1\|, x_2/\|x_2\|)$.

It is easy to check that the coefficients in the Taylor series of $F_2$ and $F_3$ are nonnegative. Combining this with Schur's Product Theorem implies $K^\infty$ is PSD (Section 3.6.4). We also denote the data kernel matrix on $(x_i)_{i=1}^n$ by $K^\infty$ and $H^\infty$. Like Arora et al. [2019a], we assume the data distribution is $(\lambda_0, \delta, n)$-non-degenerate with respect to $H^\infty$ and $K^\infty$, i.e., with probability at least $1-\delta$, the smallest eigenvalues of $H^\infty$ and $K^\infty$ are at least $\lambda_0 > 0$.

## 3.3 Data dependent complexity measure and generalization

The emergence of $G \odot K^\infty$ above gives rise to an *adaptive* kernel-like complexity measure that will determine generalization bounds. Intuitively, this complexity measure reflects the two layers. Here, we view $G$ as the Gram matrix of some "ideal" first-layer feature functions $g_k$. We measure the complexity of the prediction function via the RKHS of $A := G \odot K^\infty$, and allow a flexible choice of the features $g_k$. The $g_k$ may be viewed as feature representation of $\phi'$ in Equation (3.3): $G(x_i, x_j) := \sum g_k(x_i) g_k(x_j)$. To regularize this choice, we penalize the complexity of the features $g_k$ via the NTK norm. Alternatively, the features $g_k$ are flexible but have bounded NTK norm.

For a labeling $f^* \in \mathbb{R}^n$ of the $n$ data points and fixed $G$, this leads to the complexity

$$\zeta(f^*, G) \;=\; f^{*T} A^{-1} f^* \cdot \langle H^{\infty -1}, G \rangle \;=\; f^{*T} A^{-1} f^* \sum\nolimits_{k=1}^{m_3} \|g_k\|_{H^\infty}^2; \quad A = G \odot K^\infty, \tag{3.9}$$

where $m_3$ is the number of intermediate features. The choice of $m_3$ is discussed in Section 3.6.13. Our data-dependent complexity measure implicitly selects the $G$ (or equivalently the feature vectors $g_k$) that leads to the tightest bound, trading off data fit and function complexity:

$$\Im = \Im((x_i)_{i=1}^n, (y_i)_{i=1}^n) := \min_{f^* \in \mathbb{R}^n} \Big\{ 2nR_n(f^*) + \varpi \min_{G \succ 0} \zeta(f^*, G) \Big\}, \tag{3.10}$$

where we use a log factor $\varpi = O(\log(n)^3 + \log(1/\lambda_0))$.

To make the relation to adaptive kernel spaces even more explicit, assume that the $g_k$ are bounded as $\sum_k \|g_k\|_{H^\infty}^2 \le 1$. Then we define a family $\mathcal{K}$ of corresponding kernels of the form

$$K_{\{g\}}(x_1, x_2) = K^\infty(x_1, x_2) \big( \sum\nolimits_{g \in \{g\}} g(x_1) g(x_2) \big), \tag{3.11}$$

i.e., $\mathcal{K} := \{K_{\{g\}} |\ \{g\} \text{ finite}, \sum_{g \in \{g\}} \|g\|_{H^\infty}^2 \le 1\}$. With this notation, the complexity measure is

$$\Im((x_i), (y_i)) = \min_{f^* \in \mathbb{R}^n} \Big\{ 2nR_n(f^*) + \varpi \min_{K \in \mathcal{K}} f^{*T} K^{-1} f^* \Big\}. \tag{3.12}$$

Hence, this measure may be understood as searching for the most efficient and effective feature representation within a family of RKHSs.

We may also relate this complexity measure to the NTK-based complexity measure $y^T H^{\infty -1} y$ [Arora et al., 2019a]. For any labeling $f^*$, let $\tilde{f}^* \in \mathcal{H}_{H^\infty}$ be the function with minimum NTK norm that maps $x_i$'s to $f_i^*$'s, so $\|\tilde{f}^*\|_{H^\infty} = f^{*T} H^{\infty -1} f^*$. If we set $\{g\} = \{\tilde{f}^*/\|\tilde{f}^*\|\}$, then one can show (Section 3.6.5)

$$f^{*T} K^{-1}_{\{\tilde{f}^*/\|\tilde{f}^*\|\}} f^* \leq 4 f^{*T} (f^* f^{*T})^{-1} f^* \times \|\tilde{f}^*\|^2 = 4\|\tilde{f}^*\|^2 = 4 f^{*T} H^{\infty -1} f^*, \tag{3.13}$$

which implies

$$\Im \leq \min_{f^* \in \mathbb{R}^n} \Big\{ 2nR_n(f^*) + (4\varpi) f^{*T} H^{\infty -1} f^* \Big\}.$$

One can further set $f^* = y$ above and obtain

$$\Im \leq (4\varpi_1) y^T H^{\infty -1} y. \tag{3.14}$$

### 3.3.1 Generalization

With the complexity $\Im$ in hand, we can now state our generalization result. It assumes optimization by a Projected Stochastic Gradient Descent (PSGD), described in detail in Section 3.4.

**Theorem 6.** *Suppose we run Projected Stochastic Gradient Descent (PSGD) on the regularized empirical risk with parameters as in Section 3.4, and $|y_i| \leq B$ a.s.. Then, with high probability (e.g. 0.99) over the randomness of data, initialization and noise of the gradient steps, PSGD converges in poly$(B \vee 1/B, 1/\lambda_0, n)$ iterations to a solution*

$(W_{PSGD}, V_{PSGD})$ *with population risk bounded as*

$$R(f_{W_{PSGD}, V_{PSGD}}) \leq \frac{\Im((x_i)_{i=1}^n, (y_i)_{i=1}^n)}{n} + \frac{B^2 \varpi}{n}. \tag{3.15}$$

As a side remark, the factor 2 in front of $R_n(f^*)$ in the definition of $\Im$ in (3.12) is not special and a similar generalization bound can be obtained for any $\gamma > 1$. Substituting the upper bound on the complexity in Equation (3.14), one recovers an NTK-based generalization bound that scales with $y^T H^{\infty^{-1}} y/n$ up to log factors, which is roughly the square of the generalization bound presented in [Arora et al., 2019a]. The reason for the faster squared rate here is that we are considering smooth losses, while they work with a bounded Lipschitz loss. Indeed, it is not hard to apply a more rigorous uniform convergence analysis from [Srebro et al., 2010] to also obtain a faster squared rate for the approach used in [Arora et al., 2019a].

Since Equation (3.14) is an upper bound on our complexity, our result generalizes and tightens the NTK bound [Arora et al., 2019a]. To illustrate the flexibility of our complexity measure, we show in Section 3.6.1 a simple explicit example of functions represented as polynomial series where our bound improves upon the NTK bound. Notably, we only substitute low-rank matrices $G$ in our complexity measure for this construction. We leave further investigation of our complexity measure for arbitrary $G$'s to future work.

### 3.3.2 Underlying Concept class

Instead of data dependent generalization bounds, one may study the generalization gap with respect to some concept class. The complexity measure $\Im$ implicitly uses the following adaptive norm on the space of functions from $\mathbb{R}^d$, the infimum of the RKHS norms for the family of kernels $\mathcal{K}$:

$$\|f\|_\zeta = \inf_{K_{\{g\}} \in \mathcal{K}} \|f\|_{K_{\{g\}}}. \tag{3.16}$$

It is not hard to check that $\|.\|_\zeta$ is in fact a norm, and that the inf is achieved by a particular set $\{g\}$. Similar to the derivation of the upper bound on the complexity measure in Equation (3.14), by setting $\{g\} = \{f^*/\|f^*\|_{H^\infty}\}$, we obtain the following NTK upper bound:

$$\|f\|_\zeta \leq 4\|f\|_{H^\infty}. \tag{3.17}$$

This leads to a function-dependent generalization bound which bounds the risk of the learned network against an arbitrary function $f$ with $\|f\|_\zeta < \infty$.

**Theorem 7.** *For any measurable function $f : \mathbb{R}^d \to \mathbb{R}$, in the same setting as Theorem 6, the population risk of the trained network can be bounded as*

$$R(f_{W_{PSGD},V_{PSGD}}) \leq 2R(f) + O\big(\varpi\tfrac{\|f\|_\zeta^2+B^2}{n}\big). \tag{3.18}$$

As in the data-dependent case, the factor 2 on $R(f)$ can be reduced to any constant $\gamma > 1$.

### 3.3.3 Interaction of layers beyond the linear approximation

Here, we give a high level intuition on how the adaptivity is achieved in our regime compared to NTK. In the NTK approach, for every input $x$, the neural net $f_{W,V}(x)$ is approximated by its linear approximation at $(W^{(0)}, V^{(0)})$ (the initialized network), $f_{W,V}(x) = \langle \nabla_{W,V} f_{W^{(0)},V^{(0)}}(x), (W - W^{(0)}, V - V^{(0)})\rangle$. The NTK approximation works as long as $(W, V)$ are close enough to their initialization that the linear approximation remains accurate and the interaction of weights between layers is negligible. Specifically, the features $\phi'(x_i)$ behave almost linearly with respect to $W - W^{(0)}$ as $\|W - W^{(0)}\|$ is taken to be small and the sign pattern $\mathrm{Sgn}\big((W^{(0)} + W')x_i\big)$ is proven not to change much compared to $\mathrm{Sgn}\big(W^{(0)}x_i\big)$. Additionally, the NTK-type analysis needs the following two conditions to be satisfied: (1) the sign pattern of $\phi^{(0)}(x_i) + \phi'(x_i)$ with respect to $V^{(0)} + V'$ remains almost the same as the sign pattern of $\phi^{(0)}(x)$ with respect to $V^{(0)}$, and (2) the weight changes $W'$ and $V'$ should not interact, which means the "interaction" term, $\frac{1}{\sqrt{m_2}} a^T \mathrm{Sgn}\Big(V^{(0)}\phi^{(0)}(x_i)\Big) V'\phi'(x_i) \approx 0$, should be negligible. Therefore, the non-negligible terms for the NTK are: (1) $\frac{1}{\sqrt{m_2}} a^T \mathrm{Sgn}\Big(V^{(0)}\phi^{(0)}(x_i)\Big) V^{(0)}\phi'(x_i)$, which is almost linear in $W'$ (recall that $\phi'(x_i)$ depends on $W'$), and (2)$\frac{1}{\sqrt{m_2}} a^T \mathrm{Sgn}\Big(V^{(0)}\phi^{(0)}(x_i)\Big) V'\phi^{(0)}(x_i)$ which is linear in $V'$. This approach has two important implications: (1) it convexifies the optimization (for convex loss), as the approximation is now linear in $W$; and (2) it simplifies proving generalization, as it works with the class of functions in the RKHS space of some fixed kernel. However, this simplification leaves no room for the ability of the neural network to learn intermediate feature representations.

In our regime, in contrast, we enforce the condition $\forall j, i: \; V'_j \perp \phi^{(0)}(x_i)$ $(\star)$, which implies the second (2) above is zero, while the interaction term is not negligible any more and the network behaves similar to a quadratic function with respect to $(W', V')$ (for fixed $x_i$). Condition $(\star)$ is critical both in proving the convergence of the algorithm as well as bounding the Rademacher complexity of the class of networks with bounded weights. Rather than working with a fixed kernel, the interaction term enables us to use the first layer for representing the input in a suitable feature space, which can be interpreted as picking a suitable kernel, then use the second layer to describe the output based on those features. This is also indirectly encoded in our complexity measure. In addition to enforcing the orthogonality condition $(\star)$ (in the SGD variant), conditions for entering our regime are that the overparameterization $m_1, m_2, m_3$ and $\kappa_1, \kappa_2$ are within a specific range with respect to each other. We listed these relations in Section 3.6.3.

To illustrate the benefit of going to this more involved regime, denote the class of neural networks with bounded Frobenius norms $\|W'\| \leq \gamma_1, \|V'\| \leq \gamma_2$ by $\mathcal{G}_{\gamma_1,\gamma_2}$ (and a bit more structure which we elaborate upon in the proofs); it turns out that $\mathcal{G}_{\gamma_1,\gamma_2}$ roughly includes $\mathcal{H}_K(O(\gamma_1\gamma_2))$ for every kernel $K \in \mathcal{K}$, in the sense that each $f \in \mathcal{H}_K(O(\gamma_1\gamma_2))$ is well-approximated within $G_{\gamma_1,\gamma_2}$ to arbitrarily small error on fixed input (the error goes down with the size of the network). On the other hand, we show that the Rademacher Complexity (RC) of $\mathcal{G}_{\gamma_1,\gamma_2}$ behaves similar to the RC of the NTK class $\mathcal{H}_{H^\infty}(O(\gamma_1\gamma_2))$! As our algorithm guarantees finding a network with sufficiently small empirical risk within $\mathcal{G}_{\gamma_1,\gamma_2}$, this phenomenon underlies our adaptive generalization bounds.

Compared to previous work that provides an adaptive kernel analysis still for a two layer model [Dou and Liang, 2020] (although their analysis is for the gradient flow and non-algorithmic), our model requires an additional layer so it can, in a sense, "simulate" the process of feature extraction in one layer to be used in the next layer.

### 3.3.4 Comparison with Kernel fitting

We compare our generalization bounds with some kernel fitting rates. Given a kernel $K$ with $K(x,x) \leq 1$ for every $x : \|x\| \leq 1$, suppose we want to fit a function from $\mathcal{H}_K(B')$, i.e. having $K$-RKHS norm bounded by $B'$. In the realizable setting, when there is an underlying $f^{**} \in \mathcal{H}_K(B')$ with zero risk, Empirical Risk Minimization (ERM) enjoys a fast rate using the smoothness of the loss [Srebro et al., 2010]. The Rademacher Complexity bound $\mathfrak{R}(\mathcal{H}_K(B')) \leq O(\frac{B'}{\sqrt{n}})$ then implies

$$R(f^{\mathrm{ERM}}) \leq \tilde{O}(B'^2/n) \tag{3.19}$$

for the squared loss, which is minimax optimal up to log factors. To compare to the neural network, we substitute $f^{**}$ into Theorem 2. To relate the $B$ in our bound to $B'$, assume for simplicity of exposition that $\|f\|_\zeta = \|f\|_{K^*}$ for some $K^* \in \mathcal{K}$ (otherwise we can use a convergent sequence). Observing that $K_{\{g\}}(x,x) \leq 1$ for every kernel $K_{\{g\}} \in \mathcal{K}$, we obtain that $|f^{**}(x)| \leq \|f^{**}\|_{K^*} = \|f^{**}\|_\zeta$ (Section 3.6.6). Combining this fact with the realizability assumption, we can then upper bound the parameter $B$ in Theorem 2 by $\|f^{**}\|_\zeta$, and obtain

$$R(f_{W_{\mathrm{PSGD}},V_{\mathrm{PSGD}}}) = \tilde{O}(\|f^{**}\|_\zeta{}^2/n). \tag{3.20}$$

If we further take $K$ to be in $\mathcal{K}$, then Equation (3.20) combined with $\|f^{**}\|_\zeta \leq \|f^{**}\|_K \leq B'$ implies:

$$R(f_{W_{\text{PSGD}},V_{\text{PSGD}}}) = \tilde{O}({B'}^2/n),$$

that is, for every kernel $K \in \mathcal{K}$, our deep learning approach almost achieves the conventional kernel bound in Equation (3.19).

Repeating the uniform risk bound stated in Theorem 1 in [Srebro et al., 2010] for $\mathcal{H}_K(B')$ where $B'$ is set to all powers of two, followed by a union bound, one can easily obtain a fast rate of

$$R(f^{KRLR}) \leq \tilde{O}\Big(\frac{y^T K^{-1} y}{n} + \frac{B^2}{n}\Big), \tag{3.21}$$

for the solution of KRLR in the general case (not realizable) for the squared loss. On the other hand, for a $B$-bounded Lipschitz loss, we instead get a slow rate for KRLR:

$$R(f^{KRLR}) \leq \tilde{O}\Big(\sqrt{\frac{y^T K^{-1} y}{n}} + \frac{B}{\sqrt{n}}\Big),$$

where $B$ is an a.s. bound on $|y|$ as before. This bound is similar to Arora et al. [2019a]. Note that our data dependent generalization bound in Theorem 6 already achieves the fast rate for KRLR in (3.21) for any $K \in \mathcal{K}$. Finally, in the non-realizable case, we still have the following fast rate for ERM regarding the hypothesis class $\mathcal{H}_K(B')$ [Srebro et al., 2010]:

$$R(f^{ERM}) \leq \tilde{O}(R(f^{**}) + \tfrac{B'^2 + B^2}{n}),$$

where now $f^{**} := \text{argmin}_{f \in \mathcal{H}_K(B')} R(f)$, while Theorem 2 also implies (again for every $K \in \mathcal{K}$):

$$R(f_{W_{\text{PSGD}},V_{\text{PSGD}}}) \leq \tilde{O}\big(R(f^{**}) + \tfrac{\|f^{**}\|_\zeta^2 + B^2}{n}\big) = \tilde{O}\big(R(f^{**}) + \tfrac{B'^2 + B^2}{n}\big).$$

## 3.4 Algorithm: Projected Stochastic Gradient Descent

In this section, we describe our algorithm `PSGD`, presented as pseudocode in Figure 2, which is roughly Stochastic Gradient Descent modified to project out a low-dimensional random subspace from the second-layer weights. `PSGD` approximately runs SGD on a smoothed version of the following loss function ($\psi_1, \psi_2$ are defined in Section 3.6.2)

$$L_1(W', V') = R_n(f_{W',V'}) + \psi_1\|W'\|^2 + \psi_2\|V'\|^2.$$

Compared to standard SGD, our algorithm makes two modifications: (1) it uses randomized smoothing to alleviate the non-smoothness of the ReLUs, (2) it ensures that the weights in the second layer are orthogonal to the data features $\phi^{(0)}(x)$ computed by the first layer at initialization. This helps to control layer interactions as pointed out in Section 3.3.3. For smoothing, we add Gaussian smoothing matrices $W^\rho$ and $V^\rho$ to the weights with i.i.d. entries drawn from $\mathcal{N}(0, \beta_1^2/m_1)$ and $\mathcal{N}(0, \beta_2^2/m_2)$ respectively, for $\beta_2 = O_p((\kappa_1\sqrt{m_1})^{-1}(\kappa_2\sqrt{m_2})^{-2/3})$, $\beta_1 = O_p(m_3^2\kappa_2\sqrt{m_2}(\kappa_1\sqrt{m_1})^{-1})$. To simplify the exposition, $O_p(.)$ is hiding the dependencies on the basic parameters $B, n, 1/\lambda_0$ and log factors. Our convergence proof uses the loss with respect to this smoothed network.

For the projection, let $\Phi^\perp \subset \mathbb{R}^{m_2\times m_3}$ be the subspace of weights of the second layer whose rows are orthogonal to the first-layer data representations $\phi^{(0)}(x_i)$'s $\forall i \in [n]$ at initialization:

$$V' \in \Phi^\perp \leftrightarrow \forall j \in [m_2],\ \forall i \in [n]:\ V'_{j,}\phi^{(0)}(x_i) = 0. \tag{3.22}$$

In summary, at point $(W', V')$, the algorithm samples a random $(x_i, y_i)$ from the data, as well as smoothing matrices $W^{\rho,1}, V^{\rho,1}, W^{\rho,2}, V^{\rho,2}$. It then computes an unbiased estimate for the gradient $(\hat{\nabla}_W, \hat{\nabla}_V)$, adds additional normalized Gaussian noise matrices $\Xi_1, \Xi_2$ and moves in this direction with step size $\eta = 1/\text{poly}(n, B \vee 1/B, 1/\lambda_0)$:

$$(W', V') \leftarrow (W', V') + \eta\Big(\hat{\nabla}_W + \Xi_1/(\sqrt{m}_1\|\Xi_1\|),\ \mathtt{Proj}_{\Phi^\perp}(\hat{\nabla}_V + \Xi_2/\|\Xi_2\|)\Big). \tag{3.23}$$

**Parameters.** Our results apply to the overparameterized regime, when the size of the network, i.e. parameters $m_1, m_2, m_3$ are polynomially large in $n, B \vee 1/B, 1/\lambda_0$. This guarantees that the network has suitable function representation capacity, and PSGD is able to find a good local direction at every iteration. The regularization coefficients $\psi_1, \psi_2$ can be set with respect to any candidate $(f^*, G)$ for our complexity measure (3.9). In Section 3.6.2, we introduce a simple doubling trick that handles the case when we do not have access to an optimal candidate solution. With such an $f^*$, as we describe in Remark 1, define $\nu := \max\{R_n(\bar{f}^*)/2, B^2/n\}$, and set $\psi_1 = \nu/4$, $\psi_2 = \nu/(4\zeta(\bar{f}^*, G))$, where $\bar{f}^*$ is the projection of $f^*$ along the span of eigenvectors of $A$ with eigenvalue as large as $\Omega(1/n^2)$. We list the suitable regime for overparameterization in Section 3.6.3.

**Algorithm 2** PSGD(Projected Stochastic Gradient Descent)

```
Input: network architecture $m_1, m_2, m_3$, initialization parameters $\kappa_1, \kappa_2$, smoothing parameters $\beta_1, \beta_2$, training set $(x_i, y_i)_{i=1}^n$, label parameter $B$, $(f^*, G)$ from the complexity measure
Gaussian initialization $W_{j,k}^{(0)} \leftarrow \mathcal{N}(0, \kappa_1)$, $V_{j,k}^{(0)} \leftarrow \mathcal{N}(0, \kappa_2)$
Define parameters $\psi_1, \psi_2$, $\nu$, $\eta$, subspace $\Phi^\perp$, and objective $L_1$ as described in Section 3.4
while $L_1(W', V') > R_n(f^*) + 2\nu$ do
    Gaussian matrices $W_{j,k}^{\rho,1}, W_{j,k}^{\rho,2} \leftarrow \mathcal{N}(0, \frac{\beta_1^2}{m_1}), V_{j,k}^{\rho,1}, V_{j,k}^{\rho,2} \leftarrow N(0, \frac{\beta_2^2}{m_2})$
    Sample data $(x_i, y_i)$ uniformly at random
    Compute gradient estimates
    $\begin{cases} \hat{\nabla}_W = \dot{\ell}(f_{W'+W^{\rho,1}, V'+V^{\rho,1}}(x_i), y_i)\nabla_W f_{W'+W^{\rho,2}, V'+V^{\rho,2}}(x_i) + 2\psi_1 W', \\ \hat{\nabla}_V = \dot{\ell}(f_{W'+W^{\rho,1}, V'+V^{\rho,1}}(x_i), y_i)\nabla_V f_{W'+W^{\rho,2}, V'+V^{\rho,2}}(x_i) + 2\psi_2 V' \end{cases}$
    Move as $(W', V') \leftarrow (W', V') + \eta\Big(\hat{\nabla}_W + \Xi_1/(\sqrt{m}_1\|\Xi_1\|),\ \mathtt{Proj}_{\Phi^\perp}(\hat{\nabla}_V + \Xi_2/\|\Xi_2\|)\Big)$
Return $(W', V')$
```

## 3.5 High Level Idea of the PSGD Analysis

The reason for considering a Frobenius norm regularizer in `PSGD` is that we want the weights to remain close to their initialization so the final network is in the class $\mathcal{G}_{\gamma_1,\gamma_2}$ for suitably chosen $\gamma_1, \gamma_2$; while still reducing the nonconvex empirical loss $R_n(f_{W',V'})$. We prove convergence for `PSGD` by building on ideas from Allen-Zhu et al. [2018], with a framework based on the classic result that SGD can escape saddle points for nonconvex functions. Compared to them, we take a different approach driven by our purely data-dependent complexity measure. We augment this by a careful Rademacher complexity analysis of the class $\mathcal{G}_{\gamma_1,\gamma_2}$ in Section 3.6.11.

**Construction of a good Network** To study the loss landscape, similar to [Allen-Zhu et al., 2018], we show the existence of a good local update at reasonable points $(W', V')$, using the ideal pair $(W^*, V^*)$ that we carefully construct from our complexity measure. Here, we sketch our proof for constructing $(W^*, V^*)$. Let $(W', V')$ be the current weights of the algorithm. Fix a sample $i \in [n]$. In Section 3.6.12, we use $G$ to construct $W^*$ for the first layer weights with decomposition $W^* = \sum_{k=1}^{m_3} W_k^*$ and $O(1)$ bounded norm, such that $\phi^*(x_i)_k \coloneqq \frac{1}{\sqrt{m_1}} W^s \mathrm{Sgn}\Big((W^{(0)} + W')x_i\Big) W^* x_i$. This decomposition ensures for every $k, k' \in [m_3]$, negating $W_k^*$ only negates $\phi^*(x_i)_{k'}$ when $k' = k$ and has no effect on $\phi^*(x_i)_{k'}$ for $k' \neq k$. This way, we can easily generate any arbitrary sign flip of the entries of $\phi^*(x_i)$. We use this property to generate a suitable random descent direction.

Next, we construct a suitable weight matrix $V^*$ for the second layer which maps the features $\phi^*(x_i)$ into $f_i^*$ (recall the definition of the complexity measure). The key here is that we consider a regime where the norm of $\phi^{(0)}(x_i)$ is typically larger than that of $\phi'(x_i)$ and $\phi^*(x_i)$, so it is very likely that the sign pattern in the second layer is determined by $\phi^{(0)}(x_i)$ in most rows. In such a scenario, the condition $V_j' \perp \phi^{(0)}(x_i)$ becomes vital as the interaction of $V'$ with $\phi^{(0)}(x_i)$ is problematic for both generalization and optimization. From the standpoint of generalization, without excluding this interaction, one can exploit the large size of $\phi^{(0)}(x_i)$ and build a network within the class $\mathcal{G}_{\gamma_1,\gamma_2}$ corresponding to a complex function that overfits the data.

Indeed, we utilize the large magnitude of $\phi^{(0)}$ and its orthogonality to the rows of $V'$ in the RC bound. On the other hand, since the weights of the first layer does not affect $\phi^{(0)}(x_i)$, the interaction of $V'$ and $\phi^{(0)}(x_i)$ is problematic for the algorithm's convergence, particularly in proving the existence of a local descent direction. This is the rationale behind our orthogonality constraint (3.22).

Finally, the $\phi^*(x_i)$'s, the above control on the signs, and the fact that $\langle \phi^{(0)}(x_{i_1}), \phi^{(0)}(x_{i_2}) \rangle$ concentrates around $m_3 \mathbb{E}_{w \sim N(0, \kappa_1 I)}[\sigma(w^T x_{i_1}) \sigma(w^T x_{i_2})]$ which recovers the structure of the kernel $K^\infty$ (Section 3.2), give rise to the kernel $G \odot K^\infty$ in the second layer. Using this structure, we construct $V^*$ that maps $\phi^*(x_i)$'s to $f_i^*$'s, which has additional good properties, including $O({f^*}^T (G \odot K^\infty)^{-1} f^*)$-bounded norm, and rows that are orthogonal to $\phi^{(0)}(x_i)$'s. For more details, see Section 3.6.12.

**Nonexistence of Bad Saddle Points** Next, we want to exploit $(W^*, V^*)$ to prove the existence of a good direction along which the objective decreases locally. Moving along $(W^*, V^*)$ is the first idea, which fails as the cross terms created between $W', V^*$ and $V', W^*$ cannot be bounded effectively. Instead, we randomly perturb $W^*$ and $V^*$ in a coupled way and prove a reduction in expectation. We elaborate more on this suitable random direction. Multiplying random signs $\Sigma_k$ onto $W_k^*$, we define the sum $W_\Sigma^* = \sum_{k=1}^{m_3} \Sigma_k W_k^*$. We also multiply the same signs to the columns of $V^*$ and project it back onto $\Phi^\perp$ to obtain $V_\Sigma^*$. Then, we move in the random direction $(\sqrt{\eta} W_\Sigma^* - \eta W/2, \sqrt{\eta} V_\Sigma^* - \eta V/2)$; this update creates additional cross terms in the objective that we must bound to prove a local reduction argument. A key point here is that we prove with high probability the norm of the weights is always bounded. This norm restriction enables us to substitute terms that we do not have control over by their worst-case supremum. We refer to Section 3.6.13 for similar techniques.

**Convergence of `PSGD`** Finally, we use the fact that SGD escapes good saddle points [Ge et al., 2015b]. For proving the existence of a good random direction to escape saddle points above, we use that the norm of weights is uniformly bounded along all iterations; this bound, in fact, is looser than the bound that we show for the final weights of the network. Yet, this additional restriction cannot be addressed by the classical nonconvex theory of SGD. Consequently, we refine and adapt the proof of [Ge et al., 2015b] to incorporate this additional constraint. At a high level, Ge et al. [2015b] work with a supermartingale based on the loss value. To guarantee the additional norm restriction, it is initially tempting to apply Azuma-Hoeffding concentration to bound the upward deviations of this process. However, this fails as the process has a two-fold behavior, depending on how large the gradient is. At the core of our refinement proof here, we instead directly bound the MGF of the martingale using Doob's maximal inequality. We refer to Section 3.6.16 for more details.

## 3.6 Detailed proofs

The following contains different main sections of the proof. Lower-level lemmas may be found in Section 3.7.

### 3.6.1 Stronger Generalization bounds for polynomials

In this section, we prove an explicit generalization bound for functions represented as a polynomial sum. Note that the bounds in Arora et al. [2019a] for polynomials assume the monomials with degree larger than one to have even powers, while here we do not impose this restriction. In addition, different from Arora et al. [2019a], our bounds remain meaningful in the noisy case (recall our Theorem 2).

More specifically, we bound the $\zeta$ norm of such functions. Consider the target function $s$ with the following power series formula:

$$y = s(x) = \sum_{p=1}^{\infty} a_p (w_p^T x)^p, \tag{3.24}$$

where $a_p \in \mathbb{R}$ and $w_p \in \mathbb{R}^d$. We can write

$$s(x) = g_1(x) + \sum_{k=1}^{d} x_k g_2^k(x), \tag{3.25}$$

where $x_k$ denotes the $k$th entry of vector $x$ here and

$$g_1(x) = \sum_{p \in A_1 := \{p=1 \text{ or } p \text{ even}\}} a_p (w_p^T x)^p,$$

and for all $k \in [d]$:

$$g_2^k(x) = \sum_{p \in A_2 := \{p>2,\ p \text{ odd}\}} w_{p_k} a_p (w_p^T x)^{p-1}.$$

Then, using the Taylor series of $x(\frac{1}{4} + \frac{\arcsin(x)}{2\pi}) = \sum_{p=1}^{\infty} \gamma_p x^p$ for $|x| \leq 1$, the RKHS $\mathcal{H}(H^\infty)$ of the NTK can be identified by square-summable sequences of reals $(a_{p'})_{p'=1}^{\infty}$ with dot product

$$\langle (a_{p'})_{p'=1}^{\infty}, (b_{p'})_{p'=1}^{\infty} \rangle = \sum_{p'=1}^{\infty} \gamma_{\lambda(p')} a_{p'} b_{p'},$$

where $\lambda(p') : \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}$ such that it maps zero to zero, the first $d$ positive integers are mapped to one, the next $d^2$ ones are mapped to 2, etc. Moreover, the RKHS mapping $\Psi : \mathbb{R}^d \to \mathcal{H}(H^\infty)$ from the Euclidean space is:

$$\Psi(x) = \|x\| \Big( x'_1, ..., x'_d, (x'_{k_1} x'_{k_2})_{k_1,k_2 \in [d]}, \dots, (x'_{k_1} x'_{k_2} \dots x'_{k_p})_{k_1,\dots,k_p \in [d]}, \dots \Big),$$

where $x' = x/\|x\|$ and in the notation above we are presenting a sequence of sequences, by which we mean the inner sequences simply unfold. Using this identification, one can see using the linear representations of $g_1, g_2^k$ in $\mathcal{H}$:

$$\|g_1\|_{H^\infty}^2 = \sum_{p \in A_1} \gamma_p a_p^2 \|w_p\|_2^{2p}, \tag{3.26}$$

$$\|g_2^k\|_{H^\infty}^2 = \sum_{p \in A_2} \gamma_p {w_p}_k^2 a_p^2 \|w_p\|_2^{2(p-1)}. \tag{3.27}$$

Summing above and noting the linear representation of $g$:

$$\|g_1\|_{H^\infty}^2 + \sum_{k=1}^{d} \|g_2^k\|_{H^\infty}^2 = \sum_{p \in A_1} \gamma_p a_p^2 \|w_p\|_2^{2p} + \sum_{p \in A_2} \gamma_p a_p^2 \|w_p\|_2^{2p} = \sum_{p=1}^{\infty} \gamma_p a_p^2 \|w_p\|_2^{2p} = \|g\|_{H^\infty}^2. \tag{3.28}$$

Now for $\{g\} \coloneqq \{g'_k\}_{k=1}^{d+1} \coloneqq \{g_1\} \cup \{g_2^k\}_{k=1}^{d}$, we consider the kernel $K_{\{g\}}$. Expanding the Tailor series of $F_2(2F_3)(x) = \sum_{p=0}^{\infty} \mu_p x^p$, we find the identification $(h_{p'}^k)_{k \in [d+1], p'=0,\dots,\infty}$ with dot product

$$\sum_{p'=0}^{\infty} \mu_{\lambda(p')} \sum_{k=1}^{d+1} h_{p'}^k q_{p'}^k,$$

with RKHS map

$$\Psi_2(x) = \Big( g'_1(x), \dots, g'_{d+1}(x), x'_1 g'_1(x), \dots, x'_1 g'_{d+1}(x), \dots, x'_d g'_1(x), \dots,$$

$$x'_d g'_{d+1}(x), (x'_{k_1} x'_{k_2} g'_k(x))_{k_1,k_2 \in [d], k \in [d+1]}, \dots, (x'_{k_1} x'_{k_2} \dots x'_{k_p} g'_k(x))_{k_1,\dots,k_p \in [d], k \in [d+1]}, \dots \Big).$$

Now, we compute the norm of function $s$ with respect to $K_{\{g\}}$, combining the above representation and dot product with Equation (3.25) and the fact that we work with unit norm $x$, so $x' = x$:

$$\|s\|_{K_{\{g\}}}^2 = \mu_0 + (d+1)\mu_1. \tag{3.29}$$

Plugging the above and Equation (3.28) into the definition of $\|.\|_\zeta$ in (3.16), we conclude

$$\|s\|_\zeta^2 \leq \|s\|_{K_{\{g\}}}^2 (\|g_1\|_{H^\infty}^2 + \sum_{k=1}^{d} \|g_2^k\|_{H^\infty}^2) \leq (\mu_0 + (d+1)\mu_1) \sum_{p=1}^{\infty} \gamma_p a_p^2 \|w_p\|_2^{2p}. \tag{3.30}$$

Note that if the odd exponents (except possibly one) in the definition of $s$ in (3.24) are zero, then we could consider only the function $g_1$ and kernel $K_{g_1}$, which would have implied a bound of $\mu_0 \sum_{p=1}^{\infty} \gamma_p a_p^2 \|w_p\|_2^{2p}$.

### 3.6.2 The Doubling Trick

For the SGD optimization, we set the regularization coefficients in the loss $L_1$ as

$$\psi_1 = \nu/4, \quad \psi_2 = \nu/(4\zeta(f^*, G)), \tag{3.31}$$

with $\nu := \max\{R_n(f^*)/2, B^2/n\}$. This assumes we know the $f^*$ and $G$ that minimize the adaptation within the complexity measure (3.12). To achieve generalization bound in Theorem 6, here, we explain how to use a simple doubling trick to get over the fact that we might not know these optimal solutions $f^*$ and $G$. The proof here is based on the generalization result in Theorem 3.

**Theorem 1.** *Without explicitly knowing the exact value of the complexity measure, i.e., the optimal solution of Equation (3.12), one can still achieve the generalization bound in Theorem 6.*

**Proof of Theorem 1**

Our core generalization result is in Theorem 3. The proof of Theorem 1 is simply adding a doubling trick on top of the argument of Theorem 3. We also prove

Theorem 2 as a consequence of Theorem 1 in Section 2. In the rest of the proofs, for simplicity, we refer to $(W_{\text{PSGD}}, V_{\text{PSGD}})$ by $(W', V')$. Let $f^{**}, G^*$ be the optimal solution to (3.10). With a simple rescaling of $G^*$, we can assume $\langle H^{\infty^{-1}}, G^* \rangle = 1$. (Note that the complexity does not change by such rescaling). Now one can exploit the condition $\|y_i\|_\infty \leq B$, and consider the setting $f^* = 0$ to get the following trivial upper bound on the complexity measure:

$$\Im((x_i), (y_i)) \leq 2nB^2.$$

Therefore,

$$2nR_n(f^{**}) + f^{**}A^{-1}f^{**}(c'\varpi) \leq 2nB^2. \tag{3.32}$$

Using Equation (3.32) and the optimality of $(f^{**}, G^*)$:

$$\begin{aligned} R_n(f^{**}) &\leq B^2, \\ \zeta = \zeta(f^{**}, G^*) &\leq 2nB^2. \end{aligned} \tag{3.33}$$

Combining the first equation above with the definition of $\nu$ in Equation (3.41), we get

$$B^2/n \leq \nu \leq B^2. \tag{3.34}$$

To initialize $\psi_1$ and $\psi_2$, we use Equations (3.31) for any $f^*$ and $G$, and as a result we get a generalization bound as in Equation (3.45). However, to achieve the best possible rate characterized by our complexity measure in Theorem 6 without explicitly computing the answer of (3.12), we use a simple doubling trick; for every pair $(\zeta', \nu')$ such that $\zeta'$ is a power of 2 between $B^2$ and $2nB^2$, and $\nu'$ is a power of two between $B^2/n$ and $B^2$, we initialize $\psi_1, \psi_2$ as in Equation (3.31) and run the algorithm, then return the network which minimizes the empirical loss after the required polynomial number of steps. This is to make sure that the value of the loss will go at some point below the PSGDstopping threshold on the loss, since the stopping threshold depends on $R_n(f^{**})$ which we are not aware of. Another way to resolve this issue, to have an early stop when the value of the loss pass the threshold is to again run a doubling trick on the value of $R_n(f^{**})$ for every fixed value of $\nu$ and $\zeta$, and run PSGD with stop threshold $R_n(f^{**}) + 2\nu$ (here, $R_n(f^{**})$ is set using the doubling trick variable). This approach works because our final upper bound on the risk ignores the constants (note that the doubling trick introduce additional constants). Moreover, since $\nu \geq B^2/n$ by definition, we don't need to run $R_n(f^{**})$ over values smaller than $\Omega(B^2/n)$, since it does not change the order of $R_n(f^{**}) + 2\nu$. Particularly, combining this with the upper bound on $R_n(f^{**})$, we only need to run the doubling trick for $R_n(f^{**})$ in the interval $(\Omega(B^2/n), O(B^2))$. Now let $\nu'$ be the power of 2 within $\nu(G^*, f^{**})/2 \leq \nu' < \nu(G^*, f^{**})$. If we are in the case

$$f^{**T} A^{-1} f^{**} < B^2, \tag{3.35}$$

then for $\zeta'$ equal to the smallest power of two larger than $B^2$, when we run PSGD with pair $(\nu', \zeta')$, by Theorem 3:

$$R(f_{W',V'}) \leq 2R_n(f^*) + c''\varpi \frac{2B^2 + B^2}{n} \leq \frac{\Im((x_i)_{i=1}^n, (y_i)_{i=1}^n)}{n} + c' \frac{B^2 \varpi}{n}. \tag{3.36}$$

Because we return the minimum upper bound on the risk (the tighter lower bound of Equation (3.44)) among all such powers of two, we certainly achieve the above rate in (3.36). Otherwise, if $f^{**T} A^{-1} f^{**} \geq B^2$, let $\zeta'$ be the power of two within $f^{*T} A^{-1} f^* \leq \zeta' \leq 2 f^{*T} A^{-1} f^*$, then again it is easy to check that conditions of Theorem 3 are satisfied, hence we get the following generalization bound:

$$\begin{aligned} R(f_{W',V'}) &\leq 2R_n(f^{**}) + c''\varpi \frac{(\zeta' + B^2)}{n} \leq 2R_n(f^{**}) + c''\varpi \frac{2\zeta(f^{**}, G^*) + B^2}{n} && (3.37) \\ &\leq \frac{\Im((x_i)_{i=1}^n, (y_i)_{i=1}^n)}{n} + c' \frac{B^2 \varpi}{n}, && (3.38) \end{aligned}$$

which proves the bound of Theorem 6.

### 3.6.3 Amount of Overparameterization

In this section, to provide high-level insight, we indicate the right order of magnitude that our overparameterization should be in, with respect to one another. Note that the exact coefficients in these inequalities would depend on the basic parameters $B, 1/\lambda_0, n$, which we have avoided here for sake of simplicity. We refer the reader to our main proof (mostly Sections 3.6.12, 3.6.13) for more details.

$$\begin{aligned}
&\kappa_1\kappa_2 m_3 << 1,\\
&\kappa_2\sqrt{m_2} >> 1,\\
&\kappa_1\sqrt{m_3} >> \kappa_2\sqrt{m_2},\\
&m_1 >> m_3^4,\\
&\kappa_1\sqrt{m_1} >> m_3^{3/2},\\
&\kappa_2 << 1/\sqrt{m_3}\\
&\sqrt{m}_3\kappa_2 << 1/\sqrt{m}_3\\
&\sqrt{m}_2 >> m_3^{3/2}\kappa_1\kappa_2\\
&m_3^3(\kappa_2 m_2) << \kappa_1 m_1\\
&m_1, m_2, m_3, 1/\kappa_1, 1/\kappa_2 = \mathrm{poly}(n, B \vee 1/B, 1/\lambda_0).
\end{aligned}$$

In addition, we set the smoothing parameters as

$$\begin{aligned}
\beta_2 &:= \Theta_p\Big((\kappa_1\sqrt{m}_3)^{-1}(\sqrt{m}_2\kappa_2)^{-\frac{2}{3}}\Big),\\
\beta_1 &:= \Theta_p\Big(m_3\sqrt{m}_3/(\kappa_1\sqrt{m}_1)\Big),
\end{aligned}$$

where $\Theta_p$ only shows polynomial dependencies on the overparameterization.

### 3.6.4 PSD property of $K^\infty$

The Schur product theorem states that for PSD matrices $A$ and $B$, $A \odot B$ is also PSD. Now given an analytic function $F$ whose Tailor series coefficients are all nonnegative, Suppose we apply $F$ on some PSD matrix $A$ entrywise, denoted by $F(A)$, under the condition that the entries of $A$ are in the radius of convergence of $F$, then using Schur product theorem, it is straightforward that $F(A)$ is also PSD.

Using the above property, one can then check that the Tailor series of the defined functions $F_2$ and $F_3$ are nonnegative, hence, the application of the function $F_2(2F_3(x))$ on the gram matrix of $(x_i)_{i=1}^n$ is a PSD matrix, ( note that $\Big|\langle x_i, x_j\rangle\Big| \leq 1$ is in the convergence radius of $F_2(2F_3(x))$.) thus $K^\infty$ is indeed a kernel.

### 3.6.5 Complexity upper bound

First we mention a simple fact that hadamard product respects matrix orderings. Given PSD matrices $A, B, C$ such that $A \preceq B$, the fact that $A \odot C \preceq B \odot C$ is an easy consequence of the Schur Product Theorem; indeed, $B - A$ is PSD by definition, so $(B - A) \odot C = B \odot C - A \odot C$ is also PSD.

Next, it is easy to check that the Tailor series of $\arcsin(x)$ has all nonnegative coefficients. Therefore, for a PSD matrix $X$, as we discussed in Section 3.6.4, applying arcsin entrywise on $X$, namely $\arcsin X$, is also PSD. Setting $X$ equal to the entrywise application of $2F_3$ to the gram matrix of datapoints $(x_i)_{i=1}^n$, we realize the matrix $\arcsin\Big(2F_3\Big(\Big(\langle x_i, x_j\rangle\Big)_{1\leq i,j\leq n}\Big)\Big)$ is also PSD. Noting the definition of $K^\infty$ in Equation (3.8), we conclude that for the data kernel matrix $K^\infty$ we have

$$K^\infty \geq \frac{1}{4}\mathbb{1}\mathbb{1}^T,$$

where $\mathbb{1}$ is the all ones $n$-dimensional vector.

Combining the two mentioned facts, we can lower bound the matrix $K = K^\infty \odot G$ for any matrix $G$ as

$$K = K^\infty \odot G \geq \frac{1}{4}\mathbb{1}\mathbb{1}^T \odot G = \frac{1}{4}G.$$

Substituting the rank one matrix $f^* {f^*}^T$ for the $n$-dimensional vector $f^*$ in Equation (3.13):

$$K^{-1}_{\{\tilde{f}^*/\|\tilde{f}^*\|\}} = K^\infty \odot f^* f^{*T} / \|\tilde{f}^*\|^2 \geq \frac{1}{4} f^* f^{*T} / \|\tilde{f}^*\|^2. \tag{3.39}$$

The inequality used in (3.13) then follows from Equation (3.39).

### 3.6.6 Complexity measure and the $\zeta$-norm

This is a brief section regarding some basic properties of $\Im$ and $\|.\|_\zeta$.

First, note that the two versions of the complexity measure in Equations (3.10) and (3.11) are equivalent, as for any finite set of functions $\{g\}$, we can define the gram matrix with respect to the feature vectors of these functions on data, and for an arbitrary nonzero PSD $G$ we can consider a Cholesky factorization for $G$ as $G = \bar{X}^T \bar{X}$, then define the functions $\{g_k\}$ as the minimum-NTK norm functions which map the input to the features corresponding to $\bar{X}$. This observation further implies we can suppose the factor matrix $\bar{X}$ is in $\mathbb{R}^{n \times n}$, and there is a set of at most $n$ functions $\{g_k\}_{k=1}^n$ which corresponds to this $G$.

Next, we show that for an arbitrary function $f$, its sup norm over the unit ball is bounded by its $\zeta$ norm:

$$\sup_{\|x\|=1} |f(x)| \leq \|f\|_\zeta. \tag{3.40}$$

Note that for a kernel $K$ which satisfies $K(x,x) \leq 1$, using Cauchy Schwarz we simply obtain

$$f(x) \leq \|f\|_K \sqrt{K(x,x)} \leq \|f\|_K,$$

where recall that $\|.\|_K$ is the norm corresponding to the RKHS space of $K$. Hence, to show (3.40), it suffice to show that for all kernels $K \in \mathcal{K}$ and unit norm $x$ we have $K(x,x) \leq 1$. To see this fact, note that the norm of each $x \in \mathbb{R}^d$ in the NTK-space is $H^\infty(x,x) = \frac{1}{2}$. Therefore, for each function $g$ with bounded-NTK norm, again using

Cauchy Schwarz:

$$|g(x)| \leq \frac{1}{2}\|g\|_{H^\infty}.$$

As a result, for a family of functions $\{g\}$ with $\sum_{g\in\{g\}} \|g\|_{H^\infty}^2 \leq 1$, we have on every unit norm $x$:

$$\sum_{g\in\{g\}} g(x)^2 \leq 1.$$

On the other hand, it is easy to check that for every unit norm $x$, we have $K^\infty(x,x) \leq \frac{1}{2}$, so for every such $\{g\}$, we have by definition

$$K_{\{g\}}(x) \leq 1,$$

which completes the proof of Equation (3.40).

### 3.6.7 Core Generalization Result

In this section, we prove our core generalization result for the trained network, Theorem 3, which underlies our generalization bounds in Theorems 6 and 2. Recall that in the rest of the proofs, we refer to the solution $(W_{\mathrm{PSGD}}, V_{\mathrm{PSGD}})$ returned by PSGD simply by $(W', V')$.

**Theorem 3.** *Suppose we have a good candidate pair $(f^*, G)$ regarding our complexity measure in (3.10) that satisfies $\langle H^{\infty-1}, G\rangle \leq 1$, $f^{*T} A^{-1} f^* \leq \zeta$ (recall $A = G \odot H^\infty$), and that $f^*$ has zero projection onto the directions of eigenvectors of $A$ whose eigenvalues are smaller than $O(1/n^2)$ (the last condition can be relaxed, see the next remark). Then, for*

$$\nu = \max\{R_n(f^*)/4,\ B^2/n\}, \tag{3.41}$$

*if we are given $\nu/2 \leq \nu' \leq \nu$, and we set*

$$\psi_1 = \frac{\nu'}{4}, \tag{3.42}$$

$$\psi_2 = \frac{\nu'}{4\zeta}, \tag{3.43}$$

*then for the solution $(W', V')$ returned by* `PSGD` *we have the following generalization bound:*

$$R(f_{W',V'}) \leq \frac{5}{4} R_n(f_{W',V'}) + c''' \varpi \frac{(\zeta + B^2)}{n} \tag{3.44}$$

$$\leq \frac{5}{3} R_n(f^*) + c'' \varpi \frac{(\zeta + B^2)}{n}, \tag{3.45}$$

*for constants $c'', c'''$ and log factor $\varpi = \log(n)^3 + \log(1/\lambda_0)$.*

**Remark 1.** *Given a pair $(f^*, G)$ satisfying $\langle H^{\infty-1}, G\rangle \leq 1$, $f^{*T} A^{-1} f^* \leq \zeta$, one can project out the directions that are along the eigenvectors of $A$ with eigenvalues smaller than $\Omega(1/n^2)$ to obtain $\bar{f}^*$, then use the pair $(\bar{f}^*, G)$ in Theorem 3. This way, the third condition mentioned in Theorem 3 also becomes true. As we show in Lemma 51, by* *switching $f^*$ to $\bar{f}^*$ the quantity ${f^*}^T A^{-1} f^*$ does not increase, and the quantity $R_n(f^*)$ is multiplied by a constant $c > 1$ arbitrarily close to one, then adds up with $O(B^2/n)$. This means that the bounds in Theorem 3 for the pair $(\bar{f}^*, G)$ translates into similar bounds for $(f^*, G)$ albeit with a bit worse contants. It is straightforward to see that with small enough choice of c and careful AM-GM inequality that we apply inthe proof of Theorem 3, one can end up with the same constants regarding the pair $(f^*, G)$ as declared in Theorem 3. For a more careful discussion on this, we refer the reader to the proof of Lemma 19.*

**Proof of Theorem 3**

Almost all of our proofs in the rest are in the aim of proving Theorem 3. Crucially, to prove this Theorem, we need to establish two big results:

1. We need to show that the final network has small training loss, and is within the class $\mathcal{G}_{\gamma_1,\gamma_2}$ for some suitable $\gamma_1, \gamma_2$. This is handled by Theorem 4 in Section 3.6.10. We define the class $\mathcal{G}_{\gamma_1,\gamma_2}$ roughly as the class of networks with norm bounds $\|W - W^{(0)}\| \leq \gamma_1, \|V - V^{(0)}\| \leq \gamma_2$ where the rows of $V - V^{(0)}$ are orthogonal to the subspace $\Phi$, plus an additional structure defined in Section 3.6.11. This task, on its own, has three main steps in our proof:
   (a) we construct a "good" underlying network, Section 3.6.12
   (b) we find a "good" random direction and study the landscape of the objective, Section 3.6.13
   (c) we prove the convergence, Section 3.6.16
2. The Rademacher Complexity of the class $\mathcal{G}_{\gamma_1,\gamma_2}$ needs to be suitably bounded. This is handled by Theorem 5 in Section 3.6.11.

With access to these results, here we show how Theorem 3 follows by a simple application of the generalization bound in [Srebro et al., 2010]. Specifically, for fixed constants $z_1, z_3$ and every integer $i \geq 0$, we use Theorem 1 of [Srebro et al., 2010] for the class $\mathcal{G}_{z_1,\gamma_i}$, $\gamma_i = 2^i \times B/z_3$ with confidence probability $1 - 2^{-i}\delta_3$, which, with a union bound, implies that with probability at least $1 - \delta_3$, for every $i$ and $f_{W',V'} \in \mathcal{G}_{z_1,\gamma_i}$:

$$R(f_{W',V'}) \leq R_n(f_{W',V'}) + K(\sqrt{R_n(f_{W',V'})}(\sqrt{\log(n)^{1.5}\mathcal{R}(\mathcal{G}_{z_1,\gamma_i}) + \sqrt{\frac{b\log(1/(2^{-i}\delta_3))}{n}}}) \tag{3.46}$$

$$+ \log(n)^3\mathcal{R}(\mathcal{G}_{z_1,\gamma_i})^2 + \frac{b\log(1/(2^{-i}\delta_3))}{n}), \tag{3.47}$$

where $\ell(f_{W',V'}(x), y)$ is a.s. bounded by $b$ for function within the class $\mathcal{G}_{z_1,\gamma_i}$, and $K$ is a universal constant. In the following, we aim to further bound the Rademacher complexity $\mathcal{R}$ and parameter $b$.

Applying the AM-GM inequality with respect to ratio $z_4 > 0$ for the second term:

$$\begin{aligned}R(f_{W',V'}) &\leq (1+z_4)R_n(f_{W',V'}) + K^2/z_4\Big(\log(n)^{1.5}\mathcal{R}(\mathcal{G}_{z_1,\gamma_i}) + \sqrt{\frac{b\log(1/(2^{-i}\delta_3))}{n}}\Big)^2\\&+ \log(n)^3\mathcal{R}(\mathcal{G}_{z_1,\gamma_i})^2 + \frac{b\log(1/(2^{-i}\delta_3))}{n}\\&\leq (1+z_4)R_n(f_{W',V'}) + K^2/z_4\Big(\log(n)^{1.5}\mathcal{R}(\mathcal{G}_{z_1,\gamma_i}) + \sqrt{\frac{b\log(1/(2^{-i}\delta_3))}{n}}\Big)^2\\&+ \log(n)^3\mathcal{R}(\mathcal{G}_{z_1,\gamma_i})^2 + \frac{b\log(1/(2^{-i}\delta_3))}{n}\\&\leq (1+z_4)R_n(f_{W',V'}) + (2K^2/z_4+1)\log(n)^3\mathcal{R}(\mathcal{G}_{z_1,\gamma_i})^2 + (2K^2/z_4+1)\frac{b\log(1/(2^{-i}\delta_3))}{n}.\end{aligned} \tag{3.48}$$

Now let $\gamma^*$ be the smallest number of the form $2^iB/z_3$ (for some $i$) which is not smaller than $z_2\sqrt{\zeta}$. This definition implies

$$\gamma^* \leq \max\{2z_2\sqrt{\zeta}, B/z_3\}. \tag{3.49}$$

Now Theorem 5 in Section 3.6.11 bounds the Rademacher complexity:

$$\mathcal{R}(\mathcal{G}_{z_1,\gamma^*}) \leq \frac{2z_1\gamma^*}{\sqrt{n}}. \tag{3.50}$$

On the other hand, from Theorem 4 by setting $z_1, z_2 = \sqrt{40}$, we get $f_{W',V'} \in \mathcal{G}_{z_1,\gamma^*}$.

Moreover, from Lemma 43, for $f_{W',V'} \in \mathcal{G}_{z_1,\gamma^*}$, we have for every $\|x\| \leq 1$:

$$|f_{W',V'}(x)| \leq 2z_1\gamma^*, \tag{3.51}$$

so the loss $\ell(f_{W',V'}(x), y)$ can be bounded by $(B + 2z_1\gamma^*)^2$ using the 1 smoothness property. Therefore, for the class $\mathcal{G}_{z_1,\gamma^*}$ we can set $b = (B + 2z_1\gamma^*)^2$. Combining this with Equation (3.50) and plugging into Equation (3.48):

$$\begin{aligned} R(f_{W',V'}) \leq (1 + z_4)R_n(f_{W',V'}) + (2K^2/z_4 + 1)\log(n)^3 \frac{4z_1^2\gamma^{*2}}{n} \\ + (2K^2/z_4 + 1)\frac{(B + 2z_1\gamma^*)^2 \log(1/(2^{-i^*}\delta_3))}{n}. \end{aligned}$$

Furthermore, by definition of $\gamma^*$, we have $2^i \leq 2z_2z_3\sqrt{\zeta}/B$:

$$R(f_{W',V'}) \leq (1 + z_4)R_n(f_{W',V'}) + (2K^2/z_4 + 1)4z_1^2(\log(n)^3 + 2\log(2z_2z_3\sqrt{\zeta}/B))\frac{4z_1^2\gamma^{*2}}{n} \tag{3.52}$$

$$+ (2K^2/z_4 + 1)\frac{2B^2\log(2z_2z_3\sqrt{\zeta}/B)}{n}. \tag{3.53}$$

Now applying the upper bound on $\gamma^*$:

$$R(f_{W',V'}) \leq (1 + z_4)R_n(f_{W',V'}) + (2K^2/z_4 + 1)4z_1^2(\log(n)^3 \tag{3.54}$$

$$+ 2\log(2z_2z_3\sqrt{\zeta}/B))\frac{4z_1^2(2z_2\sqrt{\zeta} + 2B/z_3)^2}{n} \tag{3.55}$$

$$+ (2K^2/z_4 + 1)\frac{2B^2\log(2z_2z_3\sqrt{\zeta}/B)}{n}. \tag{3.56}$$

If $\zeta > B^2$, in the third term above we substitute $B$ by $\sqrt{\zeta}$. Finally, similar to the bound we stated in Equation (3.14), note that we have the following trivial bound for $\zeta$:

$$\zeta \leq y^T H^{\infty -1} y \leq 4nB^2/\lambda_0, \tag{3.57}$$

i.e. there is no point in considering larger $\zeta$'s, which implies $\log(2z_2z_3\sqrt{\zeta}/B) =$

$O(\log(n) + \log(1/\lambda_0))$. Plugging this above and picking $z_4 = 1/3$ show the proof of Equation (3.44). Furthermore, applying Equation (3.69) in Theorem 4 to the $R_n(f_{W',V'})$ term in Equation (3.44) further gives the second Equation (3.45).

**Remark 2.** *In the same setting of Theorem 3, if we have $\nu/2 \leq \nu'$ but not generally upper bounded by $\nu$, then* `PSGD` *leads to the following generalization bound:*

$$R(f_{W',V'}) \leq R_n(f^*) + \nu' + c''''\varpi \frac{(\zeta + B^2)}{n},$$

*using a similar argument as we did for Theorem 3.*

### 3.6.8 Structure of the proof, setting $m_3$, and further definitions

Throughout the proof, $(W', V')$ represents the pair of matrices of the current iteration of `PSGD`, $(W^*, V^*)$ are the "ideal" matrices that we construct in Section 3.6.12, $(W^\rho, V^\rho)$ and refers to the gaussian smoothing matrices. Importantly, note that our squared loss $\ell(f, y)$ is **zero** at $f = y$. We have tried to make the lower level proofs into sub-lemmas and create a manageable hierarchy as much as we could, to make the document more clear and readable.

Similar to the conditions in Theorem 3, through out most of the proofs we assume that we are given a pair $(f^*, G)$ with a slightly more general setting of Theorem 3:

$$\begin{aligned}
&f^{*T} A^{-1} f^* \leq \zeta_2, \text{ for } A = G \odot K^\infty, \\
&\langle G,\ H^\infty \rangle \leq \zeta_1.
\end{aligned}$$

Particularly, $\zeta_1, \zeta_2$ appear in Section 3.6.13. Because we are allowed to rescale $G$, we do not really gain much by assuming this more general setting, though we pick to work with the general setting as the abstraction makes the proof more straightforward to understand.

We refer to the parameters $B, 1/\lambda_0, n$ as the "basic parameters", $m_1, m_2, m_3, 1/\kappa_1, \kappa_2$ as the "overparameterization", and $\beta_1, \beta_2$ as the "smoothing parameters." By the phrase "having enough overparameterization" we mean it suffices to pick the overparameterization $m_1, m_2, m_3, 1/\kappa_1, 1/\kappa_2$ only **polynomially** large in the basic parameters.

Throughout the proof, we denote the change in the output of the first layer at $W^{(0)} + W' + W^\rho$ compared to the initialization value by $\phi^{(2)}(x_i)$, i.e.

$$\phi^{(2)}(x_i) = \frac{1}{\sqrt{m}_1} W^s \sigma((W^{(0)} + W' + W^\rho)x_i) - \phi^{(0)}(x_i),$$

while recall that $\phi'(x_i)$ has a similar definition except without the smoothing matrix $W^\rho$. Although our model is a three layer network, throughout the proof, we refer to the parts $W^s \frac{1}{\sqrt{m}_1}\sigma((W^{(0)} + W')x)$ and $\frac{1}{\sqrt{m}_2} a^T \sigma\Big((V^{(0)} + V')(.)\Big)$ as the "first layer" and "second layer," respectively.

Also, we sometimes refer to the binary sign pattern of vector $x$ multiplied to matrix $W$ by $D_{W,x}$ ($D_{W,x} \coloneqq \mathrm{Sgn}(Wx)$), i.e. the $j$th diagonal entry of $D_{W,x}$ is one if $W_{j,}^T x \geq 0$, and is zero otherwise. To refer to the $j$th row of $W$ as a vector, we sometimes drop the comma in $W_{j,}$ and write it as $W_j$.

For brevity, we denote the Frobenius norm $\|W\|_F$ of matrix by $\|W\|$, and the Euclidean norm of a vector $x$ by $\|x\|$. For matrices $W_1, W_2$ we denote their natural dot product by $\langle W_1, W_2 \rangle \coloneqq \mathrm{tr}(W_1^T W_2)$. We refer to the smallest eigenvalue of a matrix by $\lambda_{\min}(.)$. We write $\mathcal{R}(.)$ for the Rademacher complexity of a function class. We refer to the smoothed version of the network by $f'_{W',V'}(x)$, defined by

$$f'_{W',V'}(x) = \mathbb{E}_{W^\rho, V^\rho} f_{W'+W^\rho, V'+V^\rho}(x).$$

In the proof, we mainly work with the loss over the smoothed network $f'$, defined as

$$L(W', V') = R_n(f'_{W',V'}) + \psi_1 \|W'\|^2 + \psi_2 \|V'\|^2. \tag{3.58}$$

Our algorithm, `PSGD` can be regarded roughly as an SGD over $L$.

Similar to what we discussed in section 3.3, let the functions $\{g_k\}_{k=1}^{m_3}$ be some feature representation whose gram matrix is equal to $G$ and $\langle H^\infty, G \rangle = \sum_{k=1}^{m_3} \|g_k\|_{H^\infty}^2$. In such setting, it is not hard to check that we can assume each $g_k$ is the minimum norm NTK function which maps $(x_i)_{i=1}^n$ to $(g_k(x_i))_{i=1}^n$'s. Indeed, if this is not the case for some $g_k$, we can project the RKHS representation of $g_k$ onto the span of the representations of $(x_i)_{i=1}^n$, which can only decrease the complexity measure. Hence we can represent $g_k \in \mathcal{H}_{H^\infty}$ as a linear combination of basic functions $H^\infty(x_i, .)$ on data points:

$$\forall k \in [m_3],\ g_k(x) := \sum_{i=1}^{n} \mathcal{V}_{k,i} H^\infty(x_i, x). \tag{3.59}$$

Here, the sum of squared-$H^\infty$ norms of $\mathcal{V}_k$ is bounded as

$$\sum_k \|\mathcal{V}_k\|_{H^\infty}^2 = \sum_k \|g_k\|_{H^\infty}^2 = \langle H^\infty, G \rangle \leq \zeta_1. \tag{3.60}$$

For each $i \in [n]$, we refer to the feature representation vector $(g_k(x_i))_{k=1}^{m_3}$ on $x_i$ as $\bar{x}_i$. Note that we have the relationship

$$\bar{x}_i = (H_{i,}^\infty \mathcal{V}_k)_{k=1}^{m_3}, \tag{3.61}$$

where $H_{i,}^\infty$ is the $i$th row of $H^\infty$. In the analysis, we work with a bound $\xi$ on the quantity $\max_k \|\mathcal{V}_k\|$ which should be bounded polynomially by other basic parameters; in particular, it is defined in Lemma 19 and is used to bound a cross term in Lemma 23. However, $\max_k \|\mathcal{V}_k\|$ might not be effectively bounded for an arbitrary feature representation. Fortunately, we can remedy this by a simple trick; for every natural number $s$, one can substitute every $g_k$ by $s$ copies of $g_k/\sqrt{s}$, without changing the gram matrix $G$. Therefore, for any $\delta$, one can increase the multiset of functions $(g_k)$ to a bigger set $(\tilde{g}_k)$, by adding at most $O(\zeta_1/\delta)$ functions, making sure of the following for the new functions:

$$\forall k:\ \mathcal{V}_k^T H^\infty \mathcal{V}_k = \|\tilde{g}_k\|_{H^\infty}^2 \leq \delta. \tag{3.62}$$

(This is because $\sum_k \|\tilde{g}_k\|_{H^\infty}^2 \leq 1$). Furthermore, observe that for each gram matrix $G$, we have an $n$-dimensional feature representation $(g_k)_{k=1}^n$ for $G$ according to the Cholesky factorization. Combining these facts, we conclude, to guarantee Equation (3.62), in the worst case, we need $m_3$ to be as large as $n + O(\zeta_1/\delta)$.

Finally, observing the following inequality

$$\|\mathcal{V}_k\|^2 \leq \|\mathcal{V}_k\|^2_{H^\infty}/\lambda_0 \leq \|\tilde{g}_k\|^2_{H^\infty}/\lambda_0. \tag{3.63}$$

in order to guarantee $\max_k \|\mathcal{V}_k\| \leq \xi$ we need to take $m_3$ as large as $n + O(\zeta_1/(\xi^2\lambda_0))$, which is indeed bounded polynomially by the basic parameters because of the same condition for $1/\xi$. This computation also brings into sight an important point:

"Although each gram matrix $G$ is representable by $n$ features, in order for the algorithm to be able to find a suitable network, $m_3$ might need to be larger than $n$."

Moreover, for every $1 \leq k, i \leq n$, we define the matrices $Z_i^k \in \mathbb{R}^{md}$ as

$$Z_k^i = 1/\sqrt{m_1}\Big(W^s_{k,j}\mathbb{1}\{W_j^{(0)T} x_i\} x_i\Big)_{j=1}^{m_1}, \tag{3.64}$$

where in the above notation, $j$ is enumerating the columns of the matrix. We also define the following matrices which we use in our construction later:

$$W_k^{+T} = W^{k+T} = \sum_{i=1}^{n} \mathcal{V}_{k,i} Z_k^i, \tag{3.65}$$

and $W^+$ as

$$W^+ = \sum_{k=1}^{m_3} W^{k+}.$$

Finally, to avoid unnecessary complication, we often argue **high probability** bounds without an explicit representation of their dependency on the chance of failure (which is a negligible logarithmic factor). We also ignore all constants and log factors, and mainly work with the notation $\lesssim$ which ignores constants; we write $a \lesssim b \pm c$ as a short form for $b - O(c) \leq a \leq b + O(c)$. As there are several hierarchies of new parameters that are defined based on lower-level ones, we rename the new parameters and continue viewing them as black-box. This makes the proofs more readable, since we also do not care about the exact dependency of the underlying parameters most of the time, rather we are interested in their orders of magnitude, for example that a given parameter goes to zero polynomially fast with respect to the overparameterization, etc. Due to the large number of symbols that we have to work with, we might use a symbol more than once, of course when it is clear from the context which one we are refering to.

### 3.6.9 Proof of Theorem 2

In this section, we prove Theorem 2, stated below.

**Theorem 2.** *For any function $f : \mathbb{R}^d \to \mathbb{R}$, in the same setting as Theorem 6, the population risk of the trained network $(W', V')$ can be bounded as*

$$R(f_{W',V'}) \leq 2R(f) + O\Big(\alpha\varpi \frac{\|f\|_\zeta^2 + B^2}{n}\Big). \tag{3.66}$$

**Proof of Theorem 2**

Theorem 2 is a simple consequence of Theorem 6; for the given function $f$, we apply Theorem 2 with the smaller coefficient $\gamma = \frac{4}{3}$ for $R_n(f^*)$, regarding the complexity upper bound, by setting $f^* \coloneqq (f(x_i))_{i=1}^n$:

$$\begin{aligned} R(f_{W',V'}) &\leq \frac{4}{3}R_n(f^*) + (\alpha\varpi) \min_{K\in\mathcal{K}} \frac{{f^*}^T K^{-1} f^*}{n} + \frac{B^2\alpha\varpi}{n} \\ &= \frac{4}{3}R_n(f) + (\alpha\varpi) \min_{K\in\mathcal{K}} \frac{{f^*}^T K^{-1} f^*}{n} + \frac{B^2\alpha\varpi}{n}. \end{aligned}$$

On the other hand, because ${f^*}^T K^{-1} f^*$ is the minimum-RKHS norm of a function with respect to kernel $K$ which maps $x_i$'s to ${f^*}_i$ and $f$ is one such function, we have ${f^*}^T K^{-1} f^* \leq \|f\|_K$. This inequality implies

$$\min_{K\in\mathcal{K}} {f^*}^T K^{-1} f^* \leq \|f\|_\zeta,$$

so we obtain

$$R(f_{W',V'}) \leq \frac{4}{3}R_n(f) + (\alpha\varpi)\frac{\|f\|_\zeta^2 + B^2}{n}. \tag{3.67}$$

Therefore, it remains to bound $R_n(f)$ by $R(f)$.

As we showed in Section 3.6.6, for every input $x$ we have $f(x) \leq \|f\|_\zeta$, so for every data $(x, y)$, by the fact that $|y| \leq B$ a.s. and $\alpha$ smoothness of the loss, we have $\ell(f(x), y) \leq \alpha(\|f\|_\zeta + B)^2$. Moreover, note that the random variable $\ell(f(x), y)$ has mean $R(f)$. It is easy to check that in this setting, the variance of $\ell(f(x), y)$ is at most $R(f)\alpha(\|f\|_\zeta + B)^2$. Therefore, an application of the Bernstein inequality, we have with high probability over the dataset

$$R_n(f) \leq R(f) + O\left(\sqrt{\frac{R(f)\alpha(\|f\|_\zeta + B)^2}{n}} + \frac{\alpha(\|f\|_\zeta + B)^2}{n}\right) \leq \frac{3}{2}R(f) + O\left(\frac{\alpha(\|f\|_\zeta + B)^2}{n}\right).$$

Plugging this back to Equation (3.67) completes the proof. As a result, the learned network can compete with any function that has reasonably small $\|f\|_\zeta$:

$$R(f_{W',V'}) \leq \min_f \left\{2R(f) + O(\alpha\varpi\frac{\|f\|_\zeta^2 + B^2}{n})\right\}.$$

### 3.6.10 Optimization

In this section, we glue together

- the existence of a good random direction that we prove in Section 3.6.13
- the convergence analysis of `PSGD` that we do based on the work Ge et al. [2015b] in Section 3.6.16.

**Theorem 4.** *In the same setting of Theorem 3, assume the network $(W', V')$ returned by **`PSGD`**, has sufficient polynomially large "overparameterization". Then, for the solution $(W', V')$ returned by **`PSGD`** we have*

$$L(W', V') \leq R_n(f^*) + \nu, \tag{3.68}$$

*which further implies*

$$R_n(f_{W',V'}) \leq R_n(f^*) + 2\nu, \tag{3.69}$$

$$\|W'\|^2 \leq 40, \ \|V'\|^2 \leq 40\zeta. \tag{3.70}$$

*Moreover, for every $i \in [n], j \in [m_1], j \notin P$ for $P$ defined in 10, we have that* $\mathrm{sign}((W_j^{(0)} + W_j')^T x_i)$ *and* $\mathrm{sign}({W_j^{(0)}}^T x_i)$ *are the same.*

**Proof of Theorem 4**

Let $\Upsilon \in \mathbb{R}^{m_2(m_3-n)\times m_2 m_3}$ be a matrix whose rows are an orthonormal basis for the space of matrices whose rows are orthogonal to $\mathrm{span}(\{\phi^{(0)}(x_i)\}_{i=1}^n)$, i.e. $\Phi^\perp$, as defined in (3.22). Then, we consider a linear change of coordinates for the subspace $\Phi^\perp$, regarding the second layer weights, as $v' = \Upsilon \mathrm{vec}(V')$ where vec(.) splits out the vectorized version of a matrix. For consistent notation, we also denote $w' = W'$, so we now have a new coordinate system $(w', v') \in \mathbb{R}^{m_2(m_3-n)\times m_1 d}$ for pairs of weights $(W', V')$ such that $V' \in \Phi^\perp$. We also define the loss function

$$L^\Pi(w := (w', v')) = L(W', V'),$$

with respect to the change of coordinate.

Now it is easy to see that running PSGD on $L$ in the normal coordinates is equivalent to running stochastic gradient descent on $L^{\Pi}$ with respect to $(w', v')$. Moreover, because multiplying to matrix $\Upsilon$ is an orthonormal change of coordinates for $\Phi^{\perp}$ and because $V'$ is already in $\phi^{\perp}$ at each step of PSGD, then $\|v'\| = \|V'\|$, so the conditions $\|W'\| \leq C_1, \|V'\| \leq C_2$ are equivalent to $\|w'\| \leq C_1, \|v'\| \leq C_2$. Furthermore, by our construction, the random matrix $V^*_{\Sigma}$ is in the subspace $\Phi^{\perp}$, so the norm bounds $\|W^*\| \leq \zeta_1, \|V^*\| \leq \zeta_2$ are equivalent to $\|w^*\| \leq \zeta_1, \|v^*\| \leq \zeta_2$ for $w^* := W^*_{\Sigma}$ and $v^* := \Upsilon \text{vec}(V^*)$.

Now we apply the result of Theorem 6 on $L^{\Pi}$ with parameter $\nu$ set as $\nu'$ (recall the definition of $\nu'$ from Theorem 3), $\zeta_2 := \zeta$ and $\zeta_1 := 1$, and $\Delta := R_n(f^*)$, as defined in Theorem 3. More specifically, based on our arguments above regarding the natural isometry in the change of coordinate, any pair $(w', v')$ in the domain $\|w'\| \leq C_1, \|v'\| \leq C_2$, $L^{\Pi}(w', v') \geq R_n(f^*) + \nu'$ translates into a pair $(W', V')$ in the domain $\|W'\| \leq C_1, \|V'\| \leq C_2$, $L(W', V') \geq R_n(f^*) + \nu'$, for which by Theorem 6 there exists $(W^*_{\Sigma}, V^*_{\Sigma})$ such that

$$\mathbb{E}_{\Sigma} L(W' - \eta/2W' + \sqrt{\eta}W^*_{\Sigma}, V' - \eta/2V' + \sqrt{\eta}V^*_{\Sigma}) \leq L(W', V') - \eta\nu'/4. \tag{3.71}$$

Translating back to the change of coordinates:

$$\mathbb{E}_{\Sigma} L^{\Pi}(w' - \eta/2w' + \sqrt{\eta}w^*_{\Sigma}, v' - \eta/2v' + \sqrt{\eta}v^*_{\Sigma}) \leq L(w', v') - \eta\nu'/4. \tag{3.72}$$

Now we apply Lemma 53 to translate this into an argument about the landscape of $L^{\Pi}$. As a result, applying the bounds in Equations (3.107) and (3.127), we obtain that for $(w', v')$ such that

$$L^{\Pi}(w', v') \geq R_n(f^*) + \nu',$$

we should either have

$$\begin{aligned}\|\nabla L^{\Pi}(w',v')\| &\geq \frac{\nu/4}{4\sqrt{\|w'\|^2+\|v'\|^2}}\\ &= \frac{\nu/4}{4\sqrt{\|W'\|^2+\|V'\|^2}}\\ &= \frac{\nu}{16\sqrt{C_1^2+C_2^2}},\end{aligned}$$

or

$$\begin{aligned}\lambda_{min}\Big(\nabla^2 L^{\Pi}(w',v')\Big) &\leq -\frac{\nu/4}{2\min_{\Sigma}(\|w^*\|^2+\|v^*\|^2)}\\ &= -\frac{\nu}{2\min_{\Sigma}(\|W_{\Sigma}^*\|^2+\|V_{\Sigma}^*\|^2)}\\ &\leq -\frac{\nu}{16(\zeta_1+\zeta_2)}\\ &= -\frac{\nu}{16(1+\zeta)}.\end{aligned}$$

Next, we want to apply Theorem 7 by setting

$$\begin{aligned}\gamma &= \frac{\nu}{16(1+\zeta)},\\ \aleph_\ell &= R_n(f^*)+\nu',\end{aligned}$$

and Lipschitz parameters $\rho_1, \rho_2, \rho_3 = poly(B, C_1, C_2, m_1, m_2, m_3)$ set as described in Section 3.7.1, Theorem 9. Also, note that as prescribed by Theorem 7, we set

$$\begin{aligned}C_1 &\coloneqq \frac{\aleph+4l}{\psi_1},\\ C_2 &\coloneqq \frac{\aleph+4l}{\psi_2},\end{aligned} \tag{3.73}$$

where $l = O(1)$ depends on our desired chance of success for the algorithm, specified in Theorem 7. Finally, note that Theorem 7 needs to work with a bounded noise on the gradient whose covariance matrix is bounded between two multipliers of identity. The point of injecting extra noise to SGD in `PSGD` is in fact because of this covariance condition that we need in Theorem 7. On the other hand, note that in general, because of the gaussian smoothing that we use, the noise vector is not supported on a bounded domain, which makes it a bit harder to apply Hoeffding type concentration. To remedy this, we introduce a coupling between our unbounded noise vector for $L(W', V')$ and another noise random variable whose support is bounded, which with high probability is equal to the real noise, along all iterations. In Corollary, we further translate this coupling for the objective $L^{\Pi}$ after change of cooridnates, and write down the exact dependencies of the parameters $Q$, $\sigma_1$ and $\sigma_2$, which are all polynomial in the basic parameters and the overparameterization.

Hence, the conditions of Theorem 7 are satisfied, so we conclude that after at most $poly(\rho_1, \rho_2, \rho_3, Q, \aleph, C_1, C_2, 1/\gamma, \log(\sigma_1/\sigma_2)) = poly(B, m_1, m_2, m_3, C_1, C_2, \zeta_1, \zeta_2) = poly(n, B \vee 1/B, 1/\gamma_0)$ number of iterations, `PSGD` reach a point $w_t$ in some iteration $t$ with $L^{\Pi}(w_t) \leq \aleph_\ell$.

Translating back this $w_t = (w'_t, v'_t)$ by multiplying the $v'_t$ part to $\Upsilon^T$, we get a pair $(W'_t, V'_t)$ with objective value bounded as

$$L(W'_t, V'_t) \leq R_n(f^*) + \nu'. \tag{3.74}$$

But note that we obviously have the condition $\|W'\| \leq C_1, \|V'\| \leq C_2$ through the whole iterations, for the choice of $C_1, C_2$ in Equation (3.73). Therefore, using Lemma 43, for every $i \in [n]$:

$$|f'_{W',V'}(x_i)| = O(C_1, C_2), \tag{3.75}$$

$$|f_{W',V'}(x_i)| = O(C_1, C_2) \tag{3.76}$$

From Equations (3.76), as also stated in Theorem 9, we know that for all $i \in [n]$, $\ell(., y_i)$ is $O(C_1C_2)+B^2$-Lipschitz at points $f_{W',V'}(x_i)$ and $f'_{W',V'}(x_i)$, so we can bound the difference $|\ell(f'_{W',V'}(x_i), y_i) - \ell(f_{W',V'}(x_i), y_i)|$ by $(O(C_1C_2)+B^2)|f'_{W',V'}(x_i), y_i) - f_{W',V'}(x_i)|$, which in turn can become arbitrarily small having enough overparameterization using Lemma 44, in particular, we force it to be smaller than $O(\nu'/(B^2 + C_1C_2))$ (recall

$\nu' \geq \nu/2 \geq B^2/(2n))$. As a result, we get $|\ell(f'_{W',V'}(x_i), y_i) - \ell(f_{W',V'}(x_i), y_i)| = O(\nu)$ for every $i \in [n]$, which in turn implies $|L(W', V') - L_1(W', V')| \leq \nu$ by picking small constants, where recall that the objective $L_1$ is the same as $L$ but without the smoothing. Now applying this bound to Equation (3.74), we get

$$L_1(W'_t, V'_t) \leq R_n(f^*) + 2\nu'.$$

Therefore, as PSGD check the values of $L_1$ in the loop, it terminates at such pair $(W_t, V_t)$. From this point onward, we refer to the returned $(W'_t, V'_t)$ as just $(W', V')$.

Opening the definition of $L_1(W', V')$, we clearly get

$$R_n(f_{W',V'}) \leq L_1(W', V') \leq R_n(f^*) + 2\nu' \leq R_n(f^*) + 2\nu. \tag{3.77}$$

Furthermore, noting the setting of $\psi_1, \psi_2$ in Theorem 3 and the fact that $\nu' \geq \nu/2 \geq R_n(f^*)/8$, we get

$$\|W'\|^2 \leq \frac{4(R_n(f^*) + 2\nu')}{\nu'} \leq 40, \tag{3.78}$$

$$\|V'\|^2 \leq \frac{4\zeta(R_n(f^*) + 2\nu')}{\nu'} \leq 40\zeta, \tag{3.79}$$

which completes the proof. The fact that for every $i \in [n], j \in [m_1], j \notin P$ we have that $\mathrm{sign}((W_j^{(0)} + W'_j)^T x_i)$ and $\mathrm{sign}({W_j^{(0)}}^T x_i)$ are the same follows from Lemma 10.

### 3.6.11 Rademacher Complexity

In this section we show the proof for our Rademacher Complexity bound, which is used in Theorem 3.

**Theorem 5.** *Let $G_{\gamma_1,\gamma_2}$ be the class of neural nets with weights $(W, V)$ in our three layer setting, such that $\|W - W^{(0)}\| \leq \gamma_1$, $\|V - V^{(0)}\| \leq \gamma_2$, where for every $j \in [m_2], i \in [n]$: $V'_j \perp \phi^{(0)}(x_i)$, and for every $i \in [n], j \in [m_1], j \notin P$ for $P \subseteq [m_1]$ defined in Lemma 10, it satisfies* $\text{sign}((W_j^{(0)} + W'_j)x_i) = \text{sign}(W^{(0)}x_i)$. *Then, for large enough overparameterization, we have the following bound on the Rademacher complexity:*

$$\mathcal{R}(\mathcal{G}_{\gamma_1,\gamma_2}) \leq \frac{2\gamma_1\gamma_2}{\sqrt{n}}.$$

**Proof of Theorem 5**

Here, we do not have the smoothing matrices $W^\rho, V^\rho$ anymore. In this section, unlike the optimization section that we used $\{x'_i\}_{i=1}^n$ to denote the output of the first layer by incorporating also the smoothing matrices, here we define it without them:

$$x'_i = \frac{1}{\sqrt{m_1}} W^s \sigma((W^{(0)} + W')x_i).$$

Now define the matrices

$$Z'_i = 1/\sqrt{m_2}\Big(a_j \mathbb{1}\{V_{j,}^{(0)} x'_i \geq 0\} x'_i\Big)_{j=1}^{m_2},$$
$$Z_i'^{+} = 1/\sqrt{m_2}\Big(a_j (\mathbb{1}\{V_{j,} x'_i \geq 0\} - \mathbb{1}\{V_{j,}^{(0)} x'_i \geq 0\}) x'_i\Big)_{j=1}^{m_2}.$$

To bound the $Z_i'^{+}$ part, note that substituting $C_1$ by $\gamma_1$ in lemma 15 and assuming conditions

$$m_1 = \tilde{\Omega}(m_3^4),$$
$$\frac{2C_1^{3/2}}{\sqrt{\kappa_1}}(\frac{n^3 m_3^3}{m_1 \lambda_0})^{1/4} \leq \gamma_1,$$

(we can use this result because we do not have the smoothing matrix $W^{\rho}$ here), we get with high probability over the initialization for every $i \in [n]$:

$$\|\phi'(x_i)\| = \|x_i' - \phi^{(0)}(x_i)\| \lesssim \gamma_1. \tag{3.80}$$

Therefore, we can write

$$\begin{aligned}
|\mathrm{trace}(VZ_i'^{+})| &= \frac{1}{\sqrt{m_2}} \Big|\sum_j a_j \mathbb{1}\{\mathrm{sign}(V_{j,}x_i') \neq \mathrm{sign}(V_{j,}^{(0)}x_i')\} V_{j,}x_i'\Big| \\
&\leq \frac{1}{\sqrt{m_2}} \sum_j \mathbb{1}\{\mathrm{sign}(V_{j,}x_i') \neq \mathrm{sign}(V_{j,}^{(0)}x_i')\} |V_{j,}x_i'| \\
&\leq \frac{1}{\sqrt{m_2}} \sum_j \mathbb{1}\{|V_{j,}^{(0)}x_i'| \leq |(V_{j,} - V_{j,}^{(0)})x_i'|\} \Big(|(V_{j,} - V_{j,}^{(0)})x_i'| + |V_{j,}^{(0)}x_i'|\Big) \\
&\leq \frac{1}{\sqrt{m_2}} \sum_j \mathbb{1}\{|V_{j,}^{(0)}x_i'| \leq |(V_{j,} - V_{j,}^{(0)})x_i'|\} \Big(2|(V_{j,} - V_{j,}^{(0)})x_i'|\Big). \\
&\leq \frac{1}{\sqrt{m_2}} \sum_j \mathbb{1}\{|V_{j,}^{(0)}x_i'|, |(V_{j,} - V_{j,}^{(0)})x_i'| \leq \gamma_2^{2/3} (\frac{\kappa_2}{m_2})^{1/3} \min(\gamma_1, \|x_i'\|)\} \Big(2|(V_{j,} - V_{j,}^{(0)})x_i'|\Big) \\
&+ \frac{1}{\sqrt{m_2}} \sum_j \mathbb{1}\{|(V_{j,} - V_{j,}^{(0)})x_i'| \geq \gamma_2^{2/3} (\frac{\kappa_2}{m_2})^{1/3} \min(\gamma_1, \|x_i'\|)\} \Big(2|(V_{j,} - V_{j,}^{(0)})x_i'|\Big).
\end{aligned}$$

Now using the fact that $V_j - V_j^{(0)}$ is orthogonal to $\phi^{(0)}(x_i)$'s:

$$\begin{aligned}
LHS &\leq \frac{2\gamma_2^{2/3} (\frac{\kappa_2}{m_2})^{1/3} \gamma_1}{\sqrt{m_2}} \sum_j \mathbb{1}\{|V_{j,}^{(0)}x_i'| \leq \gamma_2^{2/3} (\frac{\kappa_2}{m_2})^{1/3} \|x_i'\|\} \\
&+ \frac{2}{\sqrt{m_2}} \sum_j \mathbb{1}\{\|V_{j,} - V_{j,}^{(0)}\| \|x_i' - \phi^{(0)}(x_i)\| \geq \gamma_2^{2/3} (\frac{\kappa_2}{m_2})^{1/3} \gamma_1\} \|V_{j,} - V_{j,}^{(0)}\| \|x_i' - \phi^{(0)}(x_i)\|.
\end{aligned}$$

Next, using the upper bound on $\|x_i' - \phi^{(0)}(x_i)\|$:

$$
\begin{aligned}
LHS &\lesssim \frac{2\gamma_2^{2/3}(\frac{\kappa_2}{m_2})^{1/3}\gamma_1}{\sqrt{m_2}} \sum_j (\mathbb{1}\{|V_{j,}^{(0)} x_i'| \le \gamma_2^{2/3}(\frac{\kappa_2}{m_2})^{1/3}\|x_i'\|\}. \\
&\quad + \frac{\gamma_1}{\sqrt{m_2}} \sum_j \mathbb{1}\{\|V_{j,} - V_{j,}^{(0)}\| \gtrsim \gamma_2^{2/3}(\frac{\kappa_2}{m_2})^{1/3}\}\|V_{j,} - V_{j,}^{(0)}\| \\
&\lesssim \frac{\gamma_2^{2/3}(\frac{\kappa_2}{m_2})^{1/3}\gamma_1}{\sqrt{m_2}} \sum_j (\mathbb{1}\{|V_{j,}^{(0)} x_i'| \le \gamma_2^{2/3}(\frac{\kappa_2}{m_2})^{1/3}\|x_i'\|\} \\
&\quad + \frac{\gamma_1}{\sqrt{m_2}} \sqrt{\sum_j \mathbb{1}\{\|V_{j,} - V_{j,}^{(0)}\|^2 \ge \gamma_2^{4/3}(\frac{\kappa_2}{m_2})^{2/3}\}} \sqrt{\sum_j \|(V_{j,} - V_{j,}^{(0)})\|^2} \\
&\le \frac{\gamma_2^{2/3}(\frac{\kappa_2}{m_2})^{1/3}\gamma_1}{\sqrt{m_2}} \Big(\sum_j \mathbb{1}\{|V_{j,}^{(0)} x_i'| \le \gamma_2^{2/3}(\frac{\kappa_2}{m_2})^{1/3}\|x_i'\|\}\Big) + \quad \frac{\gamma_2^2\gamma_1}{\sqrt{m_2}} \times (\frac{m_2}{\kappa_2})^{1/3}\frac{1}{\gamma_2^{2/3}}.
\end{aligned}
$$

Then, applying the first argument of Lemma 38, we have with high probability over the randomness of $V^{(0)}$:

$$
\begin{aligned}
LHS &\le \frac{\gamma_2^{2/3}(\frac{\kappa_2}{m_2})^{1/3}\gamma_1}{\sqrt{m_2}} (\frac{m_2}{\kappa_2}(\frac{\kappa_2}{m_2})^{1/3}\gamma_2^{2/3}) + \quad \frac{\gamma_2^2\gamma_1}{\sqrt{m_2}} \times (\frac{m_2}{\kappa_2})^{1/3}\frac{1}{\gamma_2^{2/3}} \\
&\le \frac{\gamma_2^{4/3}\gamma_1}{(\kappa_2\sqrt{m}_2)^{1/3}} + \frac{\gamma_2^{4/3}\gamma_1}{(\kappa_2\sqrt{m}_2)^{1/3}} \\
&\lesssim \frac{\gamma_2^{4/3}\gamma_1}{(\kappa_2\sqrt{m}_2)^{1/3}}.
\end{aligned}
$$

Therefore, we can write:

$$\begin{aligned}
\mathcal{R}(\mathcal{G}_{\gamma_1,\gamma_2})|_{(x_i),(y_i)} &= \frac{1}{n}\mathbb{E}_\epsilon \sup_{V,W\in S}\sum_{t=1}^{n}\epsilon_i f_{V,W}(x_i) \\
&= \frac{1}{n}\mathbb{E}_\epsilon \sup_{V,W\in S}\sum_{i=1}^{n}\epsilon_i a^T\sigma(1/\sqrt{m_2}VW^s\sigma(1/\sqrt{m_1}Wx_i)) \\
&= \frac{1}{n}\mathbb{E}_\epsilon \sup_{V,W\in S}\sum_{i=1}^{n}\epsilon_i a^T\sigma(1/\sqrt{m_2}Vx_i') \\
&= \frac{1}{n}\mathbb{E}_\epsilon \sup_{W\in S}\sup_{V\in S}\sum_{i=1}^{n}\epsilon_i \mathrm{trace}(V(Z_i' + Z_i'^{+})) \\
&\lesssim \frac{1}{n}\mathbb{E}_\epsilon \sup_{W,V\in S}\sum_{i=1}^{n}\epsilon_i \mathrm{trace}(VZ_i') + \frac{\gamma_2^{4/3}\gamma_1}{(\kappa_2\sqrt{m}_2)^{1/3}} \\
&\le \frac{1}{n}\mathbb{E}_\epsilon \sup_{W\in S}\sup_{V\in S}\mathrm{trace}(V(\sum_{i=1}^{n}\epsilon_i Z_i')) + \frac{\gamma_2^{4/3}\gamma_1}{(\kappa_2\sqrt{m}_2)^{1/3}} \\
&= \frac{1}{n}\mathbb{E}_\epsilon \sup_{W\in S}\sup_{V\in S}\mathrm{trace}((V - V^{(0)})(\sum_{i=1}^{n}\epsilon_i Z_i')) \\
&+ \frac{1}{n}\mathbb{E}_\epsilon \sup_{W\in S}\mathrm{trace}(V^{(0)}(\sum_{i=1}^{n}\epsilon_i Z_i')) + \frac{\gamma_2^{4/3}\gamma_1}{(\kappa_2\sqrt{m}_2)^{1/3}}. \qquad (3.81)
\end{aligned}$$

For the first term, for every $j \in [m_2]$, define $H_j$ to be the set of $i$'s in $[n]$ where the $j$th column of $Z_i'$ is non-zero, i.e.

$$H_j = \{i \in [n] :\ V_j^{(0)}x_i' \ge 0\}.$$

Here, we use the crucial assumption that $(V - V^{(0)})_j{}^T\phi^{(0)}(x_i) = 0$, so we can drop the $\phi^{(0)}(x_i)$ term when $x_i'$ is multiplied to $V - V^{(0)}$. Using this trick and applying Cauchy Schwarz, we bound the first term as:

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\epsilon \sup_{W,V\in S}\mathrm{trace}((V - V^{(0)})(\sum_{i=1}^{n}\epsilon_i Z_i')) \\
&\le \frac{1}{n}\mathbb{E}_\epsilon\|V - V^{(0)}\| \sup_{W\in S}\sqrt{\frac{1}{m_2}\sum_{j=1}^{m_2}\|\sum_{i\in H_j}\epsilon_i\phi^{(2)}(x_i)\|^2}.
\end{aligned}$$

Further using Jensen's inequality:

$$\begin{aligned}&\frac{1}{n}\mathbb{E}_{\epsilon}\sup_{W,V\in S}\text{trace}((V-V^{(0)})(\sum_{i=1}^{n}\epsilon_i Z_i'))\\&\leq\frac{\|V-V^{(0)}\|}{n}\sqrt{\mathbb{E}_{\epsilon}\frac{1}{m_2}\sum_{j=1}^{m_2}\sup_{W\in S}\|\sum_{i\in H_j}\epsilon_i\phi^{(2)}(x_i)\|^2}.\end{aligned}\tag{3.82}$$

Using Equation (3.111) of Lemma 15 (note that we do not have the smoothing matrix $W^{\rho}$ here, so we are allowed to use this result), we obtain

$$\|\langle W-W^{(0)},Z_i^k\rangle-\phi'(x_i)\|\lesssim\frac{2C_1^{3/2}}{\sqrt{\kappa}_1}(\frac{n^3m_3^3}{m_1\lambda_0})^{1/4},$$

where $Z_i^k$'s are defined in Equation (3.299).

Plugging this back in (3.82):

$$\begin{aligned}&\mathbb{E}_{\epsilon}\sup_{W\in S}\|\sum_{i\in H_j}\epsilon_i\phi'(x_i)\|^2\\&\leq\mathbb{E}_{\epsilon}\sup_{W\in S}\Big(\|\sum_{i\in H_j}\epsilon_i\langle W-W^{(0)},Z_i^k\rangle\|+\frac{2C_1^{3/2}}{\sqrt{\kappa}_1}(\frac{n^3m_3^3}{m_1\lambda_0})^{1/4}\Big)^2\\&\lesssim\mathbb{E}_{\epsilon}\sup_{W\in S}\|\sum_{i\in H_j}\epsilon_i\langle W-W^{(0)},Z_i^k\rangle\|^2+\frac{C_1^3}{\kappa_1}(\frac{n^3m_3^3}{m_1\lambda_0})^{1/2}\\&=\mathbb{E}_{\epsilon}\sup_{W\in S}\sum_{k=1}^{m_3}\Big(\text{trace}((W-W^{(0)})(\sum_{i\in H_j}\epsilon_i Z_i^k))\Big)^2+\frac{C_1^3}{\kappa_1}(\frac{n^3m_3^3}{m_1\lambda_0})^{1/2}.\end{aligned}\tag{3.83}$$

Now for every fixed dataset, with high probability over the randomness of $W^s$, for every $k_1\neq k_2$:

$$\begin{aligned}&\Big|\langle\sum_{i\in H_j}\epsilon_i Z_i^{k_1},\sum_{i\in H_j}\epsilon_i Z_i^{k_2}\rangle\Big|\leq\sum_{i_1,i_2\in H_j}\Big|\langle Z_{i_1}^{k_1},Z_{i_2}^{k_2}\rangle\Big|\\&=\frac{1}{m_1}\sum_{i_1,i_2\in H_j}\Big|\sum_{j=1}^{m_1}W_{k_1,j}^sW_{k_2,j}^s\langle x_{i_1},x_{i_2}\rangle\mathbb{1}\{W_{j,}^{(0)}x_{i_1}\geq 0\}\mathbb{1}\{W_{j,}^{(0)}x_{i_2}\geq 0\}\Big|\end{aligned}$$

But note that because $\langle x_{i_1},x_{i_2}\rangle\leq 1$, the variables $W_{k_1,j}^sW_{k_2,j}^s\langle x_{i_1},x_{i_2}\rangle\mathbb{1}\{W_{j,}^{(0)}x_{i_1}\geq$

$0\}\mathbb{1}\{W_{j,}^{(0)}x_{i_2} \geq 0\}$ are subgaussian with parameter one with respect to the randomness of $W^s$. Hence, with high probability over the randomness of $W^s$, we get

$$\Big|\langle \sum_{i\in H_j} \epsilon_i Z_i^{k_1}, \sum_{i\in H_j} \epsilon_i Z_i^{k_2}\rangle\Big| \lesssim \frac{1}{m_1} \sum_{i_1,i_2\in H_j} \sqrt{m_1} \leq \frac{n^2}{\sqrt{m}_1}. \tag{3.84}$$

Therefore, with high probability over the randomness of $W^{(0)}$ and $W'$ and the dataset, we get Equation (3.84). In order to get rid of the high probability argument on the dataset, we use the stronger Equation (3.300) in Lemma 55 which uniformly bounds $\langle Z_{k_1}(x), Z_{k_2}(x')\rangle$ by $\log(m_1)d/m_1$ for any $x, x'$, which in turn gives

$$\Big|\langle \sum_{i\in H_j} \epsilon_i Z_i^{k_1}, \sum_{i\in H_j} \epsilon_i Z_i^{k_2}\rangle\Big| \leq \sum_{i_1,i_2\in H_j} \Big|\langle Z_{i_1}^{k_1}, Z_{i_2}^{k_2}\rangle\Big| \lesssim \frac{n^2 d\log(m_1)}{\sqrt{m}_1},$$

with high probability, independent of the choice of dataset. This bounds is slightly worse comapred to (3.84), but still efficient for our purpose.

Furthermore, a similar bound to Equation (3.84) can be obtained in a more adversarial situation when we also take maximum against the choice of the dataset.

Note that the entries of $\sum_{i\in H_j} \epsilon_i Z_i^k$ for $1 \leq k \leq m_3$ can differ only in a sign. Therefore, their norms are all equal. Now suppose that $\mathcal{C}_j$ is the random variable of the norm of these variables:

$$\mathcal{C}_j := \| \sum_{i\in H_j} \epsilon_i Z_i^k \|.$$

Then, by substituting $r_k = \frac{1}{\mathcal{C}_j} \sum_{i\in H_j} \epsilon_i Z_i^k$ in Lemma 49, we get

$$\sum_{k=1}^{m_3} \Big(\text{trace}((W - W^{(0)})(\sum_{i\in H_j} \epsilon_i Z_i^k))\Big)^2 \leq \mathcal{C}_j^2(1 + m_3^2 O(\frac{n^2 d\log(m_1)}{\sqrt{m}_1 {\mathcal{C}_j}^2}))\|W - W^{(0)}\|_F^2 \tag{3.85}$$

$$= (\mathcal{C}_j^2 + \frac{n^2 m_3^2 d\log(m_1)}{\sqrt{m}_1})\|W - W^{(0)}\|_F^2. \tag{3.86}$$

Now recall from Equation (3.80), we have

$$\|\phi'(x_i)\| \leq \gamma_1. \tag{3.87}$$

Hence, we can apply Corollary 5.1 with $\phi^{(2)}(x_i)$ and $C_1$ substituted by $\phi'(x_i)$ and $\gamma_1$ respectively, to argue with high probability over the initialization, there exists a set $\tilde{P}_i$ such that for every $i \in [n]$ and $j \notin \tilde{P}_i$, sign of ${V_j}^T x_i'$ is the same as ${V_j^{(0)}}^T \phi^{(0)}(x_i)$, and moreover,

$$|\tilde{P}_i| \lesssim \big(\frac{C_1^2}{(m_3\kappa_1^2)}\big)^{1/3} m_2.$$

Now let

$$\bar{H}_j = \{i \in [n] : \; V_j^{(0)}\phi^{(0)}(x_i) \geq 0\}.$$

Note that for $j \notin \tilde{P} = \bigcup_i \tilde{P}_i$, we have $H_j = \bar{H}_j$. Now note that the norm of each $\sum_{i \in H_j} \epsilon_i Z_i^k$ is at most one. for each index $1 \leq \ell \leq m_1 d$, as the random variables $\sum_{i \in \bar{H}_j} \epsilon_i (Z_i^k)_\ell$ are $\sum_{i \in \bar{H}_j} (Z_i^k)_\ell^2 \leq \sum_{i \in [n]} (Z_i^k)_\ell^2$ subgaussian, we have with probability at least $1 - \frac{1}{n}$ over the randomness of $\epsilon_i$'s, for every $1 \leq \ell \leq m_1 d$ and every $1 \leq j \leq m_2$:

$$\Big|\sum_{i \in \bar{H}_j} \epsilon_i (Z_i^k)_\ell\Big| \leq \sqrt{\sum_{i \in [n]} (Z_i^k)_\ell^2 \log(m_1 d m_2 n)},$$

which implies for every $j \in [m_2]$:

$$\|\sum_{i \in \bar{H}_j} \epsilon_i Z_i^k\|^2 \leq \sum_\ell \sum_{i \in [n]} (Z_i^k)_\ell^2 \log(m_1 d) \leq n \log(m_1 d m_2 n).$$

Name this event $\mathcal{B}$, so

$$\mathbb{P}(\mathcal{B}) \leq \frac{1}{n}.$$

Note that although $H_j$ might depend on the randomness of $\epsilon_i$'s, $\bar{H}_j$ does not, and if $j \notin \tilde{P}$, we obtain

$$\mathcal{C}_j = \| \sum_{i \in \bar{H}_j} \epsilon_i Z_i^k \| \leq \sqrt{n \log(m_1 d m_2 n)}.$$

Moreover, note that we also have the following worse case bound:

$$\mathcal{C}_j = \| \sum_{i \in H_j} \epsilon_i Z_i^k \| \leq \sum_{i \in \bar{H}_j} \|Z_i^k\| \leq n.$$

Applying the last two inequalities into Equations (3.279) and (3.86):

$$\begin{aligned}
&\mathbb{E}_\epsilon \frac{1}{m_2} \sum_{j=1}^{m_2} \sup_{W \in S} \| \sum_{i \in H_j} \epsilon_i \phi'(x_i) \|^2 \\
&\leq \frac{C_1^3}{\kappa_1} (\frac{n^3 m_3^3}{m_1 \lambda_0})^{1/2} + \frac{1}{m_2} \sum_{j=1}^{m_2} \mathbb{E}_\epsilon \sup_{W \in S} \sum_{k=1}^{m_3} \Big( \operatorname{trace}((W - W^{(0)})(\sum_{i \in H_j} \epsilon_i Z_i^k)) \Big)^2 \\
&\leq \frac{C_1^3}{\kappa_1} (\frac{n^3 m_3^3}{m_1 \lambda_0})^{1/2} + \frac{1}{m_2} \mathbb{E}_\epsilon \mathbb{1}\{\mathcal{B}\} \sum_{j \in \tilde{P}} (C_j^2 + \frac{n^2 m_3^2 d \log(m_1)}{\sqrt{m_1}}) \|W - W^{(0)}\|_F^2 \\
&+ \frac{1}{m_2} \mathbb{E}_\epsilon \mathbb{1}\{\mathcal{B}\} \sum_{j \notin \tilde{P}} (C_j^2 + \frac{n^2 m_3^2 d \log(m_1)}{\sqrt{m_1}}) \|W - W^{(0)}\|_F^2 \\
&+ \frac{1}{m_2} \mathbb{E}_\epsilon \mathbb{1}\{\mathcal{B}^c\} \sum_{j=1}^{m_2} (C_j^2 + \frac{n^2 m_3^2 d \log(m_1)}{\sqrt{m_1}}) \|W - W^{(0)}\|_F^2 \\
&\leq \frac{C_1^3}{\kappa_1} (\frac{n^3 m_3^3}{m_1 \lambda_0})^{1/2} + \|W - W^{(0)}\|_F^2 \Big[ \frac{|\tilde{P}|}{m_2} \Big( n^2 + \frac{n^2 m_3^2 d \log(m_1)}{\sqrt{m_1}} \Big) + 2 \Big( n + \frac{n^2 m_3^2 d \log(m_1)}{\sqrt{m_1}} \Big) \Big] \\
&\leq \frac{C_1^3}{\kappa_1} (\frac{n^3 m_3^3}{m_1 \lambda_0})^{1/2} + \gamma_1^2 \Big[ n^3 \big( \frac{C_1^2}{(m_3 \kappa_1^2)} \big)^{1/3} + \big( \frac{C_1^2}{(m_3 \kappa_1^2)} \big)^{1/3} \frac{n^3 m_3^2 d \log(m_1)}{\sqrt{m_1}} + 2 \frac{n^2 m_3^2 d \log(m_1)}{\sqrt{m_1}} + 2n \Big].
\end{aligned} \tag{3.88}$$

Next, we analyze the term $\frac{1}{n} \mathbb{E}_\epsilon \sup_{W \in S} \operatorname{trace}(V^{(0)}(\sum_{i=1}^n \epsilon_i Z_i'))$. Noting that $\|\phi^{(0)}(x_i)\| \lesssim \kappa_1 \sqrt{m_3}$ with high probability and using Equation (3.87):

$$\| \sum_{i=1}^n \epsilon_i Z_i' \|_F \leq \sum_{i=1}^n \|Z_i'\|_F \leq \sum_{i=1}^n \|x_i'\| \leq \sum_i (\|\phi'(x_i)\| + \|\phi^{(0)}(x_i)\|) \lesssim n(\sqrt{m}_3 \kappa_1 + \gamma_1). \tag{3.89}$$

Hence

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_\epsilon \sup_{W \in S} \operatorname{trace}(V^{(0)}(\sum_{i=1}^n \epsilon_i Z_i')) &= \frac{1}{n}\mathbb{E}_\epsilon \sup_{W \in S} \sum_{i=1}^n \epsilon_i \operatorname{trace}(V^{(0)} Z_i') \\
&= \frac{1}{n}\mathbb{E}_\epsilon \sup_{W \in S} \sum_{i=1}^n \epsilon_i \Big( \frac{1}{\sqrt{m_2}} \sum_{j=1}^{m_2} a_j V_{j,}^{(0)} x_i' \mathbb{1}\{V_{j,}^{(0)} x_i' \geq 0\} \Big). \\
&= \frac{1}{n}\mathbb{E}_\epsilon \sup_{W \in S} \sum_{i=1}^n \epsilon_i \frac{1}{\sqrt{m_2}} a^T \sigma(V^{(0)} x_i') \leq \sup_{W \in S} \frac{1}{\sqrt{m_2}} a^T \sigma(V^{(0)} x_i').
\end{aligned}$$

But using Lemma 39:

$$LHS \lesssim \kappa_2 \sqrt{m_3} \|x_i'\|.$$

Applying a similar bound as we did in Equation (3.89) on $\|x_i'\|$:

$$\|x_i'\| \leq \|\phi^{(0)}(x_i)\| + \|\phi'(x_i)\| \lesssim \kappa_1 \sqrt{m_3} + \gamma_1.$$

Substituting above, we get

$$\frac{1}{n}\mathbb{E}_\epsilon \sup_{W \in S} \operatorname{trace}(V^{(0)}(\sum_{i=1}^n \epsilon_i Z_i')) \leq \kappa_2 \sqrt{m_3}(\kappa_1 \sqrt{m_3} + \gamma_1). \tag{3.90}$$

Finally, Substituting Equations (3.88) into (3.82), then combining it with (3.90) into (3.81), we obtain a bound on Rademacher complexity which holds w.h.p over both the randomness of the initialization and the dataset:

$$\begin{aligned}
\mathcal{R}(\mathcal{G}_{\gamma_1,\gamma_2})|_{x,y} \lesssim & \sqrt{\frac{C_1^3}{\kappa_1} (\frac{n^3 m_3^3}{m_1 \lambda_0})^{1/2}} \qquad (3.91) \\
& + \frac{\gamma_1 \gamma_2}{n} \sqrt{n^3 \big(\frac{C_1^2}{(m_3 \kappa_1^2)}\big)^{1/3} + \big(\frac{C_1^2}{(m_3 \kappa_1^2)}\big)^{1/3} \frac{n^3 m_3^2 d \log(m_1)}{\sqrt{m_1}} + 2\frac{n^2 m_3^2 d \log(m_1)}{\sqrt{m_1}} + 2n} \qquad (3.92) \\
& + \kappa_2 \sqrt{m_3}(\kappa_1 \sqrt{m_3} + \gamma_1) + \frac{\gamma_2^{4/3} \gamma_1}{(\kappa_2 \sqrt{m_2})^{1/3}}.
\end{aligned}$$

Having enough overparameterization, we have for every dataset $(x, y)$ (i.e. worst-case

Rademacher complexity):

$$\mathcal{R}(\mathcal{G}_{\gamma_1,\gamma_2})|_{x,y} \leq 2\gamma_1\gamma_2/\sqrt{n}. \tag{3.93}$$

Note that for the bound (3.93) to hold, the overparameterization should be picked poly large in $\gamma_1, \gamma_2$, as well as in other basic parameters. However, noting Equations (3.49) and (3.57) in the proof of Theorem 3, we set $\gamma_1 = 1, \gamma_2 \geq \Omega(B, n, 1/\gamma_0)$ in Theorem 3, so $\gamma_1\gamma_2$ is at most poly in the basic parameters. Therefore, again the overparameterization can be picked polynomially large in the basic only parameters (i.e. **independent** of $\gamma_1, \gamma_2$ or $\zeta$).

### 3.6.12 Constructing $W^*, V^*$

This section consists of two subsections; First, we prove a structural result for the first layer weights $(W', V')$ that the algorithm visits, then construct a weight matrix $W^*$ for the first layer with some good properties. Second, we do the same thing for the second layer (however, the structure of the first and second layers are completely different). Through out this section, we assume we have the norm bounds $\|W'\| \leq C_1, \|V'\| \leq C_2$.

Notably, we rely on a number of basic Lemmas more related to the representation power of the network, which we defer their proof into a later Section 3.7.2 and refer to them here as needed.

**First Layer, Construction of $W^*$**

**Lemma 10.** *Suppose $m_1 \geq 16n^2m_3^2/\lambda_0^2$. Let $P_i = \{j \in [m_1] |\ |W_j^{(0)}x_i| \leq c_2/\sqrt{m_1}\}$ and $P = \cup P_i$. During SGD iterations, suppose we have $\|W'\|_F \leq C_1$. Then, for a value $c_2$ satisfying*

$$2C_1\sqrt{nm_3}/\sqrt{\lambda_0} \leq c_2 \leq \kappa_1\lambda_0\sqrt{m_1}/(2n^2),$$

*with high probability $\forall i$:*

$$|P_i| \lesssim c_2\sqrt{m_1}/\kappa_1,$$

*and for $j \notin P$, during the whole algorithm we have*

$$\begin{aligned}
&\|W_j'\| \leq \frac{\sqrt{nm_3}C_1}{\sqrt{m_1\lambda_0}} + c_2/(4\sqrt{m_1}) \leq c_2/(2\sqrt{m_1}),\\
&c_2/\sqrt{m_2} \leq |W_j^{(0)}x_i|.
\end{aligned}$$

*So the signs of neurons outside $P$ never changes. In particular, we can set $c_2$ as small as $c_2 = C_1\sqrt{nm_3}/\sqrt{\lambda_0}$. In the rest of the proof (i.e. other sections), we set $c_2$ to this value.*

**Proof of Lemma 10**

Define the matrix

$$\tilde{Z}_k^i = \frac{1}{\sqrt{m}_1}(W^s_{k,j} x_i \mathbb{1}\{\forall i:\ W_j^{(0)T} x_i \geq c_2/\sqrt{m}_1\})_{j=1}^n.$$

Let $P_i$ be the set of indices $j$ such that $\mathbb{1}\{W_j^{(0)T} x_i \geq c_2/\sqrt{m}_1\}$ is zero. First of all, note that by Bernstein inequality:

$$|P_i| \leq c_2\sqrt{m_1}/\kappa_1 + O(\sqrt{c_2\sqrt{m_1}/\kappa_1} + 1) \lesssim c_2\sqrt{m}_1/\kappa_1.$$

Now suppose that until the current iteration of the algorithm the assumption has been true, i.e. the signs of the neurons outside of $P$ have never changed. As a result, due to the specific update of the SGD for both of the terms $\mathbb{E}_Z \ell(f_{V',W'}(x), y)$ and $\|W'\|_F^2$, if we define $W'|_P$ to be the restriction of $W'$ to indices that are not in $P$ (i.e. the columns in $P$ are equal to zero), then we can write

$$W'|_P{}^T = \sum_{k=1}^{m_3}\sum_{i=1}^{n} \alpha_{k,i}\tilde{Z}_k^i. \tag{3.94}$$

An issue here is that we also have some injected noise by `PSGD` into $W'$ which violates Equation (3.94). To handle the injected noise as well, we define the subspace $\Phi'$ of $\mathbb{R}^{m_1\times d}$ matrices to be the set of vectors with arbitrary rows for $j \in [m_1]$ with $j \in P$, while restricted to the other rows $j \notin P$ in should be in the span of $(\tilde{Z}_k^i)_{i,k}$. Then, we decompose $W'$ into subspaces $\Phi'$ and $\Phi'^{\perp}$ respectively as $W' = W'^{(1)} + W'^{(2)}$, where $W'^{(1)} \in \Phi', W'^{(2)} \in \Phi'^{\perp}$. Here, we want to prove $\|W'^{(1)}_j\| \leq c_2/(4\sqrt{m}_1)$. We handle the $\|W'^{(2)}_j\|$ part in Section 54. So instead of $W'|_P$ in Equation (3.94) we consider $W'^{(1)}|_P$:

$$W'^{(1)}|_P{}^T = \sum_{k=1}^{m_3}\sum_{i=1}^{n} \alpha_{k,i}\tilde{Z}_k^i. \tag{3.95}$$

We handle the other part $W'^{(2)}$ in Section 54. Now exactly similar to the drivation in

Lemma 47, we can state with high probability

$$\begin{aligned} C_1^2 &\geq \|W'\|^2 \geq \|W'^{(1)}\|^2 \\ &\geq \|W'^{(1)}|_P\|^2 \geq \sum_{k=1}^{m_3} \|\sum_{i=1}^{n} \alpha_{k,i} \tilde{Z}_k^i\|^2 - O(nm_3/\sqrt{m_1}) \sum_k \|\alpha_k\|^2. \end{aligned} \tag{3.96}$$

Note that we are exploiting the fact that the norm $\|W'\|$ remains bounded by $C_1$. Now using a Hoeffding bound for matrix $H^{\infty\prime}$ defined below, we write:

$$\begin{aligned} H^{\infty\prime}_{i_1,i_2} &:= \mathbb{E}_{w:\mathcal{N}(0,\mathbb{R}^d)} \mathbb{1}\{\forall i: \ |w^T x_i| \geq c_2/\sqrt{m_1}\} x_{i_1}^T x_{i_2} (\mathbb{1}\{w^T x_{i_1} \geq 0\}\mathbb{1}\{w^T x_{i_2} \geq 0\}) \\ &= \mathbb{E}_{w:\mathcal{N}(0,\mathbb{R}^d)} (\mathbb{1}\{w^T x_{i_1} \geq 0\}\mathbb{1}\{w^T x_{i_2} \geq 0\}) x_{i_1}^T x_{i_2} \\ &\pm O(\mathbb{E}\mathbb{1}\{\exists i: \ |w^T x_i| \leq c_2/\sqrt{m_1}\}(\mathbb{1}\{w^T x_{i_1} \geq 0\}\mathbb{1}\{w^T x_{i_2} \geq 0\})) x_{i_1}^T x_{i_2} \\ &= H^\infty_{i_1,i_2} \pm O(nc_2/(\sqrt{m_1}\kappa_1)\|x_{i_1}\|\|x_{i_2}\|) \\ &= H^\infty_{i_1,i_2} \pm O(nc_2/(\sqrt{m_1}\kappa_1)). \end{aligned} \tag{3.97}$$

Now opening Equation (3.96) and using the property $c_2 \leq k_1 \lambda_0 \sqrt{m_1}/(2n^2)$, we get

$$\begin{aligned} LHS &= \sum_k \sum_{i_1,i_2} \alpha_{k,i_1}\alpha_{k,i_2} \langle \tilde{Z}_k^{i_1}, \tilde{Z}_k^{i_2} \rangle - O(nm_3/\sqrt{m_1} \sum_k \|\alpha_k\|^2) \\ &= \sum_k \sum_{i_1,i_2} \alpha_{k,i_1}\alpha_{k,i_2} (H^{\infty\prime}_{i_1,i_2} \pm O(1/\sqrt{m_1})) - O(nm_3/\sqrt{m_1} \sum_k \|\alpha_k\|^2) \\ &\geq \sum_k \sum_{i_1,i_2} \alpha_{k,i_1}\alpha_{k,i_2} H^\infty_{i_1,i_2} \pm \|\alpha_k\|_1^2 O(nc_2/\sqrt{m_1}\kappa_1) - O(nm_3/\sqrt{m_1} \sum_k \|\alpha_k\|^2) \\ &\geq \sum_k \alpha_k^T H^\infty \alpha_k - O(nc_2/\sqrt{m_1}\kappa_1) \sum_k \|\alpha_k\|_1^2 - O(nm_3/\sqrt{m_1} \sum_k \|\alpha_k\|^2) \\ &\geq \sum_k \alpha_k^T H^\infty \alpha_k - O(c_2 n^2/\sqrt{m_1}\kappa_1) \sum_k \|\alpha_k\|_2^2 - O(nm_3/\sqrt{m_1} \sum_k \|\alpha_k\|^2) \\ &= (\lambda_0 - O(nm_3/\sqrt{m_1}) - O(c_2 n^2/\sqrt{m_1}\kappa_1)) \sum_k \|\alpha_k\|^2 \\ &\geq \lambda_0/2 \sum_k \|\alpha_k\|^2. \end{aligned}$$

For the last line to hold, we need enough overparameterization. This implies

$$\sum_k \|\alpha_k\|^2 \lesssim C_1^2/\lambda_0.$$

Now again, exactly similar to the derivation in Lemma 47, for $j \notin P$ we have

$$\|W'^{(1)}_j\| \leq \sqrt{nm_3}/\sqrt{m_1}\sqrt{\sum_k \|\alpha_k\|^2} \lesssim \sqrt{m_3 n}C_1/\sqrt{m_1\lambda_0},$$

which completes most of the proof. For the rest, we are left to show that for the other part $W'^{(2)}$, we have $\|W'^{(2)}_j\| \leq c_2/(4\sqrt{m_1})$, which we do in Section 54.

**Lemma 11.** *Under condition $m_3 n/\sqrt{m_1} \leq \lambda_0/4$, there exist matrices $\{W_k^*\}_{k=1}^{m_3} \in \mathbb{R}^{m_1\times d}$ s.t. for every $k \neq k' \in [m_3]$ and $i \in [n]$:*

$$\begin{aligned}
&\langle W_k^*, Z_{k'}^i\rangle = 0,\\
&\|W_k^* - W_k^+\| \lesssim \frac{n\sqrt{m_3}}{\lambda_0\sqrt{m_1}}\|\mathcal{V}_k\|_{H^\infty},\\
&|\langle W_k^*, Z_k^i\rangle - \langle W_k^+, Z_k^i\rangle| \lesssim \frac{n\sqrt{m_3}}{\lambda_0\sqrt{m_1}}\|\mathcal{V}_k\|_{H^\infty}.
\end{aligned}$$

*Furthermore, for $k_1 \neq k_2$:*

$$|\langle W_{k_1}^*, W_{k_2}^*\rangle| \leq \frac{n}{\lambda_0^2}\frac{\sqrt{m_3}}{\sqrt{m_1}}(1+\frac{\sqrt{m_3}}{\sqrt{m_1}})\|\mathcal{V}_{k_1}\|_{H^\infty}\|\mathcal{V}_{k_2}\|_{H^\infty}. \tag{3.98}$$

**Proof of Lemma 11**

Let

$$W_k^+ = \sum_i \mathcal{V}_{k,i}Z_k^i.$$

we want to compute the norm of the projection $P(W_k^+)$ of $W_k^+$ onto the subspace spanned by all $Z_{k'}^i$ for $k' \neq k$ and $i \in [n]$:

$$\|P(W_k^+)\|^2 = (\langle W_k^+, Z_{k'}^i \rangle)_{k' \neq k, i \in [n]}^T \Big( \langle Z_{k_1}^{i_1}, Z_{k_2}^{i_2} \rangle \Big)^{-1}_{(k_1,i_1),(k_2,i_2) \in [m_3]-\{k\}\times[n]} (\langle W_k^+, Z_{k'}^i \rangle)_{k' \neq k, i \in [n]}, \tag{3.99}$$

where the first and third terms are vectors and the middle term is a matrix. Now note that for each $k', k_1, k_2 \neq k$, by Hoeffding inequality:

$$\Big( \langle Z_{k'}^{i_1}, Z_{k'}^{i_2} \rangle \Big)_{i_1,i_2 \in [n]} = H^\infty + (\pm 1/\sqrt{m_1})_{i_1,i_2 \in [n]}, \tag{3.100}$$

$$\begin{aligned} \langle W_k^+, Z_{k'}^i \rangle = \langle \sum_i \mathcal{V}_{k,i} Z_k^i, Z_{k'}^i \rangle &\lesssim \frac{1}{\sqrt{m}_1} \sum_i |\mathcal{V}_{k,i}| \\ &\leq \frac{\sqrt{n}}{\sqrt{m}_1} \|\mathcal{V}_k\| \\ &\leq \frac{\sqrt{n}}{\sqrt{m_1 \lambda_0}} \|\mathcal{V}_k\|_{H^\infty}. \end{aligned} \tag{3.101}$$

Therefore,

$$\|(\langle W_k^+, Z_{k'}^i \rangle)_{k' \neq k, i \in [n]}^T\| \leq n \sqrt{\frac{m_3}{m_1 \lambda_0}} \|\mathcal{V}_k\|_{H^\infty}. \tag{3.102}$$

Now Equation (3.100) implies for small enough $m_1$

$$\lambda_{min}(\Big( \langle Z_{k'}^{i_1}, Z_{k'}^{i_2} \rangle \Big)_{i_1,i_2 \in [n]}) \geq \lambda_0/2, \tag{3.103}$$

as long as $\lambda_0 \geq 2n/m_1$. Moreover, define $\tilde{A}$ to be the block version of

$$A' = \Big( \langle Z_{k_1}^{i_1}, Z_{k_2}^{i_2} \rangle \Big)^{-1}_{(k_1,i_1),(k_2,i_2) \in [m_3]-\{k\}\times[n]},$$

i.e. for $k_1 = k_2$ they are the same but for $k_1 \neq k_2$ $\tilde{A}$ is zero. Then

$$\lambda_{min}(\tilde{A}) \geq \lambda_0/2,$$

because the eigenvalues of each block is at least $\lambda_0/2$ using Equation (3.103). But note that

$$\|A' - \tilde{A}\|_2 \le \|A' - \tilde{A}\|_F \le m_3 n/\sqrt{m_1}.$$

So as long as $m_3 n/\sqrt{m}_1 \le \lambda_0/4$, we have $\lambda_{min}(A) \ge \lambda_0/4$. Combining this fact with Equation (3.102) and plugging it into Equation (3.99), we obtain

$$\|P(W_k^+)\|^2 \lesssim \frac{n^2 m_3}{m_1 \lambda_0^2} \|\mathcal{V}_k\|_{H^\infty}^2.$$

Now define $W_k^* = W_k^+ - P(W_k^+)$. Then

$$\|W_k^* - W_k^+\| = \|P(W_k^+)\| \lesssim \frac{n\sqrt{m}_3}{\lambda_0 \sqrt{m_1}} \|\mathcal{V}_k\|_{H^\infty},$$
$$|\langle W_k^* - W_k^+, Z_k^i \rangle| \le \|P(W_k^+)\| \|Z_k^i\| \lesssim \frac{n\sqrt{m}_3}{\lambda_0 \sqrt{m_1}} \|\mathcal{V}_k\|_{H^\infty}.$$

Furthermore, note that $W_{k_2}^*$ is orthogonal to $W_{k_1}^+$ for $k_1 \neq k_2$, so

$$\begin{aligned} |\langle W_{k_1}^*, W_{k_2}^* \rangle| &= |\langle W_{k_1}^+ - P(W_{k_1}^+), W_{k_2}^* \rangle| \\ &= |\langle P(W_{k_1}^+), W_{k_2}^+ - P(W_{k_2}^+) \rangle| \\ &\le \|P(W_{k_1}^+)\| \|W_{k_2}^+ - P(W_{k_2}^+)\| \\ &\le \|P(W_{k_1}^+)\| (\|W_{k_2}^+\| + \|P(W_{k_2}^+)\|). \end{aligned} \tag{3.104}$$

But note that

$$\|W_k^+\| = \|\sum_i \mathcal{V}_{k,i} Z_k^i\| \le \sum_i |\mathcal{V}_{k,i}| \|Z_k^i\| \le \sum_i |\mathcal{V}_{k,i}| \le \sqrt{n} \|\mathcal{V}_k\|_2. \le \frac{\sqrt{n}}{\lambda_0} \|\mathcal{V}_k\|_{H^\infty}.$$

Therefore, we can bound Equation (3.104) as:

$$|\langle W_{k_1}^*, W_{k_2}^* \rangle| \le \frac{n}{\lambda_0^2} \frac{\sqrt{m}_3}{\sqrt{m_1}} (1 + \frac{\sqrt{m}_3}{\sqrt{m_1}}) \|\mathcal{V}_{k_1}\|_{H^\infty} \|\mathcal{V}_{k_2}\|_{H^\infty}.$$

**Lemma 12.** *There exists a matrix $W_k^{+2}$ such that for every $j \in P$, ${W_k^{+2}}_j = 0$, and*

$$|\mathit{trace}(W_k^{+2} Z_i^k) - \bar{x}_{i,k}| \le C_1 \sqrt{m_3} n^2 / (\lambda_0 \kappa_1 \sqrt{m_1}) \|\mathcal{V}_k\|_{H^\infty}.$$

**Proof of Lemma 12**

Define $W_k^{+2}$ to be equal to $W_k^+$ for $j \notin P$ and equal to zero vector otherwise. Then, by Lemma 47: (note that $|P_i| \leq C_1\sqrt{nm_3}\sqrt{m_1}/(\sqrt{\lambda}_0\kappa_1)$)

$$\begin{aligned}|\text{trace}(W_k^+ Z_i^k) - \text{trace}(W_k^{+2} Z_i^k)| &\leq 1/\sqrt{m}_1 \sum_{j\in P} |W_k^+{}_j x_i| \\ &\leq \frac{|P|}{\sqrt{m}_1}\|W_k^+\| \\ &\leq \sqrt{nm_3}/(m_1\sqrt{\lambda_0})\ |P| \|\mathcal{V}_k\|_{H^\infty} \\ &\leq C_1 m_3 n^2/((\lambda_0\kappa_1\sqrt{m_1})\ \|\mathcal{V}_k\|_{H^\infty}.\end{aligned}$$

Combining this with Lemma 46, the desired result follows.

**Lemma 13.** *Under condition $m_3 n/\sqrt{m}_1 \leq \lambda_0/4$, there exist matrix $W_k^*$'s exactly satisfying the same conditions in Lemma 11 but with respect to $W_k^{+2}$ instead of $W_k^+$, and moreover, for $j \in P$ we have $W_k^*{}_j = 0$.*

**Proof of Lemma 13**

We can repeat the exact same procedure of Lemma 11 for $W_k^{+2}$. Using the bound in Equation (3.97), we have

$$\begin{aligned}\left(\langle \tilde{Z}_{k'}^{i_1}, \tilde{Z}_{k'}^{i_2}\rangle\right)_{i_1,i_2\in[n]} &= H'^\infty + O(\pm 1/\sqrt{m_1})_{i_1,i_2\in[n]} \\ &= H^\infty + (\pm nc_2/\sqrt{m_1}\kappa_1)_{i_1,i_2\in[n]} + O(\pm 1/\sqrt{m_1})_{i_1,i_2\in[n]} \\ &= H^\infty + (\pm nc_2/\sqrt{m_1}\kappa_1)_{i_1,i_2\in[n]},\end{aligned}$$

so as long as

$$n^2 c_2/\sqrt{m}_1\kappa_1 = n^2 C_1\sqrt{nm_3}/(\kappa_1\sqrt{m_1\lambda_0}) \leq \lambda_0/2,$$

with similar argument as in Lemma 11, we get

$$\lambda_{min}(\left(\langle \tilde{Z}_{k'}^{i_1}, \tilde{Z}_{k'}^{i_2}\rangle\right)_{i_1,i_2\in[n]}) \geq \lambda_0/2.$$

Moreover,

$$\begin{aligned}\langle W_k^{+2}, \tilde{Z}_{k'}^i\rangle = \langle \sum_i \mathcal{V}_{k,i}\tilde{Z}_k^i, \tilde{Z}_{k'}^i\rangle &\lesssim \frac{1}{\sqrt{m_1}}\sum_i |\mathcal{V}_{k,i}| \\ &\leq \frac{\sqrt{n}}{\sqrt{m_1}}\|\mathcal{V}_k\| \\ &\leq \frac{\sqrt{n}}{\sqrt{m_1\lambda_0}}\|\mathcal{V}_k\|_{H^\infty}.\end{aligned}$$

Thus, using the same argument as before the proof is complete.

**Lemma 14.** *Suppose*

$$m_1 \geq n^7 m_3/\lambda_0.$$

*During SGD, suppose we are currently at $(V', W')$ with $W' \leq C_1$. For any matrix $W_1$, we denote the signs of the first layer imposed by $W_1$ by $D_{W_1,x_i}$. Then with high probability, there exists $W^* = \sum_{k\in[m_3]} W_k^*$ such that $W_k^*$'s is orthogonal to all other $Z_{k'}^i$'s for $k' \neq k$, and for every $i \in [n]$, we have:*

$$\|\frac{1}{\sqrt{m_1}}W^s D_{W^{(0)}+W',x_i} W^* x_i - \bar{x}_i\|_\infty \lesssim \frac{nm_3}{\sqrt{m_1}\lambda_0}\Big[1 + \frac{nC_1}{\kappa_1}\Big]\|\mathcal{V}_k\|_{H^\infty} \coloneqq \Re\|\mathcal{V}_k\|_{H^\infty}.$$

*Moreover, we have*

$$\|W_j^*\| \leq \sqrt{nm_3}/(\sqrt{m_1\lambda_0})\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2 + \frac{n\sqrt{m}_3}{\lambda_0\sqrt{m_1}}(\sum_k \|\mathcal{V}_k\|_{H^\infty})} \coloneqq \varrho\sqrt{\frac{m_3}{m_1}}, \quad (3.105)$$

*Particularly, for any diagonal sign matrix $\Sigma \in \mathbb{R}^{m_3\times m_3}$, we have*

$$\|W_\Sigma^*\|_F^2 \leq (\frac{n\sqrt{n}}{\lambda_0^2}\frac{\sqrt{m}_3}{\sqrt{m_1}}(1 + \frac{\sqrt{m}_3}{\sqrt{m_1}}) + (1 + O(n/(\lambda_0\sqrt{m}_1) + \frac{n^2 m_3}{\lambda_0^2 m_1})))\sum_k \|\mathcal{V}_k\|_{H^\infty}^2. \quad (3.106)$$

*which, by having enough overparameterization, implies*

$$\|W_\Sigma^*\|_F \leq \sqrt{2\sum_k \|\mathcal{V}_k\|_{H^\infty}^2} = \sqrt{2\zeta_1}, \quad (3.107)$$

*where*

$$W^*_\Sigma \coloneqq \sum_{k=1}^{m_3} \Sigma_k W^*_k. \tag{3.108}$$

*Moreover, we have*

$$\frac{1}{\sqrt{m_1}} W^s D_{W^{(0)}+W',x_i} W^*_\Sigma x_i = \Sigma \frac{1}{\sqrt{m_1}} W^s D_{W^{(0)}+W',x_i} W^* x_i. \tag{3.109}$$

**Proof of Lemma 14**

From Lemma 12, we have

$$|\bar{x}_{i,k} - \mathrm{trace}(W_k^{+2} Z_k^i)| \le C_1 m_3 n^2/(\lambda_0 \kappa_1 \sqrt{m_1})\ \|\mathcal{V}_k\|_{H^\infty}.$$

Combining this with the result of Lemma 13, we get:

$$\begin{aligned} |\bar{x}_{i,k} - \mathrm{trace}(W_k^* Z_k^i)| &\lesssim \Big[\frac{n\sqrt{m_3}}{\lambda_0\sqrt{m_1}} + \frac{C_1 m_3 n^2}{\lambda_0 \kappa_1 \sqrt{m_1}}\Big] \|\mathcal{V}_k\|_{H^\infty} \\ &= \frac{n m_3}{\sqrt{m_1}\lambda_0}\Big[1 + \frac{nC_1}{\kappa_1}\Big] \|\mathcal{V}_k\|_{H^\infty}. \end{aligned} \tag{3.110}$$

On the other hand, based on the property that $W^*_{k\,j} = 0$ for $j \in P$ and its orthogonal property from Lemma 13, for $j \in P$ we get

$$\begin{aligned} \frac{1}{\sqrt{m_1}} W_k^s D_{W^{(0)}+W',x_i} W^* x_i &= \frac{1}{\sqrt{m_1}} W_k^s D_{W^{(0)},x_i} W^* x_i \\ &= \mathrm{trace}(W^* Z_k^i) = \mathrm{trace}(W_k^* Z_k^i) \\ &= \frac{1}{\sqrt{m_1}} W_k^s D_{W^{(0)},x_i} W_k^* x_i, \end{aligned}$$

which combined with Equation (3.110) completes the proof. From the above, Equation (3.109) is also clear. Finally, note that by Lemma 47 we have

$$\|W^{+2}{}_j\| \leq \sqrt{nm_3}/(\sqrt{m_1\lambda_0})\sqrt{\sum_k \|\mathcal{V}_k\|^2_{H^\infty}}.$$

which Combined with Lemma 13 implies

$$\|W^*_j\| \leq \sqrt{nm_3}/(\sqrt{m_1\lambda_0})\sqrt{\sum_k \|\mathcal{V}_k\|^2_{H^\infty}} + \frac{n\sqrt{m_3}}{\lambda_0\sqrt{m_1}}(\sum_k \|\mathcal{V}_k\|_{H^\infty}) \coloneqq \varrho\sqrt{\frac{m_3}{m_1}},$$

while the other claims follows from Lemma 48 and Lemma 13, combined with Equation (3.98):

$$\begin{aligned}
\|W^*_\Sigma\|^2_F &\leq \sum_k \|W^*_k\|^2 + \sum_{k_1\neq k_2} |\langle W^*_{k_1}, W^*_{k_2}\rangle| \\
&\leq \sum_k \|W^*_k\|^2 + \frac{n}{\lambda_0^2}\frac{\sqrt{m_3}}{\sqrt{m_1}}(1 + \frac{\sqrt{m_3}}{\sqrt{m_1}})(\sum_k \|\mathcal{V}_k\|_{H^\infty})^2 \\
&\leq \sum_k \|W^*_k\|^2 + \frac{n}{\lambda_0^2}\frac{\sqrt{m_3}}{\sqrt{m_1}}(1 + \frac{\sqrt{m_3}}{\sqrt{m_1}})\sqrt{n}(\sum_k \|\mathcal{V}_k\|^2_{H^\infty}) \\
&\leq \sum_k \|W^*_k\|^2 + \frac{n\sqrt{n}}{\lambda_0^2}\frac{\sqrt{m_3}}{\sqrt{m_1}}(1 + \frac{\sqrt{m_3}}{\sqrt{m_1}})\zeta_1 \\
&\leq \frac{n\sqrt{n}}{\lambda_0^2}\frac{\sqrt{m_3}}{\sqrt{m_1}}(1 + \frac{\sqrt{m_3}}{\sqrt{m_1}})\zeta_1 + \sum_k \|W^{+2}_k\|^2 + \frac{n^2 m_3}{\lambda_0^2 m_1}\sum_k \|\mathcal{V}_k\|^2_{H^\infty} \\
&\leq \frac{n\sqrt{n}}{\lambda_0^2}\frac{\sqrt{m_3}}{\sqrt{m_1}}(1 + \frac{\sqrt{m_3}}{\sqrt{m_1}})\zeta_1 + \sum_k \|W^{+}_k\|^2 + \frac{n^2 m_3}{\lambda_0^2 m_1}\sum_k \|\mathcal{V}_k\|^2_{H^\infty} \\
&\lesssim \frac{n\sqrt{n}}{\lambda_0^2}\frac{\sqrt{m_3}}{\sqrt{m_1}}(1 + \frac{\sqrt{m_3}}{\sqrt{m_1}})\zeta_1 + (1 + O(n/(\lambda_0\sqrt{m_1}) + \frac{n^2 m_3}{\lambda_0^2 m_1}))\zeta_1.
\end{aligned}$$

Next, we move on to construct $V^*$ for the second layer.

### Second Layer, Construction of $V^*$

In this section, we present a couple of lemmas that step by step lead to the construction of $V^*$. we remind the reader that $\phi^{(0)}(x_i)$ is the output of the first layer at initialization weights, $\phi'(x_i)$ and $\phi^{(2)}(x_i)$ are the changes in the output of the first layer when $W'$ and $W' + W^\rho$ are added, respectively, and finally $\phi^*(x_i)$ is the optimal features that are generated by the matrix $W^*$ but with the sign pattern of $W^{(0)} + W'$, i.e.

$$\phi^*(x_i) = \frac{1}{\sqrt{m_1}} W^s D_{W^{(0)}+W',x_i} W^* x_i.$$

We also define $x_i'$ as

$$x_i' = \phi^{(0)}(x_i) + \phi^{(2)}(x_i) = \frac{1}{\sqrt{m_1}} W^s \sigma((W^{(0)} + W' + W^\rho)x_i).$$

To begin, we state a lemma to bound the magnitude of $\|\phi'(x_i)\|$, given that the norm of $W'$ is bounded by $C_1$ and the sign pattern $\mathrm{Sgn}\Big((W^{(0)} + W')x_i\Big)$ satisfies condition stated for the set of indices $P$ in Lemma 10. Later on, we exploit this Lemma in Lemma 42 to state bounds for $\|\phi^{(2)}(x_i)\|$.

**Lemma 15.** *Let the matrix $W'$ with norm bound $\|W'\| \leq C_1$, such that the signs of $(W_j^{(0)} + W_j')x_i$ and $W_j^{(0)}x_i$ can be different only for $j \in P$, for $P$ defined in Lemma 10. (Note that for $W'$ at every step of the algorithm, this is automatically satisfied by Lemma 10) Then*

$$\|\phi'(x_i)\| \leq \frac{2C_1^{3/2}}{\sqrt{\kappa_1}} (\frac{n^3 m_3^3}{m_1 \lambda_0})^{1/4} + (1 + O(m_3^2/\sqrt{m_1}))C_1.$$

*Particularly for large enough $m_1$ compared to $n, m_3, \lambda_0, \kappa_1, C_1$, we have*

$$\|\phi'(x_i)\| \lesssim C_1.$$

**Proof of Lemma 15**

We write

$$\begin{aligned}|\phi_k'(x_i) - \langle W', Z_k^i \rangle| &\leq 2/\sqrt{m_1} \sum_{j \in P} |W_j' x_i| \leq 2/\sqrt{m_1} \sum_{j \in P} \|W_j'\| \\ &\leq 2\sqrt{|P|}/\sqrt{m_1} \|W'\|_F \leq \frac{2C_1^{3/2}}{\sqrt{\kappa_1}} (\frac{n^3 m_3}{m_1 \lambda_0})^{1/4},\end{aligned} \tag{3.111}$$

where the last line follows from the bound on $|P|$ from Lemma 10.

On the other hand, because by Hoeffding we know that $\langle Z^i_k, Z^i_{k'}\rangle \lesssim 1/\sqrt{m_1}$ by Lemma 49, we get

$$\sum_{k=1}^{m_3}\langle W', Z^i_k\rangle^2 \leq (1+O(m_3^2/\sqrt{m_1}))\|W'\|_F^2 \leq (1+O(m_3^2/\sqrt{m_1}))C_1^2.$$

Combining this with Equation (3.111), we get

$$\begin{aligned}\|\phi'(x_i)\| &\leq \sqrt{\sum_k |\phi^{(2)}_k(x_i) - \langle W', Z^i_k\rangle|^2} + \sqrt{\sum_k \langle W', Z^i_k\rangle^2}\\ &\leq \frac{2C_1^{3/2}}{\sqrt{\kappa_1}}(\frac{n^3 m_3^3}{m_1\lambda_0})^{1/4} + (1+O(m_3^2/\sqrt{m_1}))C_1. \qquad (3.112)\end{aligned}$$

Next, we prove a structural lemma regarding the sign pattern in the second layer when we feed in $x'_i$ to it, with the important message that the dominance of sign patterns are specified by $\phi^{(0)}(x_i)$.

**Lemma 16.** *Suppose we have $m_3\kappa_1^2 \gtrsim C_1^2$, $\kappa_2\sqrt{m_2} \geq C_2$, and $m_1$ satisfies the condition on Lemma 15. If we have the condition $\|\phi^{(2)}(x_i)\| \lesssim C_1$, which happens under the high probability event $E^c$ defined in Lemma 42, then for every $i \in [n]$, there exist a subset $\tilde{P}_i$ which might depend on $W^{(0)}, V^{(0)}, W', V'$, such that*

$$|\tilde{P}_i| \lesssim \Big((\frac{C_1^2}{(m_3\kappa_1^2)})^{1/3} + (c_3 + \frac{C_1^2}{c_3^2 m_3\kappa_1^2})(\frac{C_2^2}{\kappa_2^2 m_2})^{1/3}\Big)m_2.$$

*Moreover, for every $i \in [n]$, for $j \notin \tilde{P}_{i,}$:*

$$\begin{aligned}&\frac{2}{3}|V^{(0)}_j\phi^{(0)}(x_i)| \geq |V^{(0)}_j\phi^{(2)}(x_i)| + |V'_j(\phi^{(0)}(x_i) + \phi^{(2)}(x_i))|,\\ &|V^{(0)}_j\phi^{(0)}(x_i)| \gtrsim (\frac{\kappa_2}{m_2})^{1/3}C_2^{2/3}c_3\|\phi^{(0)}(x_i)\|,\\ &|V^{(0)}_j\phi^{(0)}(x_i)| \gtrsim (\frac{\kappa_2}{m_2})^{1/3}C_2^{2/3}c_3\|x'_i\|.\end{aligned}$$

**Proof of Lemma 16**

By assumption, we know that during the algorithm, we have $\|V'\| \leq C_2$. Also, we know by Lemma 42 that under $E^c$:

$$\|\phi^{(2)}(x_i)\| \leq 2C_1.$$

Define the set

$$P_i' = \{j| \; |V_j^{(0)}\phi^{(0)}(x_i)| \leq c_3(\frac{\kappa_2}{m_2})^{1/3}C_2^{2/3}\|\phi^{(0)}(x_i)\|\} \tag{3.113}$$

and $P' = \cup P_i'$. We have

$$\mathbb{P}(|V_j^{(0)}\phi^{(0)}(x_i)| \leq c_3(\frac{\kappa_2}{m_2})^{1/3}C_2^{2/3}\|\phi^{(0)}(x_i)\|) \leq c_3(\frac{\kappa_2}{m_2})^{1/3}C_2^{2/3}/\kappa_2,$$

so by Bernstein, with high probability:

$$|P_i'| \leq m_2C_2^{2/3}c_3(\frac{\kappa_2}{m_2})^{1/3}/\kappa_2 + \sqrt{m_2C_2^{2/3}c_3(\frac{\kappa_2}{m_2})^{1/3}/\kappa_2} + 1 \lesssim c_3m_2C_2^{2/3}(\frac{\kappa_2}{m_2})^{1/3}/\kappa_2,$$

so with high prob.

$$|P_i'| \lesssim c_3C_2^{2/3}(\frac{m_2}{\kappa_2})^{2/3}. \tag{3.114}$$

On the other hand, Note that

$$\phi_k^{(0)}(x_i) = \sum_{j=1}^{m_1} 1/\sqrt{m}_1 W_{k,j}^s\sigma(W_j^{(0)}x_i) \tag{3.115}$$

is subGaussian with parameter $\sigma^2 = O(1/m_1\sum_j \sigma(W_j^{(0)}x_i)^2)$. Furthermore, note that if we compute the variance of $\phi_k^{(0)}(x_i)$ with respect to the randomness of $W^s$:

$$\mathbb{E}\phi_k^{(0)}(x_i)^2 = 1/m_1\sum_{j=1}^{m_1}\sigma(W_j^{(0)}x_i)^2 := \aleph$$

which itself concentrates around $1/2\kappa_1^2\|x_i\|^2 = 1/2\kappa_1^2$ by another Bernstein, i.e. $\aleph = 1/2\kappa_1^2(1 \pm O(1/\sqrt{m}_1))$. Therefore, by concentration of subexponential variables (Bernstein), it is not hard to see that the squared norm of the vector $\phi^{(0)}(x_i)$ is $(m_3\kappa_1^4, \kappa_2)$-subexponential and concentrates around $m_3\aleph$, i.e.

$$\|\phi^{(0)}(x_i)\|^2 = m_3\aleph \pm O(\kappa_1^2\sqrt{m}_3) = m_3\kappa_1^2/2 \pm O(m_3\kappa_1^2/\sqrt{m}_1) \pm O(\kappa_1^2\sqrt{m}_3), \quad (3.116)$$

with high probability. Combining this with the fact that $\|\phi^{(2)}(x_i)\| \lesssim C_1$ implies with high probability:

$$\frac{\|\phi^{(0)}(x_i)\|}{\|\phi^{(2)}(x_i)\|} \gtrsim \frac{\sqrt{m}_3\kappa_1}{C_1}. \quad (3.117)$$

Now define $P_i'' = \{j|\ |V_j'x_i'| \geq |V_j^{(0)}\phi^{(0)}(x_i)|/3\}$. If $j \in P_i'' - P_i'$, then by Equation (3.117), with high probability

$$\|V_j'\|\|\phi^{(2)}(x_i)\| \geq |V_j'\phi^{(2)}(x_i)| = |V_j'(\phi^{(0)}(x_i) + \phi^{(2)}(x_i))| = |V_j'x_i'|$$

$$\geq |V_j^{(0)}\phi^{(0)}(x_i)|/3 \gtrsim c_3(\frac{\kappa_2}{m_2})^{1/3}C_2^{2/3}\|\phi^{(0)}(x_i)\|,$$

or

$$\|V_j'\|^2 \gtrsim c_3^2(\frac{\kappa_2}{m_2})^{2/3}C_2^{4/3}\frac{m_3\kappa_1^2}{C_1^2}.$$

But note that $\|V'\|_F^2 \leq C_2^2$ by our assumption, which implies

$$|P_i'' - P_i'| \lesssim C_2^2/[c_3^2(\frac{\kappa_2}{m_2})^{2/3}C_2^{4/3}\frac{m_3\kappa_1^2}{C_1^2}] = \frac{C_2^{2/3}C_1^2}{c_3^2m_3\kappa_1^2}(\frac{m_2}{\kappa_2})^{2/3}. \quad (3.118)$$

Now combining Equations (3.114) and (3.118), we finally obtain

$$|P_i''| = |P_i'' - P_i'| + |P_i'| \lesssim (c_3 + \frac{C_1^2}{c_3^2m_3\kappa_1^2})C_2^{2/3}(\frac{m_2}{\kappa_2})^{2/3}.$$

Now define the set

$$P_i''' = \{j|\ |V_j^{(0)}\phi^{(2)}(x_i)| \geq |V_j^{(0)}\phi^{(0)}(x_i)|/3\}. \tag{3.119}$$

Note that for every $j \in [m_2]$, $V_j^{(0)}\phi^{(0)}(x_i)$ is gaussian with variance $\|\phi^{(0)}(x_i)\|$ over the randomness of $V_j^{(0)}$, so

$$\mathbb{P}(V_j^{(0)}\phi^{(0)}(x_i) \leq \alpha\kappa_2\|\phi^{(0)}(x_i)\|) \lesssim \alpha.$$

Therefore, if we define the set

$$Q_i = \{j \in [m_2]|\ |V_j^{(0)}\phi^{(0)}(x_i)| \leq \alpha\kappa_2\|\phi^{(0)}(x_i)\|\},$$

then for large enough $m_2$, by Bernstein with high prob.:

$$|Q_i| \lesssim \alpha m_2. \tag{3.120}$$

Now note that $\phi^{(0)}(x_i)$ is fixed during the algorithm. On the other hand, by random matrix theory, we know that with high probability, the eigenvalues of the matrix $V^{(0)}$ are in $(\kappa_2(\sqrt{m_2} - \sqrt{m_3}), \kappa_2(\sqrt{m_2} + \sqrt{m_3}))$. Therefore, even if the vector $\phi^{(2)}(x_i)$ is picked adversarialy (because it keeps changing during the algorithm), we get that with high probability over the randomness of $V^{(0)}$:

$$|V^{(0)}\phi^{(2)}(x_i)\|^2 \leq \kappa_2^2(\sqrt{m_2} + \sqrt{m_3})^2\|\phi^{(2)}(x_i)\|^2 \lesssim \kappa_2^2 m_2\|\phi^{(2)}(x_i)\|^2. \tag{3.121}$$

Moreover, because $\|\phi^{(2)}(x_i)\| \lesssim C_1$ and from Equation (3.116), with high probability over the randomness of $W^{(0)}$:

$$\frac{\|\phi^{(0)}(x_i)\|}{\|\phi^{(2)}(x_i)\|} \gtrsim \frac{\sqrt{m_3}\kappa_1}{C_1}.$$

This means that for $j \in P_i''' - Q_i$, combining these inequalities we conclude with high probability

$$|V_j^{(0)}\phi^{(2)}(x_i)| \geq |V_j^{(0)}\phi^{(0)}(x_i)|/3 \gtrsim \alpha\kappa_2\|\phi^{(0)}(x_i)\| \geq \alpha\kappa_2\frac{\sqrt{m}_3\kappa_1}{C_1}\|\phi^{(2)}(x_i)\|,$$

which combined with (3.121) implies

$$\|P_i'''\| \lesssim \frac{m_2C_1^2}{m_3\kappa_1^2\alpha^2}.$$

Balancing this term with the one in Equation (3.120), we set

$$\alpha := \frac{C_1^{2/3}}{m_3^{1/3}\kappa_1^{2/3}},$$

which implies

$$|P_i'''| \lesssim |P_i''' - Q_i| + |Q_i| \leq \big(\frac{C_1^2}{(m_3\kappa_1^2)}\big)^{1/3}m_2.$$

Defining $\tilde{P}_i = P_i'' \cup P_i'''$, we finally get

$$|\tilde{P}_i| \lesssim \Big(\big(\frac{C_1^2}{(m_3\kappa_1^2)}\big)^{1/3} + (c_3 + \frac{C_1^2}{c_3^2m_3\kappa_1^2})C_2^{2/3}(\frac{1}{\kappa_2^2m_2})^{1/3}\Big)m_2.$$

Clearly by the definition of $P_i''$ and $P_i'''$ the proof is complete.

**Corollary 5.1.** *Under the condition $\|\phi^{(2)}(x_i)\| \lesssim C_1$ (which happens under the event $E^c$ defined in Lemma 42), setting $c_3 := \frac{C_1^{2/3}}{m_3^{1/3}\kappa_1^{2/3}}(\frac{\kappa_2^2m_2}{C_2^2})^{1/3}$ in the previous Lemma, we obtain $\forall i \in [n]$:*

$$|\tilde{P}_i| \lesssim \big(\frac{C_1^2}{(m_3\kappa_1^2)}\big)^{1/3}m_2.$$

*Also for $j \notin \tilde{P}_i$, the conditions in (3.113) and (3.119) becomes the same as*

$$|W_j^{(0)}\phi^{(0)}(x_i)| \leq \kappa_2\frac{C_1^{2/3}}{m_3^{1/3}\kappa_1^{2/3}}\|\phi^{(0)}(x_i)\|.$$

*Hence, for every $i \in [n]$ and for $j \notin \tilde{P}_i$, with high probability:*

$$\frac{2}{3}|W_j^{(0)}\phi^{(0)}(x_i)| \geq |W_j^{(0)}\phi^{(2)}(x_i)| + |W_j'(\phi^{(0)}(x_i) + \phi^{(2)}(x_i))|, \tag{3.122}$$

$$|W_j^{(0)}\phi^{(0)}(x_i)| \gtrsim \kappa_2 \frac{C_1^{2/3}}{m_3^{1/3}\kappa_1^{2/3}}\|\phi^{(0)}(x_i)\| \gtrsim \kappa_2(\sqrt{m}_3\kappa_1 C_1^2)^{1/3}, \tag{3.123}$$

$$|W_j^{(0)}\phi^{(0)}(x_i)| \gtrsim \frac{C_1^{2/3}}{m_3^{1/3}\kappa_2^{2/3}}\|x_i'\|. \tag{3.124}$$

Next, we state concentration result for the gram matrix of $\phi^{(0)}(x_i)$'s.

**Lemma 17.** *For every $i_1, i_2 \in [n]$, with high probability over the randomness of $W^{(0)}$ and $V^{(0)}$ we have*

$$\langle \phi^{(0)}(x_{i_1}), \phi^{(0)}(x_{i_2})\rangle = m_3\mathbb{E}\sigma(W_j^{(0)}x_{i_1})\sigma(W_j^{(0)}x_{i_2}) \pm O(m_3\kappa_1^2/\sqrt{m}_1 + \sqrt{m}_3\kappa_1^2).$$

**Proof of Lemma 17**

First, we compute the expectation:

$$\mathbb{E}\langle \phi^{(0)}(x_{i_1}), \phi^{(0)}(x_{i_2})\rangle = 1/m_1 \sum_{j_1, j_2 \in [m_1]} \sum_{k \in [m_3]} \mathbb{E}W_{k,j_1}^s W_{k,j_2}^s \sigma(W_{j_1}^{(0)}x_{i_1})\sigma(W_{j_2}^{(0)}x_{i_2})$$

$$= 1/m_1 \sum_{j_1 \neq j_2} \mathbb{E} \sum_{k \in [m_3]} W_{k,j_1}^s W_{k,j_2}^s \sigma(W_{j_1}^{(0)}x_{i_1})\sigma(W_{j_2}^{(0)}x_{i_2}) + m_3/m_1 \sum_{j \in [m_1]} \sigma(W_j^{(0)}x_{i_1})\sigma(W_j^{(0)}x_{i_2}).$$

But $\sigma(W_{j_1}^{(0)}x_{i_1})\sigma(W_{j_2}^{(0)}x_{i_2})$ is $(m_1\kappa_1^4, \kappa_1^2)$-sub-exponential, so

$$\sum_{j \in [m_1]} \sigma(W_j^{(0)}x_{i_1})\sigma(W_j^{(0)}x_{i_2}) = m_1\mathbb{E}\sigma(W_j^{(0)}x_{i_1})\sigma(W_j^{(0)}x_{i_2}) \pm O(\sqrt{m}_1\kappa_1^2),$$

which means with high probability:

$$\mathbb{E}\langle \phi^{(0)}(x_{i_1}), \phi^{(0)}(x_{i_2})\rangle = m_3\mathbb{E}\sigma(W_j^{(0)}x_{i_1})\sigma(W_j^{(0)}x_{i_2}) \pm O(m_3\kappa_1^2/\sqrt{m}_1).$$

On the other side, we know that $\phi_k^{(0)}(x_{i_1})$ is subgaussian with parameters $\sigma^2 = 1/m_1\sum_j(W_j^{(0)}x_{i_1})^2 := \aleph_1$ and $\sigma^2 = 1/m_1\sum_j(W_j^{(0)}x_{i_2})^2 := \aleph_2$ respectively. On the other hand, we know that by Bernstein w.h.p

$$
\begin{aligned}
\aleph_1 &= 1/2\kappa_1^2(1 \pm O(1/\sqrt{m_1})),\\
\aleph_2 &= 1/2\kappa_1^2(1 \pm O(1/\sqrt{m_1})).
\end{aligned}
$$

Hence, $\phi_k^{(0)}(x_{i_1})\phi_k^{(2)}(x_{i_2})$ is $(\aleph_1\aleph_2, \sqrt{\aleph_1\aleph_2})$-subexponential, and so $\langle\phi^{(0)}(x_{i_1}), \phi^{(0)}(x_{i_2})\rangle$ is $(m_3\aleph_1\aleph_2, \sqrt{\aleph_1\aleph_2})$-subexponential. Therefore, applying another Bernstein on the top, we get

$$
\begin{aligned}
\langle\phi^{(0)}(x_{i_1}), \phi^{(0)}(x_{i_2})\rangle &= \mathbb{E}\langle\phi^{(0)}(x_{i_1}), \phi^{(0)}(x_{i_2})\rangle \pm O(\sqrt{m_3}\sqrt{\aleph_1\aleph_2})\\
&= m_3\mathbb{E}\sigma(W_j^{(0)}x_{i_1})\sigma(W_j^{(0)}x_{i_2}) \pm O(m_3\kappa_1^2/\sqrt{m_1}) \pm \frac{\sqrt{m_3}\kappa_1^2}{2}(1 \pm O(1/\sqrt{m_1}))\\
&= m_3\mathbb{E}\sigma(W_j^{(0)}x_{i_1})\sigma(W_j^{(0)}x_{i_2}) \pm O(m_3\kappa_1^2/\sqrt{m_1} + \sqrt{m_3}\kappa_1^2).
\end{aligned}
$$

Now we define the matrix $L_i \in \mathbb{R}^{m_3 \times m_2}$, with its $j$th column $L_{i,j}$ equal to

$$
\frac{a_j}{\sqrt{m_2}}\mathbb{1}\{V_j^{(0)}\phi^{(0)}(x_i) \geq 0\}\phi^*(x_i).
$$

First, we state the following lemma which characterize a concentration result for the gram matrix of $(L_i)_{i=1}^n$.

**Lemma 18.** *With high probability, we have the following approximation:*

$$
\langle L_{i_1}, L_{i_2}\rangle = \langle\phi^*(x_{i_1}), \phi^*(x_{i_2})\rangle\Big[F_2\Big(2F_3(\langle x_{i_1}, x_{i_2}\rangle)\Big) \pm O(m_1^{-1/4} + m_2^{-1/4} + m_3^{-1/4})\Big].
$$

**Proof of Lemma 18**

By Hoeffding:

$$
\begin{aligned}
\langle L_{i_1}, L_{i_2}\rangle &= 1/m_2\sum_{j\in m_2}\phi^*(x_{i_1})^T\phi^*(x_{i_2})\mathbb{1}\{V_j^{(0)}\phi^{(1)}(x_{i_1}) \geq 0\}\mathbb{1}\{V_j^{(0)}\phi^{(1)}(x_{i_2}) \geq 0\}\\
&= \phi^*(x_{i_1})^T\phi^*(x_{i_2})\Big(\mathbb{E}\mathbb{1}\{V_j^{(0)}\phi^{(1)}(x_{i_1}) \geq 0\}\mathbb{1}\{V_j^{(0)}\phi^{(1)}(x_{i_2}) \geq 0\} \pm O(1/\sqrt{m_2})\Big)\\
&= \phi^*(x_{i_1})^T\phi^*(x_{i_2})\Big(F_2\Big(\langle\phi^{(1)}(x_{i_1}), \phi^{(1)}(x_{i_2})\rangle/(\|\phi^{(1)}(x_{i_1})\|\|\phi^{(1)}(x_{i_2})\|)\Big) \pm O(1/\sqrt{m_2})\Big),
\end{aligned}
$$

where recall

$$F_2(x) = 1/4 + \arcsin(x)/2\pi,$$

measures the angle between two unit vectors based on their dot product. Now notice that according to Lemma 17, with high probability:

$$\begin{aligned}
&\langle L_{i_1}, L_{i_2}\rangle / \langle \phi^*(x_{i_1}), \phi^*(x_{i_2})\rangle \\
&= F_2\Big(\frac{m_3 \mathbb{E}\sigma(W_j^{(0)} x_{i_1})\sigma(W_j^{(0)} x_{i_2}) \pm O((m_3/\sqrt{m}_1 + \sqrt{m}_3)\kappa_1^2)}{\sqrt{(m_3 \mathbb{E}\sigma(W_j^{(0)} x_{i_1})^2 \pm O((m_3/\sqrt{m}_1 + \sqrt{m}_3)\kappa_1^2))(m_3 \mathbb{E}\sigma(W_j^{(0)} x_{i_2})^2 \pm O(...)}} \pm O(1/\sqrt{m}_2)\Big) \\
&= F_2\Big(\frac{F_3(\langle x_{i_1}, x_{i_2}\rangle) \pm O(1/\sqrt{m}_1 + 1/\sqrt{m}_3)}{1/2 \pm O(1/\sqrt{m}_1 + 1/\sqrt{m}_3)} \pm O(1/\sqrt{m}_2)\Big),
\end{aligned}$$

where recall $F_3 : [-1, +1] \to [-1/2, 1/2]$ is defined as:

$$F_3(x) \coloneqq \frac{\sqrt{1 - x^2}}{2\pi} + \frac{x}{4} + \frac{x \arcsin x}{2\pi}.$$

It is easy to see $F_3$ has the property that for unit vectors $x_1, x_2$ and $w$ sampled as standard normal:

$$F_3(\langle x_1, x_2\rangle) = \mathbb{E}\sigma(w^T x_1)\sigma(w^T x_2).$$

But because $|F_3(.)| = O(1)$, we have

$$\langle L_{i_1}, L_{i_2}\rangle / \langle \phi^*(x_{i_1}), \phi^*(x_{i_2})\rangle = F_2\Big(2F_3(\langle x_{i_1}, x_{i_2}\rangle) \pm O(1/\sqrt{m}_1 + 1/\sqrt{m}_2 + 1/\sqrt{m}_3)\Big).$$

Now notice that the derivative of $F_2$, i.e. $1/2\pi\sqrt{1 - x^2}$ is increasing in the interval $(0, 1)$, so for a fixed $\delta$, the maximum of $|F_2(x) - F_2(x - \delta)|$ happens at $x = 1$. On the other hand, by writing the first order approximation of $\arcsin(1 - t^2)$ around $t = 0$ and upper bounding its derivative in the interval $[0, 1]$, we get that for $0 \le \delta \le 1$:

$$\arcsin(1 - \delta) \ge \arcsin(1) - 2\sqrt{\delta}.$$

Therefore, $F_2(x \pm \delta) = F_2(x) \pm O(\sqrt{\delta})$. Hence:

$$\begin{aligned}\langle L_{i_1}, L_{i_2}\rangle / \langle \phi^*(x_{i_1}), \phi^*(x_{i_2})\rangle &= F_2\Big(2F_3(\langle x_{i_1}, x_{i_2}\rangle)\Big) \pm O(\sqrt{1/\sqrt{m}_1 + 1/\sqrt{m}_2 + 1/\sqrt{m}_3}) \\ &= F_2\Big(2F_3(\langle x_{i_1}, x_{i_2}\rangle)\Big) \pm O(m_1^{-1/4} + m_2^{-1/4} + m_3^{-1/4}),\end{aligned}$$

which completes the proof.

Finally, we are ready to construct the weights $V^*$ for the second layer.

### Construction of $V^*$

**Lemma 19.** *Let*

$$\Re = \frac{n\sqrt{m}_3}{\sqrt{m}_1 \lambda_0}\Big[1 + \frac{nC_1}{\kappa_1}\Big].$$

*Suppose we have the condition that for every $k \in [m_3]$:*

$$\max_k \|\mathcal{V}_k\| \leq \xi, \tag{3.125}$$

*where recall the definition of $\mathcal{V}_k$ in Equation (3.59). We assume enough overparameterization to make sure $\Re < 1$. Recall for the matrix $A$ defined by*

$$A = \Big(\langle \bar{x}_{i_1}, \bar{x}_{i_2}\rangle F_2(2F_3(\langle x_{i_1}, x_{i_2}\rangle))\Big)_{1\leq i_1, i_2 \leq n}, \tag{3.126}$$

*we have*

$$(f^*(x_i))_{i=1}^n{}^T A^{-1} (f^*(x_i))_{i=1}^n \leq \zeta_2.$$

*Then, there exists weight matrix $V^*$ which only depends on the random initializations $W^{(0)}, V^{(0)}$ (e.g. not on $V'$ and $W'$) for the second layer, such that having enough overparameterization*

$$\|V^*\|_F^2 \leq 2\zeta_2, \tag{3.127}$$

*and for every $j \in [m_2]$:*

$$\|V_j^*\|_\infty \leq \frac{(1+\Re)n\sqrt{n\zeta_2}}{\sqrt{m}_2}\xi := \varrho_3\xi/\sqrt{m}_2, \tag{3.128}$$

$$\|V_j^*\|_2 \leq \frac{1}{\sqrt{m}_2}n(1+\Re)\sqrt{\frac{\zeta_2\sum_k\|\mathcal{V}_k\|_{H^\infty}^2}{\lambda_0}} := \varrho_2/\sqrt{m}_2, \tag{3.129}$$

*and further under the high probability event $E^c$ defined in Lemma 42:*

$$|\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V',x_i}V^*\phi^*(x_i) - f^*(x_i)| \lesssim (\frac{C_1}{\sqrt{m}_3\kappa_1})^{1/3}(1+\Re)\sqrt{\zeta_2\sum_k\|\mathcal{V}_k\|_{H^\infty}^2} := \Re_3. \tag{3.130}$$

**Proof of Lemma 19**

Let

$$V^* = \sum_{i=1}^n \mathcal{V}_i^* L_i,$$

be the minimum norm vector which maps $L_i$'s to $f^*(x_i)$'s. As a result, for the matrix

$$L = \Big(\langle L_{i_1}, L_{i_2}\rangle\Big)_{i_1,i_2}$$

it is easy to see

$$\|V^*\|_F^2 = (f^*(x_i))_{i=1}^n{}^T L^{-1}(f^*(x_i))_{i=1}^n.$$

Now combining Lemmas 14 and 50, we get

$$\|\phi^*(x_i)\|_\infty \leq (1+\Re)\xi, \tag{3.131}$$

$$\|\phi^*(x_i)\| \leq (1+\Re)\sqrt{\sum_k\|\mathcal{V}_k\|_{H^\infty}^2}, \tag{3.132}$$

and

$$\langle|\phi^*(x_{i_1}), \phi^*(x_{i_2})\rangle - \langle\bar{x}_{i_1}, \bar{x}_{i_2}\rangle| \leq (2\Re + \Re^2)\sum_k\|\mathcal{V}_k\|_{H^\infty}^2.$$

Now by Lemma 18:

$$|\langle L_{i_1}, L_{i_2}\rangle - A_{i_1,i_2}| \lesssim (2\Re + \Re^2)(\sum_k \|\mathcal{V}_k\|_{H^\infty}^2)\Big|F_2\Big(2F_3(\langle x_{i_1}, x_{i_2}\rangle)\Big)\Big| + \langle \bar{x}_{i_1}, \bar{x}_{i_2}\rangle(m_1^{-1/4} + m_2^{-1/4} + m_3^{-1/4})$$
$$+ \text{ other cross term.}$$

By applying $\langle \bar{x}_{i_1}, \bar{x}_{i_2}\rangle \leq \|\bar{x}_{i_1}\|\|\bar{x}_{i_2}\|$ we get

$$\begin{aligned} LHS &\leq (\sum_k \|\mathcal{V}_k\|_{H^\infty}^2)\Big((2\Re + \Re^2)\Big|F_2\Big(2F_3(\langle x_{i_1}, x_{i_2}\rangle)\Big)\Big| + (m_1^{-1/4} + m_2^{-1/4} + m_3^{-1/4})\Big) \\ &\lesssim (\sum_k \|\mathcal{V}_k\|_{H^\infty}^2)\Big(\Re + m_1^{-1/4} + m_2^{-1/4} + m_3^{-1/4}\Big). \end{aligned}$$

Therefore,

$$\begin{aligned} \|A - \Big(\langle L_{i_1}, L_{i_2}\rangle\Big)_{i_1,i_2}\|_2 &\leq \|A - \Big(\langle L_{i_1}, L_{i_2}\rangle\Big)_{i_1,i_2}\|_F \\ &\leq n(\sum_k \|\mathcal{V}_k\|_{H^\infty}^2)\Big(\Re + m_1^{-1/4} + m_2^{-1/4} + m_3^{-1/4}\Big) := \Re_2. \end{aligned}$$

Note that $\Re_2$ naturally goes to zero (with poly dependence) as $\Re \to 0$ and $m_1, m_2, m_3$ are large enough. Now if all of the eigenvalues of the matrix $A$ are $\Omega(1/n^2)$, then if we overparameterize enough such that $\Re_2 = O(1/n^2)$ with small enough constant so that $\Re_2$ is less than half of the smallest eigenvalue of $A$, then for the $i$th eigenvalue $\lambda_i$ of $A$ and $L$ we can write

$$\lambda_i(L) \geq \lambda_i(A) - \Re_2 \geq \lambda_i(A)/2,$$

so

$$\lambda_i(L^{-1}) \leq 2\lambda_i(A^{-1}),$$

which implies the property

$$\|V^*\|_F^2 \leq 2\zeta_2. \tag{3.133}$$

However, $A$ might have very small eigenvalues. To remedie this, we use Lemma 51; we can substitute $f^*$ with some $\bar{f}^*$ such that

$$R_n(\bar{f}^*) \leq 2R_n(f^*) + \frac{B^2}{n}, \tag{3.134}$$

$$\bar{f^*}^T A^{-1}\bar{f^*} \leq {f^*}^T A^{-1} f^*, \tag{3.135}$$

where $\bar{f}^*$ is on the subspace of eigenvectors of $A$ whose eigenvalues are larger than $\Omega(1/n^2)$. But it is easy to check that in the context of Theorem 3, such substitution results in a $\bar{f}^{*T}A^{-1}\bar{f}^* \leq {f^*}^T A^{-1} f^* \leq \zeta$ and $\bar{v}(\bar{f}^*)$ parameter (as defined in (3.41)) with respect to $\bar{f}^*$ which satisfies $\bar{\nu}/2 \leq \nu$. Note that the algorithm is with respect to the setting $\nu$, however we want to exploit generalization bound with respect to $\bar{f}^*$ whose parameter is $\bar{\nu}$ as it enables us to use our analysis in this Lemma. Furthermore, note that using Equation (3.134) we can further upper bound the empirical risk of $\bar{f}^*$ with that of $f^*$, which makes it straightforward to derive a similar generalization bound as in (3.45) with respect to $f^*$, of course with a change of constants. Note that $\bar{f}^*$ is just the sum of $A$-eigenbasis directions in $f^*$ whose eigenvalues are larger than $\Omega(1/n^2)$. Hence, given a pair $(f^*, G)$, as we also point out in remark 1, we can construct the suitable pair $(\bar{f}^*, G)$ algorithmically and then use that pair to initialize the parameters of the algorithm (namely $\zeta$ and $\nu$). Otherwise, if we are not explicitly given a pair $(f^*, G)$ and instead want to run the doubling trick described in Theorem 1, we do not even have any additional computation; since using Theorem 1, within the framework of the doubling trick, the risk of the final network is competitive with respect to any choice of $(\bar{f}^*, G)$. Note that as we mentioned in Lemma 51, the constant 2 is arbitrary and can be reduced to any number less than 2, and it is easy to see that one can pick choice of constants along the way such that we end up with a factor two behind the risk (first) term in the definition of our complexity measure.

Therefore, without loss of generality we can use substitute $f^*$ by $\bar{f}^*$ and still obtain Equation (3.133).

On the other hand, the definition of $V^*$ implies

$$\frac{1}{\sqrt{m_2}} a^T D_{V^{(0)},x_i} V^* \phi^*(x_i) = f^*(x_i).$$

But note that by Corollary 5.1, under the high probability event $E^c$ defined in Lemma 42, $D_{V^{(0)},x_i}$ and $D_{V^{(0)}+V',x_i}$ can only be different in the index set $\tilde{P}_i$ and

$$|\tilde{P}_i| \lesssim \big(\frac{C_1^2}{(m_3\kappa_1^2)}\big)^{1/3} m_2,$$

Therefore, for all $i \in [n]$:

$$\begin{aligned}
&|\frac{1}{\sqrt{m_2}} a^T D_{V^{(0)}+V',x_i} V^* \phi^*(x_i) - \frac{1}{\sqrt{m_2}} a^T D_{V^{(0)},x_i} V^* \phi^*(x_i)| \\
&\le 1/\sqrt{m_2} \sum_{j\in\tilde{P}} |V_j^* \phi^*(x_i)| \\
&\le 1/\sqrt{m_2} \sum_{j\in\tilde{P}} \|V_j^*\| \|\phi^*(x_i)\| \\
&\le \frac{\sqrt{|\tilde{P}|}}{\sqrt{m_2}} \|V^*\| (1+\Re) \sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2} \\
&\lesssim \big(\frac{C_1}{\sqrt{m_3}\kappa_1}\big)^{1/3} \sqrt{\zeta_2} (1+\Re) \sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2},
\end{aligned}$$

which proves the first claim. On the other hand, we get:

$$2\zeta_2 \ge \|V^*\|_F^2 \ge \mathcal{V}^T L \mathcal{V}.$$

But because $\lambda_{min}(L) \gtrsim 1/n^2$, we get

$$\mathcal{V}^T L \mathcal{V} \gtrsim \|\mathcal{V}\|_2^2 / n^2,$$

which implies

$$\|\mathcal{V}\|_2 \lesssim n\sqrt{\zeta_2}.$$

But now using Equation (3.131), we can write

$$\begin{aligned}|V^*_{j,k}| &\leq \sum_{i=1}^{n} |\mathcal{V}_i||L_{ij,k}| \leq \frac{1}{\sqrt{m_2}} \|\phi^*(x_i)\|_\infty \sum_i |\mathcal{V}_i| \\ &\lesssim \frac{(1+\Re)\xi}{\sqrt{m_2}} \sum_i |\mathcal{V}_i| \leq (1+\Re)\xi\sqrt{n}\|\mathcal{V}\|/\sqrt{m_2} \\ &\lesssim \frac{(1+\Re)n\sqrt{n\zeta_2}}{\sqrt{m_2}}\xi,\end{aligned}$$

which proves the other part. Moreover,

$$\|V^*_j\|^2 \leq \|\mathcal{V}\|^2 \frac{1}{\sqrt{m_2}} (\sum_i \|\phi^*(x_i)\|_2^2) \leq \frac{1}{\sqrt{m_2}} n(1+\Re)\sqrt{\zeta_2 \sum_k \|\mathcal{V}_k\|^2_{H^\infty}/\lambda_0}. \quad (3.136)$$

### 3.6.13 Existence of a good direction

Our aim in this section is to show that if the objective value is above certain threshold, there exists a good random direction which reduces the objective in expectation. Particularly our aim is to prove the following theorem (informal):

**Theorem 6.** *For a given pair $(f^*, G)$ with*

$$\langle H^\infty, G\rangle \leq \zeta_1,$$
$$f^{*T}(K^\infty \odot G)^{-1} f^* \leq \zeta_2,$$
$$R_n(f^*) \leq \Delta,$$

*recall the ideal random matrices $(W^*_\Sigma, V^*_\Sigma)$ constructed in Section 3.6.12, where $\Sigma$ is a random diagonal sign matrix. Specifically, $W^*_\Sigma$ is defined in Equation (3.108), and $V^*_\Sigma$ is the projection of the rows of matrix $V^*$ onto the orthogonal subspace spanned by $(\phi^{(0)}(x_i))_{i=1}^n$.*

*Using the parameter setting for $i = 1, 2$*

$$\psi_i = \frac{\nu}{4\zeta_i}, \tag{3.137}$$

*with respect to an arbitrary parameter $\nu > 0$, then for every pair $(W', V')$ such that $\|W'\| \leq C_1, \|V'\| \leq C_2$ and*

$$L(W', V') \geq \Delta + \nu, \tag{3.138}$$

*for parameters $m_1, m_2, m_3, 1/\kappa_1, 1/\kappa_2$ polynomially large enough in $B, 1/\lambda_0, n, C_1, C_2$ and small enough step size $\eta$, we have*

$$\mathbb{E}_\Sigma L(W' - \eta/2W' + \sqrt{\eta}W^*_\Sigma, V' - \eta/2V' + \sqrt{\eta}V^*_\Sigma) \leq L(W', V') - \eta\nu/4. \tag{3.139}$$

In order to prove the above theorem, we first state and prove the following lemma which is the core of Theorem 6.

**Lemma 20.** *For matrices $(W^*, V^*)$ constructed in Section 3.6.12, specifically for their random coupling $(W^*_\Sigma, V^*_\Sigma)$ as denoted above, we have:*

$$\mathbb{E}_\Sigma \ell(f'_{(1-\eta/2)W'+\sqrt{\eta}W^*_\Sigma,(1-\eta/2)V'+\sqrt{\eta}V^*_\Sigma}(x_i), y_i) \leq (1-\eta)\ell(f'_{W',V'}(x_i), y_i) + \eta\ell(f^*(x_i), y_i) \pm \eta\wp,$$

*where $\wp$ goes to zero with polynomially large overparameterization (the exact dependence is revealed via the proof).*

### Proof of Lemma 20

For brevity, we use the notation $D_{',\rho}$ here to refer to the diagonal binary sign matrix when the input is multiplied by the sum of weight and smoothing matrices. It will be clear in the context of the equation that what the "input" and the "weight" matrices are. This notation is also defined and used in Lemma 37). Here, we bound multiple cross terms that are created as a result of moving in the random direction. To simplify the presentation and avoid confusing recursions in the proof, we have made a sublemma for each of these cross terms and has deferred its proof to Section 3.6.14. We use difference sub-indices of the symbol $\Re$ to illustrate terms that go to zero by growing the overparameterization in our architecture.

We start by using Lemma 37,

$$\begin{aligned}
&\mathbb{E}_\Sigma \ell(f'_{(1-\eta/2)W'+\sqrt{\eta}W^*_\Sigma,(1-\eta/2)V'+\sqrt{\eta}V^*_\Sigma}(x_i), y_i) \\
=&\mathbb{E}_\Sigma \ell(\mathbb{E}_{W^\rho,V^\rho} f_{(1-\eta/2)W'+\sqrt{\eta}W^*_\Sigma+W^\rho,(1-\eta/2)V'+\sqrt{\eta}V^*_\Sigma+V^\rho}(x_i), y_i) \\
=&\mathbb{E}_\Sigma \ell\Big(\mathbb{E}_{W^\rho,V^\rho} a^T D_{',\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho + \sqrt{\eta}V^*_\Sigma)W^s \\
&D_{',\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho + \sqrt{\eta}W^*_\Sigma)x_i + \Re_8\eta, y_i\Big) \\
=&\mathbb{E}_\Sigma \ell\Big(\mathbb{E}_{W^\rho,V^\rho}\Big[a^T D_{',\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s D_{',\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i \\
&+\eta a^T D_{',\rho}V^*_\Sigma W^s D_{',\rho}W^*_\Sigma x_i\Big] + \sqrt{\eta}\mathbb{E}_{W^\rho,V^\rho}\Big[a^T D_{',\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s D_{',\rho}W^*_\Sigma x_i \\
&+a^T D_{',\rho}V^*_\Sigma W^s D_{',\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i\Big] \\
&+\Re_8\eta, y_i\Big).
\end{aligned}$$

Now using the notation introduced in Lemma 27, we have

$$W^s D_{\prime,\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i = \phi^{(0)}(x_i) + (1-\eta/2)\phi^{(2)}(x_i) + \frac{\eta}{2}\phi^{(2)\prime}(x_i).$$

By Lemma 27, we have the following bound for $\phi^{(2)\prime}(x_i)$:

$$\begin{aligned}&\mathbb{E}_{W^\rho,V^\rho}\Big|\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)} + V^\rho + (1-\eta/2)V')\phi^{(2)\prime}(x_i)\Big| \\ &\quad\lesssim (\kappa_2\sqrt{m_2 m_3} + \sqrt{m}_3\beta_2 + C_2)\Re_5.\end{aligned}$$

Therefore, Combining this with Lemma 29, we get

$$\begin{aligned}
&= \mathbb{E}_\Sigma \ell\Big(\mathbb{E}_{W^\rho,V^\rho}\Big[a^T D_{\prime,\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s(\phi^{(0)}(x_i) + (1-\eta/2)\phi^{(2)}(x_i)) \\
&+ \eta a^T D_{\prime,\rho}V^*_\Sigma W^s D_{\prime,\rho}W^*_\Sigma x_i\Big] + \sqrt{\eta}\mathbb{E}_{W^\rho,V^\rho}\Big[a^T D_{\prime,\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s D_{\prime,\rho}W^*_\Sigma x_i \\
&+ a^T D_{\prime,\rho}V^*_\Sigma W^s D_{\prime,\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i\Big] \\
&\pm O((\kappa_2\sqrt{m_2 m_3} + \sqrt{m}_3\beta_2 + C_2)\Re_5\eta) \pm O(\Re_8\eta), y_i\Big) \\
&= \mathbb{E}_\Sigma \ell\Big(\mathbb{E}_{W^\rho,V^\rho}\Big[(1-\eta)a^T D_{\prime,\rho}(V^{(0)} + V' + V^\rho)W^s(\phi^{(0)}(x_i) + \phi^{(2)}(x_i)) \\
&+ \eta a^T D_{\prime,\rho}V^*_\Sigma W^s D_{\prime,\rho}W^*_\Sigma x_i\Big] + \sqrt{\eta}\mathbb{E}_{W^\rho,V^\rho}\Big[a^T D_{\prime,\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s D_{\prime,\rho}W^*_\Sigma x_i \\
&+ a^T D_{\prime,\rho}V^*_\Sigma W^s D_{\prime,\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i\Big] \\
&\pm O(\eta(\Re_6' + \Re_4 + (\sqrt{m}_3\kappa_2 + \beta_2)(C_1 + \sqrt{m}_3\beta_1))) \\
&\pm O((\kappa_2\sqrt{m_2 m_3} + \sqrt{m}_3\beta_2 + C_2)\Re_5\eta) \pm O(\Re_8\eta), y_i\Big). \\
&= \mathbb{E}_\Sigma \ell\Big((1-\eta)f'_{W',V'}(x_i) + \eta\mathbb{E}_{W^\rho,V^\rho}a^T D_{\prime,\rho}V^*_\Sigma W^s D_{\prime,\rho}W^*_\Sigma x_i \\
&+ \sqrt{\eta}\mathbb{E}_{W^\rho,V^\rho}\Big[a^T D_{\prime,\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s D_{\prime,\rho}W^*_\Sigma x_i \\
&+ a^T D_{\prime,\rho}V^*_\Sigma W^s D_{\prime,\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i\Big] \\
&\pm O(\eta(\Re_6' + \Re_4 + (\sqrt{m}_3\kappa_2 + \beta_2)(C_1 + \sqrt{m}_3\beta_1))) \\
&\pm O((\kappa_2\sqrt{m_2 m_3} + \sqrt{m}_3\beta_2 + C_2)\Re_5\eta) \pm O(\Re_8\eta), y_i\Big).
\end{aligned}$$

Moreover, using the notation $\phi^{*\prime}(x_i)$ introduced in Lemma 24 and the bound in

Lemma 26, we can rewrite the second term as:

$$
\begin{aligned}
&LHS = \mathbb{E}_\Sigma \ell\Big((1-\eta)f'_{W',V'}(x_i) + \eta\mathbb{E}_{W^\rho,V^\rho} a^T D_{\prime,\rho} V^*_\Sigma(\phi^*(x_i) + \phi^{*\prime}(x_i)) \\
&+ \sqrt{\eta}\mathbb{E}_{W^\rho,V^\rho}\Big[a^T D_{\prime,\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s D_{\prime,\rho} W^*_\Sigma x_i \\
&+ a^T D_{\prime,\rho} V^*_\Sigma W^s D_{\prime,\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i\Big] \\
&\pm O(\eta(\Re'_6 + \Re_4 + (\sqrt{m}_3\kappa_2 + \beta_2)(C_1 + \sqrt{m}_3\beta_1))) \\
&\pm O((\kappa_2\sqrt{m_2 m_3} + \sqrt{m}_3\beta_2 + C_2)\Re_5\eta) \pm O(\Re_8\eta), y_i\Big) \\
&= \mathbb{E}_\Sigma \ell\Big((1-\eta)f'_{W',V'}(x_i) + \eta\mathbb{E}_{W^\rho,V^\rho} a^T D_{\prime,\rho} V^*_\Sigma \phi^*(x_i) \\
&+ \sqrt{\eta}\mathbb{E}_{W^\rho,V^\rho}\Big[a^T D_{\prime,\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s D_{\prime,\rho} W^*_\Sigma x_i \\
&+ a^T D_{\prime,\rho} V^*_\Sigma W^s D_{\prime,\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i\Big] \\
&\pm O(\eta\Re_{10}) \pm O(\eta(\Re'_6 + \Re_4 + (\sqrt{m}_3\kappa_2 + \beta_2)(C_1 + \sqrt{m}_3\beta_1))) \pm O((\kappa_2\sqrt{m_2 m_3} + \sqrt{m}_3\beta_2 + C_2)\Re_5\eta) \\
&\pm O(\Re_8\eta), y_i\Big). \qquad (3.140)
\end{aligned}
$$

Now we write the gradient-lipshitz inequality for $\ell$ at point

$$
p^{(1)}_\Sigma := (1-\eta)f'_{W',V'}(x_i) + \eta\mathbb{E}_{W^\rho,V^\rho} a^T D_{\prime,\rho} V^*_\Sigma \phi^*(x_i) \pm \eta\wp_1,
$$

and regarding the following vector, where $\wp_1$ is the sum of all the noise terms above and goes to zero by over parameterization:

$$
\begin{aligned}
p^{(2)}_\Sigma := &\sqrt{\eta}\mathbb{E}_{W^\rho,V^\rho}\Big[a^T D_{\prime,\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s D_{\prime,\rho} W^*_\Sigma x_i \\
&+ a^T D_{\prime,\rho} V^*_\Sigma W^s D_{\prime,\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i\Big].
\end{aligned}
$$

Hence, using the 1 smoothness of $\ell(., y_i)$:

$$
LHS \leq \mathbb{E}_\Sigma \ell\Big(p^{(1)}_\Sigma\Big) + \mathbb{E}_\Sigma \dot{\ell}(p)\sqrt{\eta} p^{(2)}_\Sigma + \frac{1}{2}\eta\mathbb{E}_\Sigma(p^{(2)}_\Sigma)^2. \qquad (3.141)
$$

But note that

$$\mathbb{E}_{\Sigma}\dot{\ell}(p)\sqrt{\eta}p_{\Sigma}^{(2)} = \dot{\ell}(p)\sqrt{\eta}\mathbb{E}_{\Sigma}p_{\Sigma}^{(2)} = 0. \tag{3.142}$$

On the other hand, using the notation of Lemma 24 and the result of Lemma 25:

$$\begin{aligned}
&\mathbb{E}_{\Sigma}\Big(\mathbb{E}_{W^\rho,V^\rho}a^T D_{',\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)W^s D_{',\rho}W_{\Sigma}^{*}x_i\Big)^2 \\
&= \mathbb{E}_{\Sigma}\Big(\mathbb{E}_{W^\rho,V^\rho}a^T D_{',\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)(\Sigma\phi^*(x_i) + {\phi^*}'_{\Sigma}(x_i))\Big)^2 \\
&\leq 4\mathbb{E}_{\Sigma}\Big(\mathbb{E}_{W^\rho,V^\rho}a^T D_{',\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho)\Sigma\phi^*(x_i)\Big)^2 \\
&+ 4\mathbb{E}_{\Sigma}\Big(\mathbb{E}_{W^\rho,V^\rho}a^T D_{',\rho}(V^{(0)} + (1-\eta/2)V' + V^\rho){\phi^*}'_{\Sigma}(x_i)\Big)^2 \\
&\lesssim \Re_{12}^2 + \Re_{11}^2.
\end{aligned} \tag{3.143}$$

Moreover, using again the result on $\phi^{(2)\prime}(x_i)$ from Lemma 27 and the fact that $\phi^{(0)}(x_i)$ is orthogonal to the rows of $V_{\Sigma}^*$:

$$\begin{aligned}
&\mathbb{E}_{W^\rho,V^\rho}a^T D_{',\rho}V_{\Sigma}^{*}W^s D_{',\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i \\
&= a^T D_{',\rho}V_{\Sigma}^{*}(\phi^{(0)}(x_i) + (1-\eta/2)\phi^{(2)}(x_i)) \\
&+ \frac{\eta}{2}a^T D_{',\rho}V_{\Sigma}^{*}\phi^{(2)\prime}(x_i) \\
&= (1-\frac{\eta}{2})a^T D_{',\rho}V_{\Sigma}^{*}\phi^{(2)}(x_i) \\
&+ \frac{\eta}{2}a^T D_{',\rho}V_{\Sigma}^{*}W^s D_{',\rho}\phi^{(2)\prime}(x_i) \\
&\lesssim (1-\frac{\eta}{2})a^T D_{',\rho}V_{\Sigma}^{*}\phi^{(2)}(x_i) \pm \Re_5.
\end{aligned}$$

Combining the last Equation with Lemma 23:

$$\begin{aligned}
&\mathbb{E}_{\Sigma}\Big(\mathbb{E}_{W^\rho,V^\rho}a^T D_{',\rho}V_{\Sigma}^{*}W^s D_{',\rho}(W^{(0)} + (1-\eta/2)W' + W^\rho)x_i\Big)^2 \\
&\lesssim (1-\frac{\eta}{2})^2 a^T D_{',\rho}V_{\Sigma}^{*}W^s D_{',\rho}\phi^{(2)}(x_i) + \Re_5^2 \\
&\lesssim \Re_0^2 + \Re_5^2.
\end{aligned} \tag{3.144}$$

Combining Equations (3.143) and (3.144):

$$\mathbb{E}_{\Sigma}(p_{\Sigma}^{(2)})^2 \lesssim \Re_{11}^2 + \Re_{12}^2 + \Re_0^2 + \Re_5^2 := \wp_2. \tag{3.145}$$

Combining Equations (3.142) and (3.145), plugging into (3.141), and reopening the definition of $p_{\Sigma}^{(1)}$:

$$LHS \lesssim \mathbb{E}_{\Sigma}\ell\Big((1-\eta)f'_{W',V'}(x_i) + \eta\mathbb{E}_{W^{\rho},V^{\rho}}a^T D_{',\rho}V_{\Sigma}^{*}\phi^{*}(x_i) \pm \eta\wp_1, y_i\Big) + \eta\mu_2\wp_2. \tag{3.146}$$

Now note that we can easily bound the magnitude of the term $\eta\mathbb{E}_{W^{\rho},V^{\rho}}a^T D_{',\rho}V_{\Sigma}^{*}\phi^{*}(x_i)$ as:

$$\begin{aligned}
&|\mathbb{E}_{W^{\rho},V^{\rho}}a^T D_{',\rho}V_{\Sigma}^{*}\phi^{*}(x_i)| \leq \mathbb{E}_{W^{\rho},V^{\rho}}|a^T D_{',\rho}V_{\Sigma}^{*}\phi^{*}(x_i)| \\
&\leq \|V_{\Sigma}^{*}\|_F\|\phi^{*}(x_i)\| \leq \|V^{*}\|\|\phi^{*}(x_i)\| \leq \sqrt{2\zeta_2}(1+\Re)\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2},
\end{aligned}$$

while using Lemma 43:

$$|f'_{W',V'}(x_i)| \leq (\kappa_2\sqrt{m}_3 + \beta_2)\Big(\sqrt{m}_3\kappa_1 + C_1 + \sqrt{m}_3\beta_1\Big) + C_2(C_1 + \sqrt{m}_3\beta_1),$$

which is $O(C_1C_2)$ for enough overparameterization and smoothing parameters $\beta_1, \beta_2$ as defined in 3.6.20. Furthermore, from Equations (3.132) and (3.127), we easily see that

$$\mathbb{E}_{W^{\rho},V^{\rho}}a^T D_{',\rho}V_{\Sigma}^{*}\phi^{*}(x_i) \leq \sqrt{2\zeta_2}(1+\Re)\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2}.$$

Now taking $\eta$ small enough so that the bound $\eta\sqrt{2\zeta_2}(1+\Re)\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2}$ and $\eta\wp_1$ both also be bounded of order $O(C_1C_2)$, we observe that the term inside the argument of $\ell(.,y_i)$ Equation in (3.146) is $O(C_1C_2)$. Hence, we can use the Lipschitz parameter of $\ell$ in the interval $[-O(C_1C_2), O(C_1C_2)]$, given by Lemma 9 to take out the noise term:

$$LHS \lesssim \mathbb{E}_\Sigma \ell\Big((1-\eta)f'_{W',V'}(x_i) + \eta \mathbb{E}_{W^\rho,V^\rho} a^T D_{',\rho} V^*_\Sigma \phi^*(x_i), y_i\Big) \pm \eta\wp_1 + \eta O(C_1C_2+B)\wp_2. \tag{3.147}$$

Now by applying Lemma 28 and writing the Lipchitz property of $\ell$ at point $(1-\eta)f'_{W',V'}(x_i) = O(C_1C_2)$:

$$\begin{aligned} LHS &\lesssim \mathbb{E}_\Sigma \ell\Big((1-\eta)f'_{W',V'}(x_i) + \eta f^*(x_i) \pm \eta\Re_9, y_i\Big) \pm \eta\wp_1 + \eta O(C_1C_2+B)\wp_2 \\ &\coloneqq \mathbb{E}_\Sigma \ell\Big((1-\eta)f'_{W',V'}(x_i) + \eta f^*(x_i), y_i\Big) \pm \eta\Re_9 \pm \eta\wp_1 + \eta O(C_1C_2+B)\wp_2 \\ &= \ell\Big((1-\eta)f'_{W',V'}(x_i) + \eta f^*(x_i), y_i\Big) \pm \eta\wp, \end{aligned}$$

where the last line is just definition. Now Convexity of $\ell$ finishes the proof.

Next, using Lemma 20 we prove Theorem 6.

**Restating Theorem 6** In the same setting as Theorem 6 and having enough overparameterization such that $\wp \le \frac{\nu}{8}$ ($\wp$ defined in Lemma 20) and polynomially small enough step size $\eta$, we have

$$\mathbb{E}_\Sigma L(W' - \eta W' + \sqrt{\eta} W^*_\Sigma, V' - \eta V' + \sqrt{\eta} V^*_\Sigma) \le L(W', V') - \eta\nu/4.$$

**Proof of Theorem 6**

First, note that taking expectation w.r.t $\Sigma$:

$$\begin{aligned} \mathbb{E}_\Sigma \|(1-\eta/2)W' + \sqrt{\eta}W^*_\Sigma\|^2 &= \mathbb{E}_\Sigma (1-\eta/2)^2\|W'\|^2 + 2(1-\eta/2)\sqrt{\eta}\langle W', \sum_{k=1}^{m_3} \Sigma_k W^*_k\rangle + \eta\|\sum_{k=1}^{m_3}\Sigma_k W^*_k\|^2 \\ &= (1-\eta/2)^2\|W'\|^2 + \eta\sum_k \|W^*_k\|^2, \end{aligned}$$

which by orthogonality of $W^*_k$'s:

$$LHS = (1-\eta/2)^2\|W'\|^2 + \eta\|W^*\|^2 = (1-\eta)\|W'\|^2 + \eta\|W^*\|^2 + \eta^2\|W'\|^2.$$

Similarly for $V'$:

$$\mathbb{E}_\Sigma\|(1-\eta/2)V'+\sqrt{\eta}V^*_\Sigma\| = (1-\eta/2)^2\|V'\|^2+\eta\mathbb{E}_\Sigma\|V^*\Sigma\|^2 = (1-\eta)\|V'\|^2+\eta\|V^*\|^2+\eta^2\|V'\|^2.$$

Now using Lemma 20:

$$\begin{aligned}
&\mathbb{E}_\Sigma L(W'-\eta/2W'+\sqrt{\eta}W^*_\Sigma, V'-\eta/2V'+\sqrt{\eta}V^*_\Sigma)\\
&\le (1-\eta)\mathbb{E}_\mathcal{Z}\ell(f'_{W',V'}(x),y)+\eta\mathbb{E}_\mathcal{Z}\ell(f^*(x),y)\\
&+(1-\eta)\Big(\psi_1\|W'\|^2+\psi_2\|V'\|^2\Big)+\eta\Big(\psi_1\|W^*\|^2+\psi_2\|V^*\|^2\Big)+\eta\Big(\wp+\eta(\|W'\|^2+\|V'\|^2)\Big)\\
&\le L(W',V')-\eta\Big(L(W',V')-\Delta-\psi_1\zeta_1-\psi_2\zeta_2\Big)+\eta\Big(\wp+\eta(\|W'\|^2+\|V'\|^2)\Big),
\end{aligned}$$

which by the choice of $\psi_i$'s is equal to

$$\begin{aligned}
LHS &\le L(W',V')-\eta\Big(L(W',V')-\Delta-\nu/2\Big)+\eta\Big(\wp+\eta(\|W'\|^2+\|V'\|^2)\Big)\\
LHS &\le L(W',V')-\eta\nu/2+\eta\Big(\wp+\eta(\|W'\|^2+\|V'\|^2)\Big).
\end{aligned}$$

Moreover, using the condition

$$\wp \le \nu/8,$$

and picking $\eta$ as small as

$$\eta(\|W'\|^2+\|V'\|^2) \le \eta(C_1^2+C_2^2) \le \nu/8,$$

we finally get

$$LHS \le L(W',V')-\eta\nu/4.$$

### 3.6.14 Existence of a good direction Helper Lemmas

In this section, we state and prove the core lemmas that are used in the proof of Lemma 20. Notably, through all of this section, we assume the norm bounds $\|W'\| \leq C_1, \|V'\| \leq C_2$ and that as our usual assumption, the rows of $V'$ are orthogonal to $\phi^{(0)}(x_i)$'s for all $i \in [n]$. A notation that we use throughout the proofs is $V^*_\Sigma$ which refers to the projectiono of $V^*\Sigma$ onto the orthogonal subspace to $(\phi^{(0)}(x_i))_{i=1}^n$.

**Lemma 21.** *Let $P(.)$ be the projection operator onto the subspace spanned by $(\phi^{(0)}(x_i))_{i=1}^n$. Also, we denote the projection of rows of $V^*\Sigma$ onto the orthogonal subspace to $(\phi^{(0)}(x_i))_{i=1}^n$ by ${V^*_\Sigma}_j$. Then*

$$\mathbb{E}_\Sigma \|{V^*_\Sigma}_j - V^*_j\Sigma)\|^2 \leq \varrho_3^2\xi^2 n/m_2,$$

*with high probability*

$$\|{V^*_\Sigma}_j - V^*_j\Sigma)\| \lesssim \frac{\varrho_3\xi\sqrt{n}}{\sqrt{m}_2},$$

**Proof of Lemma 21**

By Equation (3.129), we have $\|V^*_j\|_\infty \leq \varrho_3\xi/\sqrt{m_2}$. Now suppose that $u_1, ..., u_n$ are an orthonormal basis for the subspace $span(\phi^{(0)}(x_i))_{i=1}^n$. Then

$$\mathbb{E}_\Sigma \|{V^*_\Sigma}_j - V^*_j\Sigma)\|^2 = \mathbb{E}_\Sigma \|P(V^*_j\Sigma)\|^2 = \sum_i \sum_k {V^*_j}^2_k u_{ik}^2 \leq \|V^*_j\|^2_\infty n \leq \varrho_3^2\xi^2 n/m_2.$$

Also, by Hoeffding, with high probability:

$$\|P(V^*_j\Sigma)\|^2 = \sum_i (\sum_{k=1}^{m_3} {V^*_j}_k u_{ik}\Sigma_k)^2 \lesssim n\|V^*_j\|^2_\infty,$$

which implies the second part.

**Lemma 22.** *The first cross term goes away because of the definition of $V^*_\Sigma$. (inside the expectations is zero almost surely)*

$$\mathbb{E}_{\Sigma}\Big(\mathbb{E}_{V^{\rho},W^{\rho}}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^{\rho}+V',x_i} V_{\Sigma}^{*}\phi^{(0)}(x_i)]\Big)^2 = 0.$$

**Lemma 23.** *Second cross term:*

$$\mathbb{E}_{\Sigma}\Big(\mathbb{E}_{V^{\rho},W^{\rho}}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^{\rho}+V',x_i} V_{\Sigma}^{*}\phi^{(2)}(x_i)]\Big)^2 \tag{3.148}$$

$$\lesssim \xi^2((1+\Re)^2 n\zeta_2 + \varrho_3^2 n)(C_1^2 + m_3\beta_1^2) = \Re_0^2. \tag{3.149}$$

**Proof of Lemma 23**

This time we use Equation (3.129) in Lemma 19 and Lemma 21:

$$\begin{aligned}
&\mathbb{E}_{\Sigma}\Big(\mathbb{E}_{V^{\rho},W^{\rho}}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^{\rho}+V',x_i} V_{\Sigma}^{*}\phi^{(2)}(x_i)]\Big)^2\\
&\le \mathbb{E}_{\Sigma}\Big(\mathbb{E}_{V^{\rho},W^{\rho}} 1/\sqrt{m}_2 \sum_j |V_{\Sigma j}^{*}\phi^{(2)}(x_i)|\Big)^2\\
&\le \frac{1}{m_2}\mathbb{E}_{\Sigma,V^{\rho},W^{\rho}}\Big(\sum_j |V_{\Sigma j}^{*}\phi^{(2)}(x_i)|\Big)^2\\
&\le \mathbb{E}_{V^{\rho},W^{\rho}} E_{\Sigma}\sum_j |V_{\Sigma j}^{*}\phi^{(2)}(x_i)|^2\\
&\lesssim E_{V^{\rho},W^{\rho}} E_{\Sigma}\sum_j |(V_{\Sigma j}^{*} - V_j^{*}\Sigma)\phi^{(2)}(x_i)|^2 + \sum_j |V_j^{*}\Sigma\phi^{(2)}(x_i)|^2\\
&\lesssim \mathbb{E}_{V^{\rho},W^{\rho}} E_{\Sigma}\sum_j \|V_{\Sigma j}^{*} - V_j^{*}\Sigma)\|^2\|\phi^{(2)}(x_i)\|^2 + \sum_j \|V_j^{*}\|_{\infty}^2\|\phi^{(2)}(x_i)\|_2^2\\
&\lesssim ((1+\Re)^2 n\zeta_2\xi^2 + \varrho_3^2\xi^2 n)\mathbb{E}_{V^{\rho},W^{\rho}}\|\phi^{(2)}(x_i)\|_2^2.
\end{aligned}$$

Now according to Lemma 42, we have

$$\mathbb{E}_{V^{\rho},W^{\rho}}\|\phi^{(2)}(x_i)\|^2 \lesssim {C_1}^2 + m_3{\beta_1}^2,$$

which completes the proof.

**Lemma 24.** *We get an additional term $\phi^{*\prime}(x_i)$ as a result of smoothing which we* *define as*

$$\phi^{*\prime}(x_i) = \frac{1}{\sqrt{m_1}} W^s D_{W^{(0)}+W'+W^\rho,x_i} W^*_\Sigma x_i - \phi^*{}_\Sigma(x_i). \tag{3.150}$$

*Then*

$$\mathbb{P}(\phi^{*\prime}(x_i) \neq 0) \leq m_1 \exp\{-c_2^2/(8\beta_1^2)\}.$$

*Moreover, we have the following inequality almost surely (over the randomness of $W^\rho$):*

$$\|\phi^{*\prime}(x_i)\|_\infty \lesssim \sqrt{\sum_k \|\mathcal{V}_k\|^2_{H^\infty}}.$$

**Proof of Lemma 24**

According to Lemma 10, for $j \notin P$, for every $i \in [n]$ we have

$$|(W_j^{(0)} + W_j')x_i| \geq c_2/2\sqrt{m_1}.$$

Now note that as long as the sign patterns for $j \notin P$ does not change, $\phi^{*\prime}(x_i)$ will be zero. Therefore by union bound

$$\mathbb{P}(\phi^{*\prime}(x_i) \neq 0) \leq \sum_{j=1}^{m_1} \mathbb{P}(\text{sign change in } j) \leq m_1 \mathbb{P}(|(W_j^{(0)} + W_j')x_i| \leq |W_j^\rho x_i|)$$

$$\leq m_1 \mathbb{P}(|W_j^\rho x_i| \geq c_2/(2\sqrt{m_1})).$$

But $(W_j^\rho)x_i$ is Gaussian with variance $\beta_1^2/m_1$. Hence

$$LHS \lesssim m_1 \exp\{-c_2^2/(8\beta_1^2)\},$$

which proves the first part. For the second part, according to Equation (3.106) in Lemma 14, for every $k \in [m_3]$:

$$|{\phi^{*}}'_k(x_i)| \le |\frac{1}{\sqrt{m}_1} W_k^s D_{W^{(0)}+W'+W^{\rho},x_i} W^* x_i| + |{\phi^{*}}_k(x_i)| \tag{3.151}$$

$$\le 2/\sqrt{m}_1 \sum_j \|W_j^*\| \le 2\|W^*\|_F \lesssim \sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2}, \tag{3.152}$$

which implies the second part.

**Lemma 25.** *Fourth Extra term:*

$$\mathbb{E}_{W^{\rho},V^{\rho}}[\frac{1}{\sqrt{m}_2} a^T D_{V^{(0)}+V^{\rho}+V',x_i}(V^{(0)} + V^{\rho} + (1-\eta/2)V'){\phi^{*}}'(x_i)]$$

$$\lesssim (\kappa_2\sqrt{m_2 m_3} + C_2 + \sqrt{m}_3\beta_2)\ m_1 \exp\{-c_2^2/(8\beta_1^2)\}\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2} := \Re_{11}.$$

**Proof of Lemma 25**

Note that with high probability over the randomness of $V^{(0)}$, we have $\|V^{(0)}\|_F \lesssim \sqrt{m_2 m_3}\kappa_2$. Now according to Lemma 24 and using the fact that $\|V'\|_F \le C_2$:

$$\begin{aligned}
&\le \mathbb{E}_{W^{\rho},V^{\rho}} \frac{1}{\sqrt{m}_2}\|a\|\|V^{(0)} + V^{\rho} + (1-\eta)V'\|_2\|{\phi^{*}}'(x_i)\| \\
&= \mathbb{E}_{W^{\rho},V^{\rho}}\|V^{(0)} + V^{\rho} + (1-\eta/2)V'\|_F\|{\phi^{*}}'(x_i)\| \\
&\le \sqrt{\mathbb{E}_{V^{\rho}}(\|V^{(0)} + (1-\eta/2)V'\|_F^2 + \|V^{\rho}\|_F^2)}\ m_1 \exp\{-c_2^2/(8\beta_1^2)\}\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2} \\
&\lesssim \sqrt{\|V^{(0)}\|_F^2 + \|V'\|_F^2 + m_3\beta_2^2}\ m_1 \exp\{-c_2^2/(8\beta_1^2)\}\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2} \\
&\lesssim (\kappa_2\sqrt{m_2 m_3} + C_2 + \sqrt{m}_3\beta_2)\ m_1 \exp\{-c_2^2/(8\beta_1^2)\}\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2}.
\end{aligned}$$

**Lemma 26.** *Fifth extra term:*

$$\left|\mathbb{E}_{W^{\rho},V^{\rho}}[\frac{1}{\sqrt{m}_2} a^T D_{V^{(0)}+V^{\rho}+V',x_i} V_\Sigma^* {\phi^{*}}'(x_i)]\right| \lesssim \sqrt{\zeta}_2 m_1 \exp\{-c_2^2/(8\beta_1^2)\}\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2} := \Re_{10}.$$

**Proof of Lemma 26**

Similar to the previous Lemma, the inner expectation can be bounded as:

$$\leq \mathbb{E}\frac{1}{\sqrt{m_2}}\|a\|\|V_\Sigma^*\|_F\|\phi^{*\prime}(x_i)\| \leq \mathbb{E}\|V^*\|_F\|\phi^{*\prime}(x_i)\| \lesssim \sqrt{\zeta}_2 m_1 \exp\{-c_2^2/(8\beta_1^2)\}\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2}.$$

**Lemma 27.** *We have another extra term as a product of the movement $-\eta/2W'$ in the first layer:*

$$\phi^{(2)\prime}(x_i) = \frac{2}{\eta\sqrt{m_1}}\big[W^s D_{W^{(0)}+W^\rho+W'}(W^{(0)}+W^\rho+(1-\eta/2)W')x_i-\phi^{(0)}(x_i)-(1-\eta/2)\phi^{(2)}(x_i)\big].$$

*Then*

$$\begin{aligned}
&\Big|\mathbb{E}_{W^\rho,V^\rho}\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}V_\Sigma^*\phi^{(2)\prime}(x_i)\Big| \\
&\lesssim \sqrt{\zeta_2 m_3\Big(\frac{\beta_1^2}{m_1}+\frac{c_2C_1^2}{\sqrt{m_1}\kappa_1}+m_1\exp\{-c_2^2/(8\beta_1^2)\}C_1^2\Big)} := \Re_5. && (3.153)\\
&\mathbb{E}_{W^\rho,V^\rho}\Big|\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta/2)V')\phi^{(2)\prime}(x_i)\Big| \\
&\lesssim (\kappa_2\sqrt{m_2m_3}+\sqrt{m}_3\beta_2+C_2)\Re_5. && (3.154)
\end{aligned}$$

**Proof of Lemma 27**

First we prove the following approximation argument (for all $k \in [m_3]$):

$$\mathbb{E}_{W^\rho}|\phi^{(2)'}(x_i)_k|^2 \leq \frac{\beta_1^2}{m_1}+\frac{c_2C_1^2}{\sqrt{m_1}\kappa_1}+m_1\exp\{-c_2^2/(8\beta_1^2)\}C_1^2. \qquad (3.155)$$

We have

$$LHS = \mathbb{E}_{W^\rho}\Big|\frac{1}{\sqrt{m_1}}W_k^s D_{W^{(0)}+W^\rho+W',x_i}(W^{(0)}+W^\rho)x_i - \frac{1}{\sqrt{m_1}}W_k^s D_{W^{(0)},x_i}W^{(0)}x_i\Big|^2$$

$$\lesssim \mathbb{E}_{W^\rho}\Big|\frac{1}{\sqrt{m_1}}W_k^s D_{W^{(0)}+W^\rho,x_i}(W^{(0)}+W^\rho)x_i - \frac{1}{\sqrt{m_1}}W_k^s D_{W^{(0)},x_i}W^{(0)}x_i\Big|^2$$

$$+\mathbb{E}_{W^\rho}\Big|\frac{1}{\sqrt{m}_1}W_k^s D_{W^{(0)}+W^\rho,x_i}(W^{(0)}+W^\rho)x_i - \frac{1}{\sqrt{m}_1}W_k^s D_{W^{(0)}+W^\rho+W',x_i}(W^{(0)}+W^\rho)x_i\Big|^2$$

By the independence of $W_j^\rho$'s, the first term can be upper bounded as

$$= \mathbb{E}_{W^\rho}\frac{1}{m_1}\sum_{j=1}^{m_1}\Big((W_j^{(0)}+W_j^\rho)x_i\mathbb{1}\{(W_j^{(0)}+W_j^\rho)x_i \geq 0\} - W_j^{(0)}x_i\mathbb{1}\{W_j^{(0)}x_i \geq 0\}\Big)^2 =$$

$$\leq \frac{1}{m_1}\sum_{j=1}^{m_1}\mathbb{E}_{W^\rho}|W_j^\rho x_i|^2 = \frac{1}{m_1}\sum\frac{\beta_1^2}{m_1} = \frac{\beta_1^2}{m_1}.$$

For the second term, note that for every $j \notin P$, the $j$th entries of $D_{W^{(0)}+W^\rho,x_i}$ and $D_{W^{(0)}+W^\rho+W',x_i}$ are different only if $W_j^\rho$ can make a sign change in the $j$th row, i.e. $|(W_j^{(0)}+W_j')x_i| \leq |W_j^\rho x_i|$ should happen. We denote this event for every $j \notin P$ by $\tilde{E}_j$. Furthermore, if this happens for some $j$, then the value of $(W^{(0)}+W^\rho)_j x_i$ is upper bounded by $|W_j' x_i|$. Now similar to our discussion in Lemma 24 and using the result of Lemma 10:

$$\mathbb{P}(\cup_{j\notin P}\tilde{E}_j) = \mathbb{P}(\text{sign change in some } j \notin P) \leq \sum_{j\notin P}^{m_1}\mathbb{P}(\text{sign change in } j)$$

$$\leq m_1\mathbb{P}(|(W_j^{(0)}+W_j')x_i| \leq |W_j^\rho x_i|) \leq m_1\mathbb{P}(|W_j^\rho x_i| \geq c_2/(2\sqrt{m}_1)).$$

But note that $(W_j^\rho)x_i$ is Gaussian with variance $\beta_1^2/m_1$. Hence

$$LHS \lesssim m_1\exp\{-c_2^2/(8\beta_1^2)\},$$

So finally we can write

$$\lesssim \frac{\beta_1^2}{m_1} + \mathbb{E}_{W^\rho}\frac{1}{m_1}(\sum_{j\in P}|W_j' x_i|)^2 + E_{W^\rho}\frac{1}{m_1}(\mathbb{1}\{\cup_{j\notin P}\tilde{E}_j\}\sum_{j\notin P}|W_j' x_i|)^2$$

$$\lesssim \frac{\beta_1^2}{m_1} + \frac{|P|}{m_1}\|W'\|^2 + \mathbb{P}(\cup_{j\notin P}\tilde{E}_j)\|W'\|^2$$

$$\leq \frac{\beta_1^2}{m_1} + \frac{c_2 C_1^2}{\sqrt{m}_1\kappa_1} + m_1\exp\{-c_2^2/(8\beta_1^2)\}C_1^2.$$

which completes the proof for Equation (3.155). This immediately implies

$$\mathbb{E}_{W^\rho}\|\phi^{(2)\prime}(x_i)\| \le \sqrt{\mathbb{E}_{W^\rho}\|\phi^{(2)\prime}(x_i)\|^2} \le \sqrt{m_3(\frac{\beta_1^2}{m_1} + \frac{c_2 C_1^2}{\sqrt{m_1}\kappa_1} + m_1 \exp\{-c_2^2/(8\beta_1^2)\}C_1^2)}.$$

Now we first prove Equation (3.153):

$$\Big|\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i} V_\Sigma^* \phi^{(2)\prime}(x_i)]\Big| \le \mathbb{E}_{W^\rho}\frac{1}{\sqrt{m_2}}\|a\|\|D_{V^{(0)}+V^\rho+V',x_i} V_\Sigma^*\|_F\|\phi^{(2)\prime}(x_i)]\|\|$$

$$\le \|V_\Sigma^*\|_F \mathbb{E}_{W^\rho}\|\phi^{(2)\prime}(x_i)]\|\| \le \|V^*\|_F \mathbb{E}_{W^\rho}\|\phi^{(2)\prime}(x_i)]\|\|$$

$$\lesssim \sqrt{\zeta_2 m_3(\frac{\beta_1^2}{m_1} + \frac{c_2 C_1^2}{\sqrt{m_1}\kappa_1} + m_1 \exp\{-c_2^2/(8\beta_1^2)\}C_1^2)}.$$

To prove Equation (3.154):

$$\begin{aligned}
&\mathbb{E}_{W^\rho,V^\rho}\Big|\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta/2)V')\phi^{(2)\prime}(x_i)\Big| \\
&\lesssim \mathbb{E}_{W^\rho,V^\rho}\frac{1}{\sqrt{m_2}}\|a\|\|D_{V^{(0)}+V^\rho+(1-\eta/2)V',x_i}(V^{(0)}+V^\rho+(1-\eta/2)V')\|_F\|\phi^{(2)\prime}(x_i)\| \\
&\lesssim \mathbb{E}_{W^\rho,V^\rho}\frac{1}{\sqrt{m_2}}\|a\|\|V^{(0)}+V^\rho+(1-\eta/2)V'\|_F\|\phi^{(2)\prime}(x_i)\| \\
&\lesssim \frac{1}{\sqrt{m_2}}\|a\|\sqrt{\mathbb{E}_{V^\rho}(\|V^{(0)}\|^2+\|V^\rho\|_F^2+\|(1-\eta/2)V'\|_F^2)}\mathbb{E}_{W^\rho}\|\phi^{(2)\prime}(x_i)\| \\
&\lesssim \sqrt{(\kappa_2^2 m_2 m_3 + m_3\beta_2^2 + C_2^2)}\Re_5 \lesssim (\kappa_2\sqrt{m_2 m_3} + \sqrt{m}_3\beta_2 + C_2)\Re_5.
\end{aligned}$$

**Lemma 28.** *Closeness condition:*

$$\mathbb{E}_\Sigma\Big|\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i} V_\Sigma^* {\phi^*}_\Sigma(x_i)] - f^*(x_i)\Big| \lesssim \Re_9,$$

*where*

$$\Re_9 \coloneqq \varrho_3\xi\sqrt{n}(1+\Re)\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2} + \Re_3 \tag{3.156}$$

$$+ m_2\Big(\exp\{-(m_2\kappa_2^2 C_2^4)^{1/3}/(2\beta_2^2)\} + m_1\exp\{-C_1^2/(8m_3\beta_1^2)\}\Big)\sqrt{\zeta_2}(1+\Re)\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2}. \tag{3.157}$$

**Proof of Lemma 28**

Note that by Corollary 5.1 and according to the proof of Equation 3.130 in Lemma 19, if for every $j \notin \tilde{P}$ we don't have a sign change in $D_{V^{(0)}+V^\rho+V',x_i}V^*\phi^*(x_i)$, then get

$$|\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}V^*\phi^*(x_i) - f^*(x_i)| \leq \Re_3.$$

Also, note that we need the event $E^c$ (defined in Lemma 42) to happen in order to be able to use Corrolary 5.1. Hence, given a $W^\rho$ for which $E^c$ happens, we upper bound the probability of sign change with respect to the randomness of $V^\rho$. We define the following event with respect to the randomness of $V^\rho$ when conditioned on a $W^\rho$ for which $E^c$ happens ($P_i$'s are defined in Lemma 16):

$$SC \coloneqq \{\exists j \notin P_i \text{ s.t.} |V_j^\rho x_i'| \gtrsim (\frac{\kappa_2}{m_2})^{1/3} C_2^{2/3}\|x_i'\|\}.$$

Now from the result in Corollary 5.1 we have $\leq \mathbb{1}\{\text{sign change in } \ j \notin P_i\} \leq \mathbb{1}\{SC\}$. Therefore,

$$E\mathbb{1}\{\text{sign change}\} \leq \mathbb{1}\{SC\} \leq \sum_{j \notin P_i} \mathbb{P}(|V_j^\rho x_i'| \gtrsim (\frac{\kappa_2}{m_2})^{1/3} C_2^{2/3}\|x_i'\|)$$

$$\leq m_2 \mathbb{P}(|V_j^\rho x_i'| \gtrsim (\frac{\kappa_2}{m_2})^{1/3} C_2^{2/3}\|x_i'\|).$$

But note that $(V_j^\rho)x_i'$ is Gaussian with variance $\beta_2^2\|x_i'\|^2/m_2$. Hence

$$LHS \lesssim m_2 \exp\{-(m_2\kappa_2^2 C_2^4)^{1/3}/(2\beta_2^2)\}. \tag{3.158}$$

Now let $D$ be a sign matrix random variable such that if $E^c$ and $SC^c$ both happens, then it is equal to the valid sign matrix $D_{V^{(0)}+V^\rho+V',x_i}$, and otherwise it is equal to an arbitrary valid sign matrix in the case when both $E^c$ and $SC^c$ happen. Now using

Equation (3.117) we have with high probability over the initialization:

$$
\begin{aligned}
&\mathbb{E}_\Sigma\Big|\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}V^*_\Sigma {\phi^*}_\Sigma(x_i)] - f^*(x_i)\Big| \\
&\leq \mathbb{E}_\Sigma\Big|\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^*_\Sigma - V^*\Sigma){\phi^*}_\Sigma(x_i)] \\
&+ \mathbb{E}_\Sigma\Big|\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}V^*\Sigma {\phi^*}_\Sigma(x_i)] - f^*(x_i) \\
&\leq \mathbb{E}_{W^\rho,V^\rho}\mathbb{E}_\Sigma\Big|\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^*_\Sigma - V^*\Sigma){\phi^*}_\Sigma(x_i)\Big| \\
&+ \mathbb{E}_\Sigma\Big|\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V^\rho+V',x_i}V^*\phi^*(x_i)] - f^*(x_i)\Big| \\
&\leq \mathbb{E}_{W^\rho,V^\rho}\mathbb{E}_\Sigma\frac{1}{m_2}\sum_j \|V^*_{\Sigma j} - V^*_j\Sigma\|\|{\phi^*}_\Sigma(x_i)\| \\
&+ \mathbb{E}_\Sigma\Big|\mathbb{E}_{W^\rho,V^\rho}[\Big(\frac{1}{\sqrt{m_2}}a^T D V^*\phi^*(x_i) - f^*(x_i)\Big) \\
&+ \mathbb{1}\{SC \cup E\}\Big(\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V',x_i}V^*\phi^*(x_i) - D\Big)]\Big| \\
&\leq \mathbb{E}_{W^\rho,V^\rho}\mathbb{E}_\Sigma\frac{1}{m_2}\sum_j \|V^*_{\Sigma j} - V^*_j\Sigma\|\|{\phi^*}_\Sigma(x_i)\| \\
&+ \mathbb{E}_\Sigma\Big|\mathbb{E}_{W^\rho,V^\rho}\Big[\Big(\frac{1}{\sqrt{m_2}}a^T D V^*\phi^*(x_i) - f^*(x_i)\Big)\Big| \\
&+ \mathbb{E}_\Sigma\mathbb{E}_{W^\rho,V^\rho}\Big|\mathbb{1}\{SC \cup E\}\Big(\frac{1}{\sqrt{m_2}}a^T D_{V^{(0)}+V',x_i}V^*\phi^*(x_i) - \frac{1}{\sqrt{m_2}}a^T D V^*\phi^*(x_i)\Big)\Big]\Big| \\
&\leq \mathbb{E}_{W^\rho,V^\rho}\frac{1}{\sqrt{m_2}}\sum_j \sqrt{\mathbb{E}_\Sigma\|V^*_{\Sigma j} - V^*_j\Sigma\|^2}\|\phi^*(x_i)\| \\
&+ \Re_3 + 2\mathbb{P}(SC \cup E)\max_{D'}\Big|\frac{1}{\sqrt{m_2}}a^T D' V^*\phi^*(x_i)\Big|.
\end{aligned}
$$

Now note that for any sign matrix $D'$, we have the following bound:

$$
\Big|\frac{1}{\sqrt{m_2}}a^T D' V^*\phi^*(x_i)\Big| \leq \frac{1}{\sqrt{m_2}}\|a\|\|V^*\|_F\|\phi^*(x_i)\| \lesssim \sqrt{\zeta_2}(1+\Re)\sqrt{\sum_k \|\mathcal{V}_k\|^2_{H^\infty}}.
$$

Also, applying a union bound and using Lemmas 42

$$
\begin{aligned}
\mathbb{P}(SC \cup E) &\leq \mathbb{P}(SC) + \mathbb{P}(E) \\
&\lesssim \exp\{-(m_2\kappa_2^2 C_2^4)^{1/3}/(2\beta_2^2)\} + m_1\exp\{-C_1^2/(8m_3\beta_1^2)\}.
\end{aligned}
$$

Hence, also applying Lemma 49, we further write

$$LHS \lesssim \varrho_3 \xi \sqrt{n}(1+\Re)\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2} + \Re_3$$
$$+ m_2\Big(\exp\{-(m_2\kappa_2^2 C_2^4)^{1/3}/(2\beta_2^2)\} + m_1 \exp\{-C_1^2/(8m_3\beta_1^2)\}\Big)\sqrt{\zeta_2}(1+\Re)\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2}.$$

**Lemma 29.** *Suppose we have $m_3\kappa_1^2 \geq C_1^2$. Then, for the following basic term we have:*

$$\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta/2)V')(\phi^{(0)}(x_i)+(1-\eta/2)\phi^{(2)}(x_i))$$

$$\lesssim (1-\eta)\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+V')(\phi^{(0)}(x_i)+\phi^{(2)}(x_i))$$
$$\pm\eta\Big(\Re_6' + \Re_4 + (\sqrt{m}_3\kappa_2+\beta_2)(C_1+\sqrt{m}_3\beta_1)\Big),$$

*where*

$$\Re_4 := C_2(C_1+\sqrt{m}_3\beta_1)m_2\exp\{-C_2^{4/3}(\sqrt{m}_2\kappa_2)^{2/3}/8\beta_2^2\} + \frac{C_2^{1/3}}{(\sqrt{m}_2\kappa_2)^{1/3}}C_2(C_1+\sqrt{m}_3\beta_1),$$
$$\Re_6' := m_1\exp\{-C_1^2/(8m_3\beta_1^2)\}\sqrt{m}_3\kappa_1(\sqrt{m}_2+\beta_2)+\Re_6,$$

*and $\Re_6$ is defined in Lemma 31.*

**Proof of Lemma 29**

First, note that by orthogonality of $\phi^{(0)}(x_i)$ to the rows of $V'$:

$$\begin{aligned}
&LHS - \eta/2\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)\phi^{(0)}(x_i)\\
&= LHS - \eta/2\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta/2)V')\phi^{(0)}(x_i)\\
&= (1-\eta/2)\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta/2)V')(\phi^{(0)}(x_i)+\phi^{(2)}(x_i))]\\
&= (1-\eta/2)^2\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+V')(\phi^{(0)}(x_i)+\phi^{(2)}(x_i))]\\
&+ (1-\eta/2)(\eta/2)\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)(\phi^{(0)}(x_i)+\phi^{(2)}(x_i))] \quad (3.159)
\end{aligned}$$

But note that for the second term:

$$\begin{aligned}
&\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)(\phi^{(0)}(x_i)+\phi^{(2)}(x_i))]\\
&\lesssim \mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho,x_i}(V^{(0)}+V^\rho)(\phi^{(0)}(x_i)+\phi^{(2)}(x_i))]\\
&\pm \frac{1}{\sqrt{m}_2}\sum_{j:\text{ sign change}}|V'_j(\phi^{(0)}(x_i)+\phi^{(2)}(x_i))|\\
&= \mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho,x_i}(V^{(0)}+V^\rho)(\phi^{(0)}(x_i)+\phi^{(2)}(x_i))]\\
&\pm \frac{1}{\sqrt{m}_2}\sum_{j:\text{ sign change}}|V'_j\phi^{(2)}(x_i)|. \qquad (3.160)
\end{aligned}$$

Now conditioned on $x'_i$, by the result of Lemma 40 we know there exists a set of indices $O \subseteq [m_2]$, s.t. $|O| \le \frac{C_2^{2/3}}{(\sqrt{m}_2\kappa_2)^{2/3}}m_2$ and for $j \notin O$ we have

$$|V_j^{(0)}x'_i| \ge \frac{C_2^{2/3}(\sqrt{m}_2\kappa_2)^{1/3}}{\sqrt{m}_2}\|x'_i\|$$

and

$$|V'_jx'_i| \le \frac{C_2^{2/3}(\sqrt{m}_2\kappa_2)^{1/3}}{2\sqrt{m}_2}\|x'_i\|.$$

Now for $j \in [m_2]$, define the event

$$R_j = \{|W_j^\rho x'_i| \ge \frac{C_2^{2/3}(\sqrt{m}_2\kappa_2)^{1/3}}{2\sqrt{m}_2}\|x'_i\|\},$$

and $R = \cup_j R_j$. First, note that using Gaussian tail bound, $R$ is a rare event:

$$\mathbb{P}(R) \leq \sum_j \mathbb{P}(R_j) \leq m_2 \exp\{-C_2^{4/3}(\sqrt{m}_2\kappa_2)^{2/3}/8\beta_2^2\}.$$

Now for $j \notin O$ and under $R^c$, clearly we have that the signs of $(V_j^{(0)} + V_j^\rho)x_i'$ and $(V_j^{(0)} + V_j^\rho + V_j')x_i'$ are the same. Therefore, applying Lemma 42, we can argue under $R^c$:

$$\mathbb{E}_{W^\rho,V^\rho}\frac{1}{\sqrt{m}_2}\sum_{j\text{: sign change}}|V_j'\phi^{(2)}(x_i)| \leq \mathbb{E}_{W^\rho}\frac{1}{\sqrt{m}_2}\sum_{j\in O}|V_j'\phi^{(2)}(x_i)|$$

$$\leq \sqrt{\frac{|O|}{m_2}}\|V'\|E_{W^\rho}\|\phi^{(2)}(x_i)\| \leq \frac{C_2^{1/3}}{(\sqrt{m}_2\kappa_2)^{1/3}}C_2(C_1 + \sqrt{m}_3\beta_1).$$

Hence, overall, using Cauchy-Shwartz

$$\mathbb{E}_{W^\rho,V^\rho}\frac{1}{\sqrt{m}_2}\sum_{j\text{: sign change}}|V_j'\phi^{(2)}(x_i)| \leq \|V'\|\mathbb{E}_{W^\rho}\|\phi^{(2)}(x_i)\|\mathbb{P}(R)+\frac{C_2^{1/3}}{(\sqrt{m}_2\kappa_2)^{1/3}}C_2(C_1+\sqrt{m}_3\beta_1)$$

$$\leq C_2(C_1 + \sqrt{m}_3\beta_1)m_2\exp\{-C_2^{4/3}(\sqrt{m}_2\kappa_2)^{2/3}/8\beta_2^2\} + \frac{C_2^{1/3}}{(\sqrt{m}_2\kappa_2)^{1/3}}C_2(C_1 + \sqrt{m}_3\beta_1) := \Re_4. \tag{3.161}$$

On the other hand, using Lemma 39, we have with high probability over the randomness of initialization

$$\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)},x_i}V^{(0)}\phi^{(2)}(x_i) \leq \sqrt{m}_3\kappa_2\|\phi^{(2)}(x_i)\|.$$

Hence:

$$\begin{aligned}
&\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho,x_i}(V^{(0)}+V^\rho)\phi^{(2)}(x_i)] \\
&\leq \mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)},x_i}V^{(0)}\phi^{(2)}(x_i)+\frac{1}{\sqrt{m}_2}\sum_j |V_j^\rho\phi^{(2)}(x_i)|] \\
&\leq \mathbb{E}_{W^\rho}\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)},x_i}V^{(0)}\phi^{(2)}(x_i)+\beta_2\mathbb{E}_{W^\rho}\|\phi^{(2)}(x_i)\| \\
&\lesssim (\sqrt{m}_3\kappa_2+\beta_2)\mathbb{E}_{W^\rho}\|\phi^{(2)}(x_i)\| \\
&\lesssim (\sqrt{m}_3\kappa_2+\beta_2)(C_1+\sqrt{m}_3\beta_1). \qquad (3.162)
\end{aligned}$$

Combining Equations (3.161) and (3.162) into Equation (3.160):

$$\left|\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)\phi^{(2)}(x_i)]\right| \lesssim \Re_4+(\sqrt{m}_3\kappa_2+\beta_2)(C_1+\sqrt{m}_3\beta_1). \qquad (3.163)$$

Moreover, for the first term in (3.159), using Equation (3.162) and Lemmas 42 and Lemma 39 we have

$$\begin{aligned}
&|\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+V')\phi^{(2)}(x_i)]| \\
&\lesssim |\mathbb{E}_{W^\rho,V^\rho}\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}}V^{(0)}\phi^{(2)}(x_i)|+\mathbb{E}_{W^\rho,V^\rho}\frac{1}{\sqrt{m}_2}\sum_j |V_j^\rho\phi^{(2)}(x_i)|+\frac{1}{\sqrt{m}_2}\sum_j |V_j'\phi^{(2)}(x_i)| \\
&\lesssim \kappa_2\sqrt{m}_3(C_1+\beta_1\sqrt{m}_3)+(C_2+\beta_2)\mathbb{E}_{W^\rho}\|\phi^{(2)}(x_i)\| \\
&\lesssim \kappa_2\sqrt{m}_3(C_1+\beta_1\sqrt{m}_3)+(C_2+\beta_2)(C_1+\sqrt{m}_3\beta_1). \qquad (3.164)
\end{aligned}$$

Substituting Equations (3.163) and (3.164) into Equation (3.159), we finally get

$$
\begin{aligned}
LHS \; - \; & \eta/2\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)\phi^{(0)}(x_i) \\
& \lesssim (1-\eta/2)^2\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+V')\phi^{(2)}(x_i)] \\
& \pm \frac{\eta}{2}\Big(\Re_4 + (\sqrt{m}_3\kappa_2+\beta_2)(C_1+\sqrt{m}_3\beta_1)\Big) \\
& \lesssim (1-\eta)\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+V')\phi^{(2)}(x_i)] \\
& \pm \frac{\eta^2}{4}\Big|\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+V')\phi^{(2)}(x_i)]\Big| \\
& \pm \frac{\eta}{2}\Big(\Re_4 + (\sqrt{m}_3\kappa_2+\beta_2)(C_1+\sqrt{m}_3\beta_1)\Big) \\
& \lesssim (1-\eta)\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+V')\phi^{(2)}(x_i)] \\
& \pm \eta^2(\kappa_2\sqrt{m}_3(C_1+\beta_1\sqrt{m}_3)+(C_2+\beta_2)(C_1+\sqrt{m}_3\beta_1)) \\
& \pm \eta\Big(\Re_4 + (\sqrt{m}_3\kappa_2+\beta_2)(C_1+\sqrt{m}_3\beta_1)\Big). \qquad (3.165)
\end{aligned}
$$

Now by picking $\eta$ small enough so that the second term is dominated by the third term we get:

$$LHS - \eta/2\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)\phi^{(0)}(x_i) \qquad (3.166)$$

$$\lesssim (1-\eta)\mathbb{E}_{V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+V')\phi^{(2)}(x_i)] \qquad (3.167)$$

$$\pm \eta\Big(\Re_4 + (\sqrt{m}_3\kappa_2+\beta_2)(C_1+\sqrt{m}_3\beta_1)\Big). \qquad (3.168)$$

Now we aim to bound the term $\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)\phi^{(0)}(x_i)$. First assume that we are in the event $E^c$ defined in Lemma 42, i.e. we have $\|\phi^{(2)}(x_i)\| \lesssim C_1$. Conditioned on such $W^\rho$, we now work with the randomness of the initialization and $V^\rho$. Note that the random matrix $V^{(0)}+V^\rho$ jointly over the randomness of $V^\rho$ and the initialization is also Gaussian, and its variance is

$$\kappa_2^2 \le \kappa_2^2 + \frac{\beta_2^2}{m_2} \le 2\kappa_2^2, \qquad (3.169)$$

where the inequality follows from the fact that $\kappa_2 \geq \frac{1}{\sqrt{m_2}}$ and $\beta_2 \leq 1$. Therefore, applying Lemma 31 for the random matrix $V^{(0)}$ in the Lemma as $V^{(0)} + V^\rho$ here, the bound does not change up to constants because of the inequality (3.169). Hence, with high probability, lets say with prob. $1 - \delta_1$ this time over both the randomness of initialization and $V^\rho$:

$$\mathcal{L} = \Big|\frac{1}{\sqrt{m_2}} a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)} + V^\rho)\phi^{(0)}(x_i)\Big| \leq \Re_6 \tag{3.170}$$

This means that with probability at least $1 - \sqrt{\delta_1}$ over the random initialization, then we have (3.170) with prob. at least $1 - \sqrt{\delta_1}$ over the randomenss of $V^\rho$. We name the latter high probability statement as $(\star)$. Moreover, note that by Lemma 41 and assuming $m_3 \log(m_2) < m_2$, we have the following almost surely bound (also note that $V^\rho_j \phi^{(0)}(x_i)$ is Gaussian with std $\frac{\beta_2}{\sqrt{m_2}}\|\phi^{(0)}(x_i)\|$):

$$\begin{aligned}
&\mathbb{E}_{V^\rho}\Big|\frac{1}{\sqrt{m_2}} a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)} + V^\rho)\phi^{(0)}(x_i)\Big| && (3.171)\\
&= \mathbb{E}_{V^\rho}\Big|\frac{1}{\sqrt{m_2}} a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)} + V^\rho)\phi^{(0)}(x_i)\Big| && (3.172)\\
&\leq \frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2} |V_j^{(0)}\phi^{(0)}(x_i)| + \mathbb{E}_{V^\rho}\frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2} |V_j^\rho \phi^{(0)}(x_i)| && (3.173)\\
&\lesssim \|\phi^{(0)}(x_i)\| \sup_{\|x'\|=1} \frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2} |V_j^{(0)}x'| + \frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}\frac{\beta_2}{\sqrt{m_2}}\|\phi^{(0)}(x_i)\| && (3.174)\\
&\lesssim \|\phi^{(0)}(x_i)\|(\sqrt{m_2} + \beta_2). && (3.175)
\end{aligned}$$

Furthermore, note because each variable $|V^\rho_j \phi^{(0)}(x_i)|$ is $\frac{\beta_2}{\sqrt{m_2}}\|\phi^{(0)}(x_i)\|$ subGaussian. Therefore, $\mathcal{L}$ is subGaussian with parameter $\|\phi^{(0)}(x_i)\|\beta_2$ with respect to the randomness of $V^\rho$. Now the point is that the high probability argument in $(\star)$ is much stronger than what one can get from the subGaussian ineqaulity with parameter $\|\phi^{(0)}(x_i)\|\beta_2$ (with the corresponding expectation term $\|\phi^{(0)}(x_i)\|(\sqrt{m_2} + \beta_2)$). However, the disadvantage of $(\star)$ is that it only works for a fixed $\delta_1$. In other words, at least it is not obvious from this argument that why for a fixed $W^{(0)}$ in a high probaiblity region of the random initialization, whether we can send $\delta_1$ to zero by growing the constant behind $\mathfrak{R}_6$ with logarithmic rate $\log(1/\delta)$. This makes our job hard for bounding the expectation with respect to $V^\rho$ if we only wish to rely on $(\star)$. Therefore, we combine it with the inequality that we get from the subGaussian parameter that we introudced above. More rigorously, we define the thresholding parameter

$$\mho := \|\phi^{(0)}(x_i)\|(\sqrt{m}_2 + \beta_2) + \|\phi^{(0)}(x_i)\|\beta_2 \log(\|\phi^{(0)}(x_i)\|(\sqrt{m}_2 + \beta_2)/\mathfrak{R}_6)$$

$$= \Theta\Big(\|\phi^{(0)}(x_i)\|(\sqrt{m}_2 + \beta_2 \log(\|\phi^{(0)}(x_i)\|(\sqrt{m}_2 + \beta_2)/\mathfrak{R}_6))\Big),$$

for which we have

$$\mathbb{E}\Big[\mathcal{L}|\ \mho \le \mathcal{L}\Big]\mathbb{P}(\mho \le \mathcal{L}) \lesssim \mathfrak{R}_6.$$

we divide the range of values for $\mathcal{L}$ into three parts:

$$\begin{aligned}\mathbb{E}[\mathcal{L}] = \mathbb{E}&\Big[\mathcal{L}|\ \mathcal{L} \le \mathfrak{R}\Big]\mathbb{P}(\mathcal{L} \le \mathfrak{R}_6)\\ &+ \mathbb{E}\Big[\mathcal{L}|\ \mathfrak{R}_6 \le \mathcal{L} \le \mho\Big]\mathbb{P}\Big(\mathfrak{R}_6 \le \mathcal{L} \le \mho\Big)\\ &+ \mathbb{E}\Big[\mathbb{L}|\ \mho \le \mathcal{L}\Big]\mathbb{P}(\mho \le \mathcal{L})\\ &\le E\Big[\mathcal{L}|\ \mathcal{L} \le \mathfrak{R}_6\Big] + \mathbb{P}(\mathfrak{R}_6 \le \mathcal{L} \le \mho) + \mathfrak{R}_6\\ &\lesssim \mathfrak{R}_6 + \sqrt{\delta}_1\mho.\end{aligned}$$

Now by choosing $\delta_1 \lesssim 1/\mho$, we conclude with high probability over initialization and conditioned on $W^\rho$'s such that $E^c$ happens we have

$$\mathbb{E}_{V^\rho}\Big|\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)} + V^\rho)\phi^{(0)}(x_i)\Big| = \mathbb{E}[\mathcal{L}] \lesssim \mathfrak{R}_6.$$

Finally, we integrate also with respect to $W^\rho$. To control the random variable when $E$ happens, we use the bound in (3.175) and the fact that $E$ is a rare event due to Lemma 42:

$$\mathbb{E}_{W^\rho,V^\rho}\Big|\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)\phi^{(0)}(x_i)\Big| \lesssim \mathbb{P}(E)\|\phi^{(0)}(x_i)\|(\sqrt{m}_2+\beta_2)+\mathbb{P}(E^c)\mathfrak{R}_6$$

$$\leq m_1 \exp\{-C_1^2/(8m_3\beta_1^2)\}\sqrt{m}_3\kappa_1(\sqrt{m}_2 + \beta_2) + \Re_6 := \Re_6'.$$

Substituting this into (3.168) the proof is finally complete.

**Lemma 30.** *Third cross term: with high probability over initialization, we have*

$$\mathbb{E}_\Sigma\Big(\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta)V')\Sigma\phi^*(x_i)]\Big)^2$$

$$\lesssim \xi^2(m_1\exp\{-C_1^2/(8m_3\beta_1^2)\}(\kappa_2^2 m_2 m_3 + \beta_1^2 m_3) + \Re_7^2 C_2^2) := \Re_{12}^2.$$

**Proof of Lemma 30**

Note that the way we defined the matrix $W^*$ and as a result $\phi^*(x_i)$ only depends on the randomness of $W^{(0)}$, not on $W'$ or the randomness of $V^{(0)}$. Now using Equation (3.131) and Jensen inequality we can write (for vector $v$, the notation $v^{2\odot}$ is another vector with each entry as the second power of the corresponding entry in $v$):

$$\begin{aligned}
&= \mathbb{E}_\Sigma\Big(\mathbb{E}_{W^\rho,V^\rho}[\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta)V')\Sigma\phi^*(x_i)]\Big)^2\\
&\leq \mathbb{E}_\Sigma\mathbb{E}_{W^\rho,V^\rho}\Big(\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta)V')\Sigma\phi^*(x_i)\Big)^2\\
&= \mathbb{E}_{W^\rho,V^\rho}\mathbb{E}_\Sigma\Big(\langle\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta)V')\ ,\ \Sigma\phi^*(x_i)\rangle\Big)^2\\
&= \mathbb{E}_{W^\rho,V^\rho}\Big\langle\Big(\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta)V')\Big)^{2\odot}\ ,\ \phi^*(x_i)^{2\odot}\Big\rangle\\
&\leq \mathbb{E}_{W^\rho,V^\rho}\Big\|\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta)V')\Big\|_2^2\ \Big\|\phi^*(x_i)\Big\|_\infty^2\\
&\leq \xi^2\mathbb{E}_{W^\rho,V^\rho}\Big\|\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho+(1-\eta)V')\Big\|_2^2\\
&\leq 2\xi^2\mathbb{E}_{W^\rho,V^\rho}\Big\|\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)\Big\|_2^2 + 2\xi^2\mathbb{E}_{W^\rho,V^\rho}\Big\|\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(1-\eta)V'\Big\|_2^2\\
&\leq \xi^2\mathbb{E}_{W^\rho,V^\rho}\Big\|\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)\Big\|_2^2 + \xi^2(1-\eta)^2\|V'\|_F^2\\
&\leq \xi^2\mathbb{E}_{W^\rho,V^\rho}\Big\|\frac{1}{\sqrt{m}_2}a^T D_{V^{(0)}+V^\rho+V',x_i}(V^{(0)}+V^\rho)\Big\|_2^2 + \xi^2(1-\eta)^2C_2^2.
\end{aligned}$$

Now under the event $E^c$ defined in Lemma 42 we get that $\|\phi^{(2)}(x_i)\| \lesssim C_1$, so we can bound the above as

$$\leq \xi^2 \mathbb{E}_{V^\rho} \sup_{\|V'\| \leq C_2, V' \perp \phi^{(0)}(x_i), \|x'\| \leq C_1} \frac{1}{m_2} \Big\| \sum_j a_j \mathbb{1}\{(V_j^{(0)} + V_j^\rho + V_j')(\phi^{(0)}(x_i) + x') \geq 0\}(V_j^{(0)} + V_j^\rho) \Big\|^2 \tag{3.176}$$

$$+ \xi^2 (1-\eta)^2 C_2^2. \tag{3.177}$$

Now defining

$$\mathcal{L}_2 \coloneqq \frac{1}{m_2} \Big\| \sum_j a_j \mathbb{1}\{(V_j^{(0)} + V_j^\rho + V_j')(\phi^{(0)}(x_i) + x') \geq 0\}(V_j^{(0)} + V_j^\rho) \Big\|^2,$$

to bound the first term, we want to apply Lemma 32 using the same trick that we did in the proof of Lemma 29. Note that $\mathcal{L}_2$ is the same term as $\Gamma^2_{x',V'}$ in Lemma 32 except that it is defined with respect to $V^{(0)} + \mathcal{V}^\rho$ instead of $V^{(0)}$. On the other hand, note that $V^{(0)} + V^\rho$ has Gaussian entries with variance $\kappa_2^2 + \frac{\beta_2^2}{m_2}$ and we know $\kappa_2^2 \leq \kappa_2^2 + \frac{\beta_2^2}{m_2} \leq 2\kappa_2^2$, which means the argument of Lemma 32 holds true here up to constants:

$$\sup_{\|V'\| \leq C_2, V' \perp \phi^{(0)}(x_i), \|x'\| \leq C_1} \mathcal{L}_2 \lesssim \Re_7^2.$$

This holds with probability say $1 - \delta_2$ over the randomness of both $V^{(0)}$ and $V^\rho$. Therefore, with probability $1 - \sqrt{\delta}_2$ over the initialization, then with probability at least $1 - \sqrt{\delta}_2$ over the randomness of $V^\rho$ we have the above. Moreover, with a simple Cauchy-Swuartz we get the following almost surely bound:

$$\mathcal{L}_2 \lesssim \|V^{(0)}\|_F^2 + \|V^\rho\|_F^2. \tag{3.178}$$

Now the variable $\|V^\rho\|^2$ is subexponential with parameter $(\beta_1^4 m_3^2, \beta_1^2 m_3)$. Furthermore, with high probability we have $\|V^{(0)}\|_F^2 \lesssim m_2 m_3 \kappa_2^2$. Therefore, taking

$$\mho_2 \coloneqq \Theta\Big(\kappa_2^2 m_2 m_3 + \beta_1^2 m_3 \log\big((\kappa_2^2 m_2 m_3 + \beta_1^2 m_3)/\Re_7\big)\Big),$$

then one can easily see by the subexponential tail:

$$\begin{aligned}&\mathbb{E}[\mathcal{L}_2|\ \mathcal{L}_2 \geq \mho_2] = \Theta\Big(\mho_2\Big),\\ &\mathbb{P}(\mathcal{L}_2 \geq \mho_2) \leq \Re_7^2/\mho_2.\end{aligned}$$

Hence, we can apply the same trick as Lemma 29 as

$$\begin{aligned}\mathbb{E}[\mathcal{L}_2] = \mathbb{E}&\Big[\mathcal{L}_2|\ \mathcal{L}_2 \leq \Re_7^2\Big]\mathbb{P}(\mathcal{L}_2 \leq \Re_7^2)\\ &+ \mathbb{E}\Big[\mathcal{L}_2|\ \Re_7^2 \leq \mathcal{L}_2 \leq \mho_2\Big]\mathbb{P}\Big(\Re_7^2 \leq \mathcal{L}_2 \leq \mho_2\Big)\\ &+ \mathbb{E}\Big[\mathcal{L}_2|\ \mho_2 \leq \mathcal{L}_2\Big]\mathbb{P}(\mho_2 \leq \mathcal{L}_2)\\ &\lesssim E\Big[\mathcal{L}_2|\ \mathcal{L}_2 \leq \Re_7^2\Big] + \mathbb{P}(\Re_7^2 \leq \mathcal{L}_2 \leq \mho_2) + \Re_7^2\\ &\lesssim \Re_7^2 + \sqrt{\delta_2}\mho_2.\end{aligned}$$

Now taking $\delta_2 \lesssim \Re_7^4/\mho_2^2$, we finally get that conditioned on $W^\rho$'s where $E$ happens, then

$$\mathbb{E}_{V^\rho}\mathcal{L}_2 \leq \Re_7^2.$$

On the other hand, to handle the case when $E$ happens, we can use the bound in (3.178) as it does not depend on the occurrence of $E$ as well:

$$\begin{aligned}\mathbb{E}_{W^\rho,V^\rho}\mathcal{L}_2 &\leq \mathbb{P}(E)\mathbb{E}_{V^\rho}(\|V^{(0)}\|^2 + \|V^\rho\|^2) + \mathbb{P}(E^c)\Re_7^2\\ &\lesssim m_1 \exp\{-C_1^2/(8m_3\beta_1^2)\}(\kappa_2^2 m_2 m_3 + \beta_1^2 m_3) + \Re_7^2.\end{aligned}$$

Plugging this back into (3.177) we finally get

$$LHS \lesssim \xi^2(m_1 \exp\{-C_1^2/(8m_3\beta_1^2)\}(\kappa_2^2 m_2 m_3 + \beta_1^2 m_3) + \Re_7^2) + \xi^2 C_2^2.$$

### 3.6.15 Bounding the worst-case Senario

**Lemma 31.** *Suppose $m_3 \geq \log(m_2)$ and $\sqrt{m}_3\kappa_1 \gtrsim C_1$. We define the sign matrices $D^{x'}_{V^{(0)}+V',x_i}$ and $D^{x'}_{V^{(0)},x_i}$ with respect to the multiplications*

$$(V^{(0)} + V')(\phi^{(0)}(x_i) + x'),$$

*and*

$$V^{(0)}(\phi^{(0)}(x_i) + x').$$

*Then, with high probability:*

$$\sup_{\|x'\| \lesssim C_1, \|V'\|_F \leq C_2, V' \perp \phi^{(0)}} \frac{1}{\sqrt{m}_2} a^T D^{x'}_{V^{(0)}+V',x_i} V^{(0)}\phi^{(0)}(x_i)$$
$$\lesssim \Big( \frac{(C_1C_2)^{4/3}}{(\sqrt{m}_2\kappa_2)^{1/3}(\sqrt{m}_3\kappa_1)^{1/3}} + \frac{(C_1C_2)^{2/3} m_3^{2/3} (\kappa_1\kappa_2)^{1/3}\sqrt{\log(m_2)}}{m_2^{1/3}} \Big) \Big( 1 + \log(m_2) \frac{C_1^{1/3}(\kappa_2\sqrt{m}_2)^{2/3}}{C_2^{2/3}(\kappa_1\sqrt{m}_3)^{1/3}} \Big)$$
$$+ \frac{m_3^{3/2}\kappa_1\kappa_2}{\sqrt{m}_2}\sqrt{\log(m_2)}(\log(m_3) + \log(\log(m_2))) + \kappa_1\kappa_2\sqrt{m_3\log(m_2)} := \Re_6.$$

**Proof of Lemma 31**

Consider a cover for the euclidean ball of radius $C_1$ in $\mathbb{R}^{m_3}$ with precision $\epsilon$, i.e. $B_{C_1}(\epsilon)$. So for every $x' \in \mathbb{R}^{m_3}$, there exists an $x \in B_{C_1}(\epsilon)$ such that $\|x - x'\| \leq \epsilon$, and $|B_{C_1}(\epsilon)| \lesssim (\frac{1}{\epsilon})^{m_3}$. Now fix $x'$ and $x$. We have

$$\Gamma_{x',V'} := \frac{1}{\sqrt{m}_2} a^T D^{x'}_{V^{(0)}+V',x_i} V^{(0)}\phi^{(0)}(x_i) = \frac{1}{\sqrt{m}_2}\sum_{j=1}^{m_2} a_j \mathbb{1}\{(V_j^{(0)}+V_j')(\phi^{(0)}(x_i)+x') \geq 0\} V_j^{(0)}\phi^{(0)}(x_i).$$

Now by a union bound, because each variable $V_j^{(0)}\phi^{(0)}(x_i)$ is Gaussian with parameter $\kappa_2\|\phi^{(0)}(x_i)\|$ and using Equation (3.116), with high probability we have for every $j \in [m_2]$:

$$V_j^{(0)}\phi^{(0)}(x_i) \lesssim \kappa_2\|\phi^{(0)}(x_i)\|\sqrt{\log(m_2)} \lesssim \kappa_1\kappa_2\sqrt{m_3\log(m_2)}. \tag{3.179}$$

Therefore, by Hoeffding over the randomness of the Bernoulli variables $a_j$, for a fixed $x'$ with high probability:

$$\Gamma_{x'} := \frac{1}{\sqrt{m}_2} \sum_{j=1}^{m_2} a_j \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i) + x') \geq 0\} V_j^{(0)} \phi^{(0)}(x_i) \lesssim \kappa_1 \kappa_2 \sqrt{m_3 \log(m_2)}.$$

On the other hand, We know that the VC-dimension of the class of binary functions with respect to halfspaces in $\mathbb{R}^{m_3}$ is $m_3 + 1$. Therefore, the set of different sign patterns in matrices $D^{x'}_{V^{(0)}, x_i}$ is bounded by $m_2^{m_3+1}$, i.e. for

$$\mathcal{D} = \{D^{x'}_{V^{(0)}, x_i} | \ x' \in \mathbb{R}^{m_3}\},$$

we have

$$|\mathcal{D}| \lesssim m_2^{m_3+1}.$$

Therefore, by taking a union bound over all sign matrices in $\mathcal{D}$, we get with high probability

$$\sup_{x'} \Gamma_{x'} \lesssim \kappa_1 \kappa_2 \sqrt{m_3 \log(m_2)} \sqrt{\log({m_2}^{m_3+1})} = \kappa_1 \kappa_2 m_3 \log(m_2). \tag{3.180}$$

Now for a threshold $r$ which satisfies

$$r \geq 2\sqrt{m}_3 \kappa_2 \epsilon, \tag{3.181}$$

we define

$$\mathcal{J}_{x,r} = \{j \in [m_2] | \ |V_j^{(0)}(\phi^{(0)}(x_i) + x)| \leq r\}.$$

Now by Equation (3.116)and the assumption of the Lemma $\sqrt{m}_3 \kappa_1 \gtrsim C_1$, we have

$$\|\phi^{(0)}(x_i) + x\| \leq \|\phi^{(0)}(x_i)\| + \|x\| \lesssim \sqrt{m}_3 \kappa_1 + C_1. \tag{3.182}$$

$$\|\phi^{(0)}(x_i) + x\| \geq \|\phi^{(0)}(x_i)\| - \|x\| \gtrsim \sqrt{m}_3 \kappa_1 - C_1 \gtrsim \sqrt{m}_3 \kappa_1. \tag{3.183}$$

Hence, $V_j^{(0)}(\phi^{(0)}(x_i) + x)$ is Gaussian with standard deviation at least $\Omega(\kappa_2\sqrt{m}_3\kappa_1)$. Therefore,

$$\mathbb{P}(|V_j^{(0)}(\phi^{(0)}(x_i) + x)| \leq r) \lesssim \frac{r}{\sqrt{m}_3\kappa_1\kappa_2}.$$

This implies

$$\mathbb{E}[|\mathcal{J}_{x,r}|] \lesssim \frac{r}{\sqrt{m}_3\kappa_1\kappa_2} m_2. \tag{3.184}$$

On the other hand, note that $|\mathcal{J}_{x,r}|$ is the sum of $m_2$ Bernoulli random variables, so it is subGaussian with parameter $m_2$. Therefore, with high probability

$$|\mathcal{J}_{x,r}| \lesssim \frac{r}{\sqrt{m}_3\kappa_1\kappa_2} m_2 + \sqrt{m}_2.$$

Now taking maximum over all $x \in B_{C_1}(\epsilon)$ and exploiting the subGaussian tail of the random variables, we get with high probability

$$\max_{x\in B_{C_1}(\epsilon)} |\mathcal{J}_{x,r}| \lesssim \frac{r}{\sqrt{m}_3\kappa_1\kappa_2} m_2 + \sqrt{m_2 \log(|B_{C_1}(\epsilon)|)} \lesssim \frac{r}{\sqrt{m}_3\kappa_1\kappa_2} m_2 + \sqrt{m_2 m_3 \log(1/\epsilon)}. \tag{3.185}$$

Moreover, consider a threshold $1 < \theta$, such that $e^{-\theta^2/8} \leq m_2/m_3$, and define the following set of indices

$$\mathcal{J}^{(2)}_{x',\theta} \coloneqq \{j \in [m_2] |\ |V_j^{(0)} x'| \geq \theta\kappa_2 C_1\}.$$

Then, using Lemma 38 and noting the fact that the standard deviation of Gaussians in $V^{(0)}$ is $\kappa_2$ and that $\|\phi^{(2)}(x_i)\| \leq C_1$, with high probability:

$$\sup_{x':\ \|x'\|=1} |\mathcal{J}^{(2)}_{x',\theta}| \leq m_3(\log(m_3) + \log(\log(m_2))). \tag{3.186}$$

Now note that for each $j \in [m_2]$, $\|V_j^{(0)}\|^2$ is subexponential with parameters $(m_3\kappa_2^4, \kappa_2^2)$, which means that with high probability:

$$\max_j \|V_j^{(0)}\|^2 \lesssim m_3\kappa_2^2 + \sqrt{m}_3\kappa_2^2\sqrt{\log(m_2)} + \kappa_2^2 \log(m_2).$$

But with condition $m_3 \geq \log(m_2)$, we can further upper bound it as

$$\max_j \|V_j^{(0)}\|^2 \lesssim m_3 \kappa_2^2.$$

Now for fixed $x, x'$, for $j \in \mathcal{J}_{x,r}$ we have

$$\begin{aligned} |V_j^{(0)}(\phi^{(0)}(x_i) + x')| &\leq |V_j^{(0)}(\phi^{(0)}(x_i) + x)| + |V_j^{(0)}(x' - x)| \\ &\leq |V_j^{(0)}(\phi^{(0)}(x_i) + x)| + \|V_j^{(0)}\| \|x' - x\| \\ &\lesssim r + \sqrt{m}_3 \kappa_2 \epsilon. \end{aligned}$$

On the other hand, for $j \notin \mathcal{J}_{x',\theta}^{(2)}$:

$$|V_j^{(0)} x'| \leq \theta \kappa_2 C_1. \tag{3.187}$$

Therefore, for $j \in \mathcal{J}_{x,r} - \mathcal{J}_{x',\theta}^{(2)}$:

$$|V_j^{(0)} \phi^{(0)}(x_i)| \leq |V_j^{(0)}(\phi^{(0)}(x_i) + x')| + |V_j^{(0)} x'| \lesssim r + \sqrt{m}_3 \kappa_2 \epsilon + \theta \kappa_2 C_1. \tag{3.188}$$

In a similar fashion, if $j \notin \mathcal{J}_{x,r}$, then using assumption (3.181):

$$|V_j^{(0)}(\phi^{(0)}(x_i) + x')| \geq |V_j^{(0)}(\phi^{(0)}(x_i) + x)| - |V_j^{(0)}(x - x')| \gtrsim r - \sqrt{m}_3 \kappa_2 \epsilon \geq r/2. \tag{3.189}$$

Hence, using the fact that $\phi^{(0)}(x_i)$ is orthogonal to $V'_j$:

$$\begin{aligned}
&\left|\mathbb{1}\{(V_j^{(0)} + V'_j)(\phi^{(0)}(x_i) + x') \geq 0\} - \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i) + x') \geq 0\}\right| \\
&\leq \mathbb{1}\{|V'_j(\phi^{(0)}(x_i) + x')| \gtrsim |V_j^{(0)}(\phi^{(0)}(x_i) + x')|\} \\
&\leq \mathbb{1}\{|V'_j x'| \gtrsim |V_j^{(0)}(\phi^{(0)}(x_i) + x')|\} \\
&\leq \mathbb{1}\{\|V'_j\|\|x'\| \gtrsim |V_j^{(0)}(\phi^{(0)}(x_i) + x')|\} \\
&\leq \mathbb{1}\{\|V'_j\|C_1 \gtrsim |V_j^{(0)}(\phi^{(0)}(x_i) + x')|\} \\
&\leq \mathbb{1}\{\|V'_j\| \gtrsim \frac{|V_j^{(0)}(\phi^{(0)}(x_i) + x')|}{2C_1} + \frac{r}{4C_1}\}. \qquad (3.190)
\end{aligned}$$

Now by triangle inequality and Equations (3.190), (3.188), (3.187) and the fact that

$\|V'\|_F \leq C_2$, we can write:

$$
\begin{aligned}
&|\Gamma_{x'} - \Gamma_{x',V'}| \\
&\leq \frac{1}{\sqrt{m}_2} \sum_{j \in \mathcal{J}_{x,r} - \mathcal{J}^{(2)}_{x',\theta}} \Big| \mathbb{1}\{(V_j^{(0)} + V_j')(\phi^{(0)}(x_i) + x') \geq 0\} - \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i) + x') \geq 0\} \Big| |V_j^{(0)} \phi^{(0)}(x_i)| \\
&+ \frac{1}{\sqrt{m}_2} \sum_{j \notin (\mathcal{J}_{x,r} \cup \mathcal{J}^{(2)}_{x',\theta})} \Big| \mathbb{1}\{(V_j^{(0)} + V_j')(\phi^{(0)}(x_i) + x') \geq 0\} - \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i) + x') \geq 0\} \Big| |V_j^{(0)} \phi^{(0)}(x_i)| \\
&+ \frac{1}{\sqrt{m}_2} \sum_{j \in \mathcal{J}^{(2)}_{x',\theta}} \Big| \mathbb{1}\{(V_j^{(0)} + V_j')(\phi^{(0)}(x_i) + x') \geq 0\} - \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i) + x') \geq 0\} \Big| |V_j^{(0)} \phi^{(0)}(x_i)| \\
&\leq \frac{1}{\sqrt{m}_2} |\mathcal{J}_{x,r} - \mathcal{J}^{(2)}_{x',\theta}| \max_{j \in \mathcal{J}_{x,r} - \mathcal{J}^{(2)}_{x',\theta}} |V_j^{(0)} \phi^{(0)}(x_i)| \\
&+ \frac{1}{\sqrt{m}_2} \sum_{j \notin (\mathcal{J}_{x,r} \cup \mathcal{J}^{(2)}_{x',\theta})} \mathbb{1}\{\|V_j'\| \gtrsim \frac{|V_j^{(0)}(\phi^{(0)}(x_i) + x')|}{2C_1} + \frac{r}{4C_1}\} |V_j^{(0)} \phi^{(0)}(x_i)| \\
&+ \frac{1}{\sqrt{m}_2} |J^{(2)}_{x',\theta}| \max_{j \in m_2} |V_j^{(0)} \phi^{(0)}(x_i)| \\
&\leq \frac{1}{\sqrt{m}_2} |\mathcal{J}_{x,r} - \mathcal{J}^{(2)}_{x',\theta}| \max_{j \in \mathcal{J}_{x,r} - \mathcal{J}^{(2)}_{x',\theta}} |V_j^{(0)} \phi^{(0)}(x_i)| \\
&+ \frac{1}{\sqrt{m}_2} \sum_{j \notin (\mathcal{J}_{x,r} \cup \mathcal{J}^{(2)}_{x',\theta})} \mathbb{1}\{\|V_j'\| \gtrsim \frac{|V_j^{(0)}(\phi^{(0)}(x_i) + x')|}{2C_1} + \frac{r}{4C_1}\} (|V_j^{(0)}(\phi^{(0)}(x_i) + x')| + \theta \kappa_2 C_1) \\
&+ \frac{1}{\sqrt{m}_2} |J^{(2)}_{x',\theta}| \max_{j \in m_2} |V_j^{(0)} \phi^{(0)}(x_i)| \\
&\lesssim \frac{1}{\sqrt{m}_2} |\mathcal{J}_{x,r} - \mathcal{J}^{(2)}_{x',\theta}| \max_{j \in \mathcal{J}_{x,r} - \mathcal{J}^{(2)}_{x',\theta}} |V_j^{(0)} \phi^{(0)}(x_i)| \\
&+ \frac{1}{\sqrt{m}_2} \sum_{j \notin (\mathcal{J}_{x,r} \cup \mathcal{J}^{(2)}_{x',\theta})} \mathbb{1}\{\|V_j'\| \gtrsim \frac{r}{C_1}\} (C_1 \|V_j'\| + \theta \kappa_2 C_1) \\
&+ \frac{1}{\sqrt{m}_2} |J^{(2)}_{x',\theta}| \max_{j \in m_2} |V_j^{(0)} \phi^{(0)}(x_i)| \\
&\lesssim \frac{1}{\sqrt{m}_2} |\mathcal{J}_{x,r} - \mathcal{J}^{(2)}_{x',\theta}| (r + \sqrt{m}_3 \kappa_2 \epsilon + \theta \kappa_2 C_1) + \frac{C_1}{\sqrt{m}_2} \sqrt{\#\Big(j :\ \|V_j'\| \gtrsim \frac{r}{C_1}\Big) \|V'\|_F^2}
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{\sqrt{m}_2} \#\Big(j : \ \|V_j'\| \gtrsim \frac{r}{C_1}\Big)\theta\kappa_2 C_1 + \frac{1}{\sqrt{m}_2}|\mathcal{J}^{(2)}_{x',\theta}| \max_{j\in m_2} |V_j^{(0)}\phi^{(0)}(x_i)| \\
&\lesssim \frac{1}{\sqrt{m}_2}|\mathcal{J}_{x,r} - \mathcal{J}^{(2)}_{x',\theta}|(r + \sqrt{m}_3\kappa_2\epsilon + \theta\kappa_2 C_1) + \frac{C_1^2 C_2^2}{\sqrt{m}_2 r} \\
&+ \frac{C_1^3 C_2^2}{\sqrt{m}_2 r^2}\theta\kappa_2 + \frac{1}{\sqrt{m}_2}|\mathcal{J}^{(2)}_{x',\theta}| \max_{j\in m_2} |V_j^{(0)}\phi^{(0)}(x_i)|.
\end{aligned}
$$

Now using Equations (3.197), (3.186), (3.179), and (3.181), and the bound on $|\mathcal{J}^{(2)}_{x',\theta}|$ from Lemma 38, we write

$$
\begin{aligned}
&\lesssim \frac{1}{\sqrt{m}_2}|\mathcal{J}_{x,r}|(r + \theta\kappa_2 C_1) + \frac{1}{\sqrt{m}_2}|\mathcal{J}^{(2)}_{x',\theta}|\kappa_1\kappa_2\sqrt{m_3\log(m_2)} + \frac{C_1^2C_2^2}{\sqrt{m}_2 r} + \frac{C_1^3C_2^2}{\sqrt{m}_2 r^2}\theta\kappa_2 \\
&\leq \frac{1}{\sqrt{m}_2}\Big(\frac{r}{\sqrt{m}_3\kappa_1\kappa_2}m_2 + \sqrt{m_2 m_3\log(1/\epsilon)}\Big)(r + \theta\kappa_2 C_1) \\
&+ \frac{1}{\sqrt{m}_2}\Big(m_3(\log(m_3) + \log(\log(m_2)))\Big)\kappa_1\kappa_2\sqrt{m_3\log(m_2)} + \frac{C_1^2C_2^2}{r\sqrt{m}_2} + \frac{C_1^3C_2^2}{\sqrt{m}_2 r^2}\theta\kappa_2. \\
&\leq \frac{1}{\sqrt{m}_2}\Big(\frac{r}{\sqrt{m}_3\kappa_1\kappa_2}m_2 + \sqrt{m_2 m_3\log(1/\epsilon)}\Big)(r + \theta\kappa_2 C_1) \\
&+ \frac{m_3^{3/2}\kappa_1\kappa_2}{\sqrt{m}_2}\sqrt{\log(m_2)}(\log(m_3) + \log(\log(m_2))) + \frac{C_1^2C_2^2}{r\sqrt{m}_2} + \frac{C_1^3C_2^2}{\sqrt{m}_2 r^2}\theta\kappa_2. \qquad (3.191)
\end{aligned}
$$

Now setting

$$
r^* := (C_1C_2)^{2/3}\frac{m_3^{1/6}(\kappa_1\kappa_2)^{1/3}}{m_2^{1/3}}.
$$

By this choice, from (3.191) we obtain

$$
\begin{aligned}
|\Gamma_{x'} - \Gamma_{x',V'}| \leq \Big(&\frac{(C_1C_2)^{4/3}}{(\sqrt{m}_2\kappa_2)^{1/3}(\sqrt{m}_3\kappa_1)^{1/3}} + \frac{(C_1C_2)^{2/3}m_3^{2/3}(\kappa_1\kappa_2)^{1/3}\sqrt{\log(1/\epsilon)}}{m_2^{1/3}}\Big)\Big(1+\theta\frac{C_1^{1/3}(\kappa_2\sqrt{m}_2)^{2/3}}{C_2^{2/3}(\kappa_1\sqrt{m}_3)^{1/3}}\Big) \\
&+ \frac{m_3^{3/2}\kappa_1\kappa_2}{\sqrt{m}_2}\sqrt{\log(m_2)}(\log(m_3) + \log(\log(m_2))).
\end{aligned}
$$

Now we set

$$
\theta^* := 3\log(m_2),
$$

which also satisfies the condition of Lemma 38 and combining with Equation (3.180), we get that with high probability

$$\begin{aligned}|\Gamma_{x',V'}| &\leq |\Gamma_{x',V'} - \Gamma_{x'}| + |\Gamma_{x'}| \\ &\lesssim \Big(\frac{(C_1C_2)^{4/3}}{(\sqrt{m}_2\kappa_2)^{1/3}(\sqrt{m}_3\kappa_1)^{1/3}} + \frac{(C_1C_2)^{2/3}m_3^{2/3}(\kappa_1\kappa_2)^{1/3}\sqrt{\log(1/\epsilon)}}{m_2^{1/3}}\Big) \\ &\times \Big(1 + \log(m_2)\frac{C_1^{1/3}(\kappa_2\sqrt{m}_2)^{2/3}}{C_2^{2/3}(\kappa_1\sqrt{m}_3)^{1/3}}\Big) \\ &+ \frac{m_3^{3/2}\kappa_1\kappa_2}{\sqrt{m}_2}\sqrt{\log(m_2)}(\log(m_3) + \log(\log(m_2))) + \kappa_1\kappa_2\sqrt{m_3\log(m_2)},\end{aligned}$$

where $\|x'\| \lesssim C_2$ and $\|V'\|_F \leq C_2$, $\forall j : \ V_j'\phi^{(0)}(x_i) = 0$. We also need to satisfy condition (3.181), which regarding this choice for $\theta = \theta^*$ becomes

$$r^* = (C_1C_2)^{2/3}\frac{m_3^{1/6}(\kappa_1\kappa_2)^{1/3}}{m_2^{1/3}} \geq 2\sqrt{m}_3\kappa_2\epsilon, \tag{3.192}$$

for which it suffices to set

$$\epsilon^* := (C_1C_2)^{2/3}\frac{\kappa_1^{1/3}}{2(m_2m_3)^{1/3}\kappa_2^{2/3}},$$

Substituting this choice of $\epsilon$ above and picking the overparameterization large enough to dominate the magnitude of $C_1, C_2$ so that $\log(1/\epsilon^*) \lesssim \log(m_2)$, the proof is complete.

**Lemma 32.** *Under the following condition*

$$(\sqrt{m}_2\kappa_2)^{1/3}(\sqrt{m}_3\kappa_1)^{2/3} \geq \log^{-7/6}(m_2)(C_1C_2)^{1/3},$$

*with high probability we have*

$$\begin{aligned}&\sup_{\|x'\|\lesssim C_1, \|V'\|_F \leq C_2, V' \perp \phi^{(0)}} \frac{1}{\sqrt{m}_2}\|\sum_j a_j \mathbb{1}\{(V_j^{(0)}+V_j')(\phi^{(0)}(x_i)+x') \geq 0\}V_j^{(0)}\| \\ &\lesssim \sqrt{m}_3\kappa_2\log(m_2) + \frac{(\sqrt{m}_2\kappa_2)^{1/3}}{(\sqrt{m}_3\kappa_1)^{2/3}}(C_1C_2)^{2/3}\log^{1/6}(m_2) := \Re_7.\end{aligned}$$

**Proof of Lemma 32**

Similar to Lemma 31, define the helper functions $\Gamma_{x'}$ and $\Gamma_{x',V'}$ as

$$\Gamma_{x',V'} = \frac{1}{\sqrt{m}_2}\|\sum_j a_j \mathbb{1}\{(V_j^{(0)}+V_j')(\phi^{(0)}(x_i)+x') \geq 0\}V_j^{(0)}\|, \tag{3.193}$$

$$\Gamma_{x'} = \frac{1}{\sqrt{m}_2}\|\sum_j a_j \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i)+x') \geq 0\}V_j^{(0)}\|. \tag{3.194}$$

First we bound $\sup_{x'}\Gamma_{x'}$. To this end, note that because $V_j^{(0)} \in \mathbb{R}^{m_3}$ and the VC-dimension of half-planes is $m_3+1$, then by Sauer's Lemma, the set

$$\mathcal{D} = \{D_{V^{(0)},x_i}^{x'} |\ x' \in \mathbb{R}^{m_3}, \|x'\| \lesssim C_1\}$$

of all sign pattern matrices has cardinality at most

$$|\mathcal{D}| \leq m_2^{m_3+1}.$$

Now note that with high probability, the entries of the matrix $V^{(0)}$ are all less than $O(\kappa_2\sqrt{\log(m_2m_3)})$. On the other hand, for each fixed sign pattern $D_{V^{(0)},x_i}^{x'}$, we have for the sum with respect to this sign pattern:

$$\|\frac{1}{\sqrt{m}_2}\sum_j a_j \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i)+x') \geq 0\}V_j^{(0)}\|^2 \tag{3.195}$$

is $(m_3\kappa_2^4\log^2(m_2m_3), \kappa_2^2\log(m_2m_3))$ sub-exponential with respect to the randomness of $a$, because each entry of the vector $\frac{1}{\sqrt{m}_2}\sum_j a_j \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i)+x') \geq 0\}V_j^{(0)}$ is $(\kappa_2\sqrt{\log(m_2m_3)})$-subGaussian. Therefore, with high probability we have

$$\begin{aligned}
&\|\frac{1}{\sqrt{m}_2}\sum_j a_j \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i)+x') \geq 0\}V_j^{(0)}\|^2 \\
&\leq \mathbb{E}_a[\|\frac{1}{\sqrt{m}_2}\sum_j a_j \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i)+x') \geq 0\}V_j^{(0)}\|^2] + \text{deviation} \\
&\lesssim m_3\kappa_2^2\log(m_2m_3) + \sqrt{m}_3\kappa_2^2\log(m_2m_3) + \kappa_2^2\log(m_2m_3).
\end{aligned}$$

Similarly, if we take a union bound over all sign matrices in $\mathcal{D}$ and using the fact that $m_2 > m_3$:

$$\begin{aligned}
&\sup_{x'}\Gamma_{x'}^2 = \sup_{x'\in\mathbb{R}^{m_3}}\|\frac{1}{\sqrt{m}_2}\sum_j a_j \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i)+x') \geq 0\}V_j^{(0)}\|^2 \\
&\lesssim m_3\kappa_2^2\log(m_2m_3) + \sqrt{m}_3\kappa_2^2\log(m_2m_3)\sqrt{\log(m_2^{m_3+1})} + \kappa_2^2\log(m_2m_3)\log(m_2^{m_3+1}) \\
&\lesssim m_3\kappa_2^2\log^2(m_2),
\end{aligned}$$

which implies

$$\sup_{x'}\Gamma_{x'} \lesssim \sqrt{m}_3\kappa_2\log(m_2). \tag{3.196}$$

Moreover, defining $\mathcal{J}_{v,x}$ similar to Lemma 31 and using the similar approach we get with high probability

$$\max_{x\in B_{C_1}(\epsilon)}|\mathcal{J}_{x,r}| \lesssim \frac{r}{\sqrt{m}_3\kappa_1\kappa_2}m_2 + \sqrt{m_2\log(|B_{C_1}(\epsilon)|)} \lesssim \frac{r}{\sqrt{m}_3\kappa_1\kappa_2}m_2 + \sqrt{m_2m_3\log(1/\epsilon)}. \tag{3.197}$$

Now for simplifying the analysis, we assume that for indices $j \in \mathcal{J}_{x,r}$ we can change the sign pattern with no cost on $V'$, i.e. we can pick any subset of them. Therefore, we first compute a high probability upper bound on the following quantity:

$$\frac{1}{\sqrt{m_2}} \sup_{S \subset \mathcal{J}_{x,r}, \pm \text{signs}} \| \sum_{j \in S} \pm V_j^{(0)} \|.$$

If we form the matrix $V^{(0)}(\mathcal{J}_{x,r})$ be the matrix which only keeps the rows with indices in $\mathcal{J}_{x,r}$, then the above quantity can be computed as

$$\frac{1}{\sqrt{m_2}} \sup_{S \subset \mathcal{J}_{x,r}, \pm \text{signs}} \| \sum_{j \in S} \pm V_j^{(0)} \| = \frac{1}{\sqrt{m_2}} \sup_{v \in \{0,1,-1\}^{|\mathcal{J}_{x,r}|}} \| v^T V^{(0)}(\mathcal{J}_{x,r}) \|$$
$$\leq \frac{1}{\sqrt{m_2}} \lambda_{\max}(V^{(0)}(\mathcal{J}_{x,r})) \sup \|v\| \leq \frac{1}{\sqrt{m_2}} \lambda_{\max}(V^{(0)}(\mathcal{J}_{x,r})) |\mathcal{J}_{x,r}|,$$

where $\lambda_{max}$ is the maximum singular value of the matrix. Now by random matrix theory, we know for a fixed $x$ and arbitrary $t \geq 0$, the following argument holds:

$$\mathbb{P}(\lambda_{\max}(V^{(0)}(\mathcal{J}_{x,r}))/\kappa_2 \gtrsim \sqrt{m_3} + \sqrt{|\mathcal{J}_{x,r}| + t}) \leq 2e^{-ct^2}. \tag{3.198}$$

Therefore, as $|\mathcal{D}| \leq m_2^{m_3+1}$, we get with high probability

$$\max_{x \in B_{C_1}(\epsilon)} \lambda_{\max}(V^{(0)}(\mathcal{J}_{x,r})) \lesssim \kappa_2(\sqrt{m}_3 + \sqrt{|\mathcal{J}_{x,r}|} + \sqrt{\log(m_2^{m_3+1})})$$
$$\leq \kappa_2 \sqrt{\log(m_2) m_3} + \kappa_2 \sqrt{|\mathcal{J}_{x,r}|}.$$

Therefore, with high probability

$$\sup_{x \in B_{C_1}(\epsilon)} \frac{1}{\sqrt{m_2}} \sup_{S \subset \mathcal{J}_{x,r}} \| \sum_{j \in S} V_j^{(0)} \| \leq \frac{\kappa_2}{\sqrt{m_2}} (\sqrt{\log(m_2) m_3 |\mathcal{J}_{x,r}|} + |\mathcal{J}_{x,r}|). \tag{3.199}$$

On the other hand, as in Equation (3.189) in the proof of Lemma 31, for $j \notin \mathcal{J}_{x,r}$ we have:

$$|V_j^{(0)}(\phi^{(0)}(x_i) + x')| \geq |V_j^{(0)}(\phi^{(0)}(x_i) + x)| - |V_j^{(0)}(x - x')| \gtrsim r - \sqrt{m}_3 \kappa_2 \epsilon. \tag{3.200}$$

Picking

$$\epsilon^* := \frac{r}{2\sqrt{m_3}\kappa_2},$$

we get for $j \notin \mathcal{J}_{x,r}$

$$|V_j^{(0)}(\phi^{(0)}(x_i) + x')| \gtrsim r.$$

Now similar to the derivation in (3.190) we have

$$\begin{aligned}
&\left|\mathbb{1}\{(V_j^{(0)} + V_j')(\phi^{(0)}(x_i) + x') \geq 0\} - \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i) + x') \geq 0\}\right| \\
&\leq \mathbb{1}\{\|V_j'\|C_1 \gtrsim |V_j^{(0)}(\phi^{(0)}(x_i) + x')|\} \\
&\leq \mathbb{1}\{\|V_j'\| \gtrsim \frac{r}{C_1}\}.
\end{aligned}$$

Hence, because $\|V'\|_F \leq C_2$, the number of indices for which $\mathbb{1}\{(V_j^{(0)} + V_j')(\phi^{(0)}(x_i) + x') \geq 0\} \neq \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i) + x') \geq 0\}$ is at most $l = \frac{(C_1 C_2)^2}{r^2}$. Therefore, we bound the following quantity to use in the analysis:

$$\sup_{S \subset [m_2] \ \& \ |S| \leq l, \pm \text{signs}} \|\sum_{j \in S} \pm V_j^{(0)}\|. \tag{3.201}$$

But if we define for $m_2 < j \leq 2m_2$,

$$V_j^{(0)} = -V_{j-m_2}^{(0)},$$

then

$$\sup_{S \subset [m_2] \ \& \ |S| \leq l, \pm \text{signs}} \|\sum_{j \in S} \pm V_j^{(0)}\| \leq \sup_{S \subset [2m_2] \ \& \ |S| \leq l} \|\sum_{j \in S} V_j^{(0)}\|.$$

Now note that each entry of $\sum_{j \in S} V_j^{(0)}$ is $\sqrt{l}\kappa_2$ subGaussian. Hence, the quantity

$\| \sum_{j\in S} V_j^{(0)} \|^2$ is $(m_3 l^2 \kappa_2^4, l\kappa_2^2)$ subexponential. Therefore, we have with high probability

$$\begin{aligned}
\sup_{|S|\le l} \| \sum_{j\in S} V_j^{(0)} \|^2 &\lesssim \mathbb{E} \| \sum_{j\in S} V_j^{(0)} \|^2 + \sqrt{m}_3 l \kappa_2^2 \sqrt{\log \binom{2m_2}{l}} + l\kappa_2^2 \log \binom{2m_2}{l} \\
&\lesssim m_3 l \kappa_2^2 + \sqrt{m}_3 l \kappa_2^2 \sqrt{\log \binom{2m_2}{l}} + l\kappa_2^2 \log \binom{2m_2}{l} \\
&\lesssim m_3 l \kappa_2^2 + \sqrt{m}_3 l^{3/2} \kappa_2^2 \sqrt{\log(m_2)} + l^2 \kappa_2^2 \log(m_2).
\end{aligned}$$

Hence

$$\sup_{|S|\le l} \| \sum_{j\in S} V_j^{(0)} \| \lesssim \sqrt{m}_3 \sqrt{l} \kappa_2 + l\kappa_2 \sqrt{\log(m_2)}.$$

Now using Equation , we can write

$$\begin{aligned}
&|\Gamma_{x'} - \Gamma_{x',V'}| \\
&\le \frac{1}{\sqrt{m}_2} \| \sum_{j\in \mathcal{J}_{x,r}} (\mathbb{1}\{(V_j^{(0)} + V_j')(\phi^{(0)}(x_i) + x') \ge 0\} - \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i) + x') \ge 0\}) V_j^{(0)} \| \\
&+ \frac{1}{\sqrt{m}_2} \| \sum_{j\notin \mathcal{J}_{x,r}} (\mathbb{1}\{(V_j^{(0)} + V_j')(\phi^{(0)}(x_i) + x') \ge 0\} - \mathbb{1}\{V_j^{(0)}(\phi^{(0)}(x_i) + x') \ge 0\}) V_j^{(0)} \| \\
&\le \frac{1}{\sqrt{m}_2} \sup_{S\subset \mathcal{J}_{x,r}, \pm\text{signs}} \| \sum_{j\in S} \pm V_j^{(0)} \| \\
&+ \frac{1}{\sqrt{m}_2} \sup_{S\subset [m_2] \ \& \ |S| \le (\frac{C_1 C_2}{r})^2, \pm\text{signs}} \| \sum_{j\in S} \pm V_j^{(0)} \| \\
&\le \frac{\kappa_2}{\sqrt{m}_2} ( \sqrt{\log(m_2) m_3 |\mathcal{J}_{x,r}|} + |\mathcal{J}_{x,r}|) + \frac{1}{\sqrt{m}_2} \Big( \sqrt{m}_3 \sqrt{l} \kappa_2 + l\kappa_2 \sqrt{\log(m_2)} \Big) \\
&\le \frac{\kappa_2}{\sqrt{m}_2} \Big( \frac{r m_2}{\sqrt{m}_3 \kappa_1 \kappa_2} + \sqrt{m_2 m_3 \log(\frac{\sqrt{m}_3 \kappa_2}{r})} \Big)^{1/2} ( \sqrt{\log(m_2) m_3} + \sqrt{|\mathcal{J}_{x,r}|}) \\
&+ \frac{1}{\sqrt{m}_2} \Big( \sqrt{m}_3 (C_1 C_2 / r) \kappa_2 + (C_1 C_2 / r)^2 \kappa_2 \sqrt{\log(m_2)} \Big) \\
&\lesssim \frac{\kappa_2}{\sqrt{m}_2} \Big( \frac{r m_2}{\sqrt{m}_3 \kappa_1 \kappa_2} + \sqrt{m_2 m_3 \log(\frac{\sqrt{m}_3 \kappa_2}{r})} \Big) \\
&+ \frac{1}{\sqrt{m}_2} \Big( \sqrt{m}_3 (C_1 C_2 / r) \kappa_2 + (C_1 C_2 / r)^2 \kappa_2 \sqrt{\log(m_2)} \Big).
\end{aligned}$$

Combining this with (3.196):

$$|\Gamma_{x',V'}| \lesssim \sqrt{m}_3\kappa_2 \log(m_2) \tag{3.202}$$

$$+ \frac{r\sqrt{m}_2}{\sqrt{m}_3\kappa_1} + \kappa_2\sqrt{m_3 \log(\frac{\sqrt{m_3\kappa_2}}{r})} + \frac{1}{\sqrt{m}_2}\Big(\sqrt{m}_3(C_1C_2/r)\kappa_2 + (C_1C_2/r)^2\kappa_2\sqrt{\log(m_2)}\Big). \tag{3.203}$$

Now setting

$$r^* := \frac{m_3^{1/6}(\kappa_1\kappa_2)^{1/3}}{m_2^{1/3}}(C_1C_2)^{2/3}\log^{1/6}(m_2),$$

we get

$$LHS \lesssim \sqrt{m}_3\kappa_2\log(m_2) + \frac{(\sqrt{m}_2\kappa_2)^{1/3}}{(\sqrt{m}_3\kappa_1)^{2/3}}(C_1C_2)^{2/3}\log^{1/6}(m_2) \tag{3.204}$$

$$+ \kappa_2 m_3^{1/2}\log^{1/2}(m_2) + \frac{m_3^{1/3}\kappa_2^{2/3}(C_1C_2)^{1/3}}{m_2^{1/6}\kappa_1^{1/3}\log^{1/6}(m_2)} \tag{3.205}$$

$$\lesssim \sqrt{m}_3\kappa_2\log(m_2) + \frac{(\sqrt{m}_2\kappa_2)^{1/3}}{(\sqrt{m}_3\kappa_1)^{2/3}}(C_1C_2)^{2/3}\log^{1/6}(m_2) \tag{3.206}$$

$$+ \frac{m_3^{1/2}\kappa_2(C_1C_2)^{1/3}}{(\sqrt{m}_2\kappa_2)^{1/3}(\sqrt{m}_3\kappa_1)^{1/3}\log^{1/6}(m_2)}. \tag{3.207}$$

Now under the condition

$$(\sqrt{m}_2\kappa_2)^{1/3}(\sqrt{m}_3\kappa_1)^{2/3} \geq \log^{-7/6}(m_2)(C_1C_2)^{1/3},$$

The final term is dominated by the first term, which finally completes the proof.

### 3.6.16 Convergence

The goal of this section is to prove Theorem 7.

**Theorem 7.** *Letting $\aleph = 4B^2$, by Corollary 8.1, we have $L^{\Pi}(0) \leq \aleph$. We define the domain $\mathcal{D}_l := \{\|w'\| \leq C_1 := \frac{\aleph+4l}{\psi_1},\ \|v'\| \leq C_2 := \frac{\aleph+4l}{\psi_2}\}$. For a large enough constant $l = O(1)$ and function $L^{\Pi}(w := (w', v')) : \mathbb{R}^N \to \mathbb{R}$. Moreover, suppose $L^{\Pi}$ is $\rho_1$ Lipschitz, $\rho_2$ gradient Lipschitz, and $\rho_3$ hessian Lipschitz in the domain $\mathcal{D}_l$ ($\rho_1, \rho_2, \rho_3 \geq 1$), in the sense that their first, second, and third directional derivatives in an arbitrary unit direction is bounded by the corresponding parameters. Suppose we have access to the gradient of $L^{\Pi}$ at each point in $\mathcal{D}_l$ plus a zero mean noise vector $\pounds$ such that $\sigma_1^2 I \leq \mathbb{E}\pounds\pounds^T \leq \sigma_2^2 I$ and $\|\pounds\| \leq Q$ almost surely. Also, suppose for a threshold $\aleph_\ell \leq \aleph$, if $L^{\Pi}(w) \geq \aleph_\ell$ and $w \in \mathcal{D}_l$, then we have at least one of the following conditions holds:*

$$(1)\ \|\nabla L^{\Pi}(w)\| \geq \frac{\nu}{16\sqrt{C_1^2 + C_2^2}}, \tag{3.208}$$

$$(2)\ \lambda_{min}(\nabla^2 L^{\Pi}(w)) \leq -\gamma. \tag{3.209}$$

Then starting from $w_0 = 0$, with probability at least 0.999 after at most $poly(\rho_1, \rho_2, \rho_3, Q, \aleph, C_1, C_2, 1/\gamma, \log(\sigma_1/\sigma_2))$ number of iterations, we reach a point $w_t$ such that $L^{\Pi}(w_t) \leq \aleph_\ell$.

Our proof here is a refined version of that in Ge et al. [2015a]. As we mentioned in section 3.5, the key fact that we are using in the other parts of our proof is a uniform upper bound $\|w'\| \leq C_1$, $\|v'\| \leq C_2$ which is unjustified by only naively using Ge et al. [2015a]. Here, first we restate a refined version of Lemmas 14 and 16 in Ge et al. [2015a] in Lemmas 33 and 35 respectively, and then use them to also bound the upward deviations of $L^{\Pi}$. Moreover, to avoid writing repeated proofs and overwhelm the reader, we mostly treat the arguments in Lemma 16 of Ge et al. [2015a] as blackbox and use them for our purpose here. A point to mention before we start, unlike Lemma 14 of Ge et al. [2015a] where the dependency on other parameters than the step size $\eta$ is more explicit, Lemma 16 hides the dependencies on all the other parameters (which is polynomial). Here, we follow the same style.

We refer to the trajectory of the steps of algorithm by $(w_t)_{t\geq 0}$. In Lemmas of this section, To avoid introducing new notation and complicating things, we refer to the current point of the algorithm by $w_0$, while for the next point of the algorithm we use $w_1$ (in Lemma 33), and $w_T$ (in Lemma 35) respectively. Also, similar to Ge et al. [2015a], $\tilde{O}$ and $\tilde{\Omega}$ below means we are looking at the dependency on $\eta$.

**Lemma 33.** *Suppose $L^{\Pi}(w_0) \leq \aleph + 2l$, and consider a parameter $\chi > 1$ which can be set arbitrarily. For every point $w_0$ such that $L^{\Pi}(w_0) \leq \aleph + 2l$, $\|\nabla L^{\Pi}(w_0)\| \geq 2\sqrt{\eta(Q^2 + \sigma_2^2 N)\rho_2\rho_1^2(2\chi + \frac{1}{2})}$, then for $w_1 := w_0 - \eta(\nabla L^{\Pi}(w_0) + \pounds)$ and random variable $\mathfrak{R}_1$ (depending on $w_0$) defined as*

$$\mathbb{E}L^{\Pi}(w_1) - L^{\Pi}(w_0) = -\eta^2\mathfrak{R}_1^2, \tag{3.210}$$

*we have $\mathfrak{R}_1 = \tilde{\Omega}(1)$ a.s., and almost surely:*

$$\Big|L^{\Pi}(w_1) - L^{\Pi}(w_0)\Big| \leq \eta\mathfrak{R}_1/\sqrt{\rho_2\chi}.$$

*(the expectation is over the randomness of $\pounds$).*

This lemma is a tuned version of Lemma 14 in Ge et al. [2015a]. First, note that the condition $L^{\Pi}(w_0) \leq \aleph + 2l$ assures the smoothness coefficients $\rho_1, \rho_2$ and $\rho_3$ for $L^{\Pi}$ by Corollary **??**. We follow similar to Ge et al. [2015a] (picking $\eta < 1/(2\rho_2)$):

$$\begin{aligned}\mathbb{E}L^{\Pi}(w_1) - L^{\Pi}(w_0) &\leq -\frac{\eta}{2}\|\nabla L^{\Pi}(w_0)\|^2 + \frac{\eta^2\sigma_2^2\rho_2 N}{2}\\ &\leq -\frac{\eta}{4}\|\nabla L^{\Pi}(w_0)\|^2 - \eta^2(\sigma_2^2 N + Q^2)\rho_2\rho_1^2(2\chi + \frac{1}{2}) + \frac{\eta^2\sigma_2^2\rho_2 N}{2}\\ &\leq -\frac{\eta}{4}\|\nabla L^{\Pi}(w_0)\|^2 - 2\eta^2 Q^2\rho_2\rho_1^2\chi. \end{aligned} \tag{3.211}$$

where we used the fact that $\rho_1 \geq 1$. On ther other hand, $L^{\Pi}$ is $\rho_1$ Lipcshitz, so we have almost surely

$$
\begin{aligned}
&|L^{\Pi}(w_1) - L^{\Pi}(w_0)| \leq \rho_1\eta(\|\nabla L^{\Pi}(w_0) + \pounds\|) \leq \rho_1\eta(\|\nabla L^{\Pi}(w_0)\| + \|\pounds\|) \\
&\leq \rho_1\eta(\|\nabla L^{\Pi}(w_0)\| + Q). \qquad (3.212)
\end{aligned}
$$

(To be completely precise, we should justify that we can write the Lipschitz inequality at point $w_0$, we also need to make sure that $w_1$ remains in the domain that we have the Lipschitz parameter in, i.e. $\mathcal{D}_l$. To see why this is true, see the next Corollary).

Therefore

$$
(\rho_2\chi)\Big|L^{\Pi}(w_1) - L^{\Pi}(w_0)\Big|^2 \leq 2\eta^2\rho_1^2\rho_2\chi\|\nabla L^{\Pi}(w_0)\|^2 + 2\rho_2\chi\eta^2\rho_1^2Q^2. \qquad (3.213)
$$

Taking

$$
\eta \leq (8\rho_1^2\rho_2\chi)^{-1}, \qquad (3.214)
$$

we get from Equation (3.211):

$$
\mathbb{E}L^{\Pi}(w_1) - L^{\Pi}(w_0) \leq -2\eta^2\rho_1^2\rho_2\chi\|\nabla L^{\Pi}(w_0)\|^2 - 2\rho_2\chi\eta^2.\rho_1^2Q^2. \qquad (3.215)
$$

Combining Equations (3.213) and (3.215), we see that

$$
(\rho_2\chi)\Big|L^{\Pi}(w_1) - L^{\Pi}(w_0)\Big|^2 \leq -(\mathbb{E}L^{\Pi}(w_1) - L^{\Pi}(w_0)).
$$

Hence, if we define

$$
\mathbb{E}L^{\Pi}(w_1) - L^{\Pi}(w_0) := -\eta^2\mathfrak{R}_1^2,
$$

we get

$$
\Big|L^{\Pi}(w_1) - L^{\Pi}(w_0)\Big| \leq \eta\mathfrak{R}_1/\sqrt{\rho_2\chi}, \qquad (3.216)
$$

and from Equation (3.215), that

$$\mathfrak{R}_1^2 \geq 2\rho_1^2\rho_2\chi\|\nabla L^{\Pi}(w_0)\|^2 + 2\rho_2\chi\rho_1^2 Q^2 \geq 2\rho_2\chi\rho_1^2 Q^2 = \tilde{\Omega}(1).$$

Moreover, because the function is $\rho_1$-Lipshitz at the domain point $w_0$, we get from Equation (3.212):

$$-\eta^2\mathfrak{R}_1^2 = \mathbb{E}L^{\Pi}(w_1) - L^{\Pi}(w_0) \geq -\eta\rho_1(\|\nabla L^{\Pi}(w_0)\| + Q) \geq -\eta\rho_1(\rho_1 + Q) \geq -\frac{1}{\rho_2\chi},$$

by taking

$$\eta \leq (\rho_1(\rho_1 + Q)\rho_2\chi)^{-1},$$

which implies

$$\eta\mathfrak{R}_1 \leq \frac{1}{\sqrt{\rho_2\chi}}.$$

This, combined with Equation (3.216) and triangle inequality implies:

$$\Big|L^{\Pi}(w_1) - \mathbb{E}L^{\Pi}(w_1)\Big| \leq \eta\mathfrak{R}_1/\sqrt{\rho_2\chi} + \eta^2\mathfrak{R}_1^2 \leq 2\eta\mathfrak{R}_1/\sqrt{\rho_2\chi}. \tag{3.217}$$

**Lemma 34.** *As long as the value of the function at some $w$ is bounded by $\aleph + 2l$ ($L^{\Pi}(w) \leq \aleph + 2l$), then $\eta$ can be picked small enough (polynomially in other parameters), namely $\eta \leq l/(\|\nabla L^{\Pi}(w_0)\| + Q)$, so that the change of the function by a step is bounded by $l$.*

Let $\psi = \min\{\psi_1, \psi_2\}$. First, note that as the function is bounded by $\aleph + 2l$, we have the Lipschitz parameter $\rho_1$, hence $\|\nabla L^{\Pi}(w)\| \leq \rho_1$. Therefore, the change in $w$ in a step is bounded as

$$\|\nabla L^{\Pi}(w) + \pounds\| \leq Q + \rho_1.$$

So by picking

$$\eta \leq \Big(\sqrt{\frac{\aleph + 3l}{\psi}} - \sqrt{\frac{\aleph + 2l}{\psi}}\Big)/(Q + \rho_1),$$

we guarantee that the value of $w$ after a step remains in the ball of radius $\sqrt{\frac{\aleph+3l}{\psi}}$, hence we still have the smoothing parameters even after one step. Therefore, now we can use the Lipcshitz parameter $\rho_1$ to bound the value of the function after one step as it is written in Equation Equation (3.212). Using this Equation, it is enough to pick $\eta$ as small as:

$$\eta \leq l/(\|\nabla L^{\Pi}(w_0)\| + Q), \tag{3.218}$$

so that the change in the function would be at most $l$ as desired.

**Lemma 35.** *For a fixed point $w_0$ s.t. $L^{\Pi}(w_0) \leq \aleph + 2l$, suppose we pick $\eta$ small enough such that*

$$\S(\eta) := 2\sqrt{\eta(Q^2 + \sigma_2^2 N)\rho_2\rho_1^2(2\chi + \frac{1}{2})} < \frac{\nu}{16\sqrt{C_1^2 + C_2^2}}.$$

*Then, note that for $\|\nabla L^{\Pi}(w_0)\| \leq \S(\eta)$, condition 3.209 implies:*

$$\lambda_{min}\Big(\nabla^2 L^{\Pi}(w_0)\Big) \leq \gamma.$$

*Then, using the notation $\mathfrak{E}_T$ for the high probability event corresponding to Equations (36) and (44) in Ge et al. [2015a], for small enough $\eta$ (polynomially small w.r.t other parameters), for $\mathfrak{R}_2$ defined as*

$$\mathbb{E}[L^{\Pi}(w_T) - L^{\Pi}(w_0)]\mathbb{1}\{\mathfrak{E}_T\} = -\mathfrak{R}_2^2\eta^2, \tag{3.219}$$

*we have almost surely*

$$\Big|[L^{\Pi}(w_T) - L^{\Pi}(w_0)]\mathbb{1}\{\mathfrak{E}_T\}\Big| \leq \mathfrak{R}_2\eta/\sqrt{\rho_2\chi}. \tag{3.220}$$

*Note that in the expectations above $w_0$ is assumed fixed. Furthermore, we can assume $\mathbb{P}(\mathfrak{E}_T) \geq 1 - \tilde{O}(\eta^5)$.*

This Lemma is a tuned version of Lemma 16 in Ge et al. [2015a]. We change a couple of things here. First, we consider an implicit coupling that if $w_t$ exits $\mathcal{D}_l$

we do not move it anymore, i.e. $w_{t'} = w_t$, $\forall t' \geq t$, which means the noise vectors also becomes zero, i.e. $\mathscr{L}_{t'} = 0, \forall t' \geq t$. This way, the sequence of noise vectors remain bounded by $Q$, because if $w_t$ is inside $\mathcal{D}_l$, then by assumption $\|\mathscr{L}_t\| \leq Q$, while otherwise $\mathscr{L}_t = 0$. We denote the event that the sequence $w_0, \ldots w_T$ remain in $D_l$ by $\mathcal{E}_T$, where $T$ is defined in Lemma 16 of Ge et al. [2015a].

Note that we also have the smoothing parameters $\rho_1, \rho_2, \rho_3$ for all $(w_t)$ because of this coupling. In fact, we will use a more strict coupling; we consider the event $\mathfrak{E}_T$ to be the high probability event corresponding to the bounds in Equations (44) and (36) of Ge et al. [2015a] holding for all $t \leq T$; We will see that $\mathfrak{E}_T \subseteq \mathcal{E}_T$ at the end of this proof, but for now we assume it is true. An important point to note here is that in Ge et al. [2015a], $\mathbb{P}(\mathfrak{E}_T)$ is bounded by $O(\eta^2)$. However, the exponent dependency of $\eta$ in this bound comes from Azuma-Hoeffding type inequalities, particularly used in Equations (60) and (42) in Ge et al. [2015a], in which by considering larger constants one can easily get higher exponents. For our analysis, a bit stronger dependence of $\eta^5$ is required.

Also, because the distribution of our noise depends on the point $w$, our sequence of noise vectors $(\mathscr{L}_t)$ is a martingale instead of being i.i.d, so we apply Azuma-Hoeffding inequality instead of the simple Hoeffdings in Lemma 16 of Ge et al. [2015a]. (because we are also sampling a random $(x_i, y_i)$ to compute the estimate of the gradient, this could be simplified to the case where we compute the actual gradient and then inject an i.i.d noise vector in each step, but it is an overhead to compute the actual gradient, so here we choose to analyze the more complicated case.)

Next, notice the definition of $\Lambda$ and $\tilde{\Lambda}$ right after Equation (66) in Ge et al. [2015a], which in our notation translates to

$$\tilde{\Lambda} := \nabla L^{\Pi}(w_0)^T \tilde{\delta} + \frac{1}{2}\tilde{\delta}^T \mathcal{H} \tilde{\delta},\ \Lambda = \nabla L^{\Pi}(w_0)^T \delta + \frac{1}{2}\delta^T \mathcal{H} \delta + \tilde{\delta}^T \mathcal{H} \delta + \frac{\rho_3}{6}\|\tilde{\delta} + \delta\|^3. \tag{3.221}$$

where

$$\tilde{\delta} = \tilde{w}_T - w_0,\ \delta = w_T - \tilde{w}_T,$$

for $(\tilde{w_t})$ which is a coupled sequence with $(w_t)$ as defined in Ge et al. [2015a]. Note that we apply the coupling for the sequence $\tilde{w_t}$ as well, i.e. if $w_{t+1} = w_t$, we also set $\tilde{w}_{t+1} = \tilde{w}_t$.

To show Equation (3.219), we want to use Equation (67) in Ge et al. [2015a], though we only use the expansion for the first term which is under $\mathbb{1}\{\mathcal{E}_T\}$, i.e.

$$\mathbb{E}[L^{\Pi}(w_T) - L^{\Pi}(w_0)]\mathbb{1}_{\mathfrak{E}_T} = \mathbb{E}\tilde{\Lambda}\mathbb{1}_{\mathfrak{E}_T} + \mathbb{E}\Lambda\mathbb{1}_{\mathfrak{E}_T}. \tag{3.222}$$

First of all, as it is mentioned in Lemma 16 of [Ge et al., 2015a], in the case where the noise vector ${\sigma_1}^2 I \le \mathbb{E}\pounds\pounds^T \le {\sigma_2}^2 I$ instead of having $\mathbb{E}\pounds\pounds^T = \sigma^2 I$ for a fixed $\sigma$, in order to still get a negative term of order $\eta$ in Equation (68) of [Ge et al., 2015a], we just need the size of $T_{\max}$ to be as large as $O(\frac{1}{\gamma\eta}(\log d + \log\frac{\sigma_2}{\sigma_1}))$, and it does not change the order of $\eta$ in any other part of Lemma 16. Now similar to Equation (68) of [Ge et al., 2015a], if w.l.o.g we assume the smallest eigenvalue $\gamma_0$ corresponds to $i = 1$:

$$\mathbb{E}\tilde{\Lambda}\mathbb{1}_{\mathfrak{E}_T} \le \frac{1}{2}\sum_{i=1}^{N}\lambda_i \sum_{\tau=0}^{T-1}\mathbb{1}_{\{\lambda_i<0\}}(1-\eta\lambda_i)^{2\tau}\eta^2\sigma_1^2\mathbb{P}(\mathfrak{E}_T) + \tag{3.223}$$

$$\frac{1}{2}\sum_{i=1}^{N}\lambda_i \sum_{\tau=0}^{T-1}\mathbb{1}_{\{\lambda_i\ge 0\}}(1-\eta\lambda_i)^{2\tau}\eta^2\sigma_2^2 \tag{3.224}$$

$$\le \frac{\eta^2}{2}\Big[\sigma_2^2\frac{N-1}{\eta} - \gamma_0\sigma_1^2\mathbb{P}(\mathfrak{E}_T)\sum_{\tau=0}^{T-1}(1+\eta\gamma_0)^{2\tau}\Big] \le -\frac{\eta\sigma_1^2}{2}. \tag{3.225}$$

where in the last line we use the fact that $\mathbb{P}(\mathfrak{E}_T) \le 1/2$ plus the additional $\log(\sigma_2/\sigma_1)$ factor. Second, note that our threshold $\S(\eta)$ for the size of gradient in Lemmas 33 and 35 has the same order of $\eta$ compared to that of Lemmas 14 and 16 in Ge et al. [2015a]. Therefore, the arguments in Lemma 16 that considers the order of $\eta$ and treat the other parameters as constants is true here as well. Hence, we still have Equation (69) of Ge et al. [2015a] which is under the event $\mathcal{E}_T$. Applying it to Equation (3.222), Hence, finally by a similar derivation of Equation (67) in Ge et al. [2015a]:

$$\mathbb{E}[L^{\Pi}(w_T) - L^{\Pi}(w_0)]\mathbb{1}\{\mathfrak{E}_T\} \leq -\tilde{\Omega}(\eta). \tag{3.226}$$

Next, we turn to prove the second bound (3.220). Combining Equations (36) and (44) in [Ge et al., 2015a], we get with high probability (we use the final high probability parameter of Lemma 16 which is the result of a union bound over all the high probability arguments which is equivalent to the occurrence of $\mathfrak{E}_T$), i.e. when $\mathfrak{E}_T$ happens,

$$\|w_T - w_0\| \leq \tilde{O}(\eta^{\frac{1}{2}} \log \frac{1}{\eta}). \tag{3.227}$$

Picking $\eta$ small enough such that for the bound above we have

$$O(\eta^{\frac{1}{2}} \log \frac{1}{\eta}) \leq \sqrt{\frac{\aleph + 3l}{\psi}} - \sqrt{\frac{\aleph + 2l}{\psi}},$$

we get for every $\bar{w}$ in the line connecting $w_0$ to $w_T$ :

$$\|\bar{w}\| \leq \sqrt{\frac{\aleph + 3l}{\psi}},$$

which implies that $L^{\Pi}$ has the smoothing parameters $\rho_1, \rho_2, \rho_3$ along $w_0$ to $w_T$. Therefore, by the $\rho_2$-gradient smoothness property of $L^{\Pi}$:

$$\|\nabla L^{\Pi}(\bar{w}) - \nabla L^{\Pi}(w_0)\| \leq \rho_2 \|w_0 - \bar{w}\| \leq O(\rho_2 \eta^{\frac{1}{2}} \log \frac{1}{\eta}).$$

Combining the assumption of the Lemma $\|\nabla L^{\Pi}(w_0)\| \leq \tilde{O}(\eta^{\frac{1}{2}})$, we get

$$\|\nabla L^{\Pi}(\bar{w})\| = \tilde{O}(\eta^{1/2} \log(1/\eta)).$$

(The last $\tilde{O}$ also hides the dependency on $\rho_2$). Now integrating over the derivative along the direction from $w_0$ to $w_T$:

$$L^{\Pi}(w_T) = L^{\Pi}(w_0) + \int_{t=0}^{1} \nabla L^{\Pi}(tw_0 + (1-t)w_T)^T (w_T - w_0) dt.$$

Therefore, using (3.227) one more time, under the event $\mathcal{E}_T$:

$$\begin{aligned} |L^{\Pi}(w_T) - L^{\Pi}(w_0)| &\leq \int \left| \nabla L^{\Pi}(tw_0 + (1-t)w_T)^T (w_T - w_0) \right| dt \\ &\leq \int_0^1 \|\nabla L^{\Pi}(tw_0 + (1-t)w_T\| \|w_T - w_0\| dt \\ &\leq \tilde{O}(\eta^{1/2} \log 1/\eta) \|w_T - w_0\| = \tilde{O}(\eta \log^2 1/\eta). \end{aligned}$$

Hence

$$\left| [L^{\Pi}(w_T) - L^{\Pi}(w_0)] \mathbb{1}\{\mathfrak{E}_T\} \right| \leq \tilde{O}(\eta \log^2 1/\eta), \tag{3.228}$$

which

Now comparing Equations (3.226) and (3.228), it is clear that one can pick $\eta$ small enough (again polynomially small in the other parameters) such that for some random variable $\mathfrak{R}_2$, which also depends on $\eta$, so that equations (3.219) and (3.220) hold.

It remains to show $\mathcal{E}_T \subseteq \mathfrak{E}_T$. This is desirable as up until now we have only proved (3.219) and (3.220) for the coupled sequence (which does not move outside the ball $D_l$), but we know that under the event $\mathcal{E}_T$, the coupled sequence and the original sequence are the same, which automatically implies the conclusion for the original sequence. Notice that the bound in (3.228) is an a.s. upper bound on the change of the function value under the event $\mathcal{E}_T$ for every $1 \leq t \leq T$. Therefore, by picking $\eta$ small enough (polynomially) s.t. the quantity $O(\eta \log(1/\eta)^2)$ in Equation (3.228) is bounded by $l$, we again make sure that the value of function during these steps changes by at most $l$ compared to $w_0$, i.e. for every $1 \leq t \leq T$:

$$\left| [L^{\Pi}(w_t) - L^{\Pi}(w_0)] \mathbb{1}\{\mathfrak{E}_t\} \right| \leq l, \tag{3.229}$$

hence, remains bounded by $\aleph + 3l$. This implies $\mathfrak{E}_T \subseteq \mathcal{E}_T$ as promised.

### 3.6.17 Process from a higher view: definition of the $(X)$ sequence

The goal here is to find a $w^*$ with $L^\Pi(w^*) \leq \aleph_\ell$ using Lemmas 33 and 35 (recall the definition of $\aleph_\ell$ from Theorem 7). The main result of this section is Lemma 36. For this purpose we define a useful coupling: to begin, as done in Ge et al. [2015a], define a sequence of times $\tau_i$ inductively in the following way: To define $\tau_{i+1}$ based on $\tau_i$, if the condition

$$\aleph_\ell \leq L^\Pi(w_{\tau_{i+1}}) \leq \aleph + 2l \tag{3.230}$$

does not hold, then just set $\tau_{i+1} = \tau_i \star(1)$. Otherwise, using the conditions (3.209), we are either in the situation of Lemma 33 or Lemma 35 by setting the value of $w_0$ in these Lemmas as $w_0 = w_{\tau_i}$. In the first case, define $\tau_{i+1} = \tau_i + 1 \star(2)$. In the latter case, Let $\mathfrak{E}_T$ be the same high probability event that we consider in Lemma (35), which happens when the aggregate behavior of the noise vectors is normal, as a result of which $w$ remains close to the starting point $w_0$. Note that from Lemma 35, we know $\mathbb{P}(\mathfrak{E}_T) \geq 1 - O(\eta^5)$. Now if the event $\mathfrak{E}_T$ happens, define $\tau_{i+1} \coloneqq \tau_i + T \star(3)$, for $T$ also from Lemma 35 and defined originally in Lemma 16 of Ge et al. [2015a], while otherwise, define $\tau_{i+1} = \tau_i \star(4)$. Moreover, if $\mathfrak{E}_T$ does not happen, define the rest of $\tau_{i'}$'s equal to $\tau_i$: $\tau_{i'} = \tau_i$ for every $i' \geq i$. At the same time, we define the monotone increasing events $\{\mathcal{G}_i\}$, where $\mathcal{G}_i$ happens in the case $\star(4)$, and $\mathcal{G}_{i+1}$ happens in case $\star(4)$. Also, $\mathcal{G}_i$ happens if any of the previous $\mathcal{G}_{i'}$'s happen for $i' < i$; in other words, $\mathcal{G}_i$ is included in $\mathcal{G}_{i+1}$. We use these events to bound the probability that the process remains above $\aleph_\ell$. Moreover, define the sequence of random variables $(X_i)$ as $X_i \coloneqq L^\Pi(w_{\tau_i})$. Note that by Lemma 34 and Equation (3.229) in Lemma 35, we have $\aleph_\ell - l \leq X_i \leq \aleph + 3l$. The key idea behind defining $X_i$'s is that we want to bound the MGF of $L^\Pi(w_t)$, without worrying about falling out of the assumptions of Lemmas 33 and 35. With the definition of $(X_i)$ and $\mathcal{G}_i$, we are ready to state the theorem which roughly says the sequence $\tau_i$ will most likely stop after a number of steps.

**Lemma 36.** *Let $\mathcal{Q}_R \coloneqq \bigcup_{i=1}^{\infty} \Big( \{ \ \tau_i \leq R\} \cap \bar{\mathcal{G}}_i \Big)$. Then, for some*

$$R = \frac{O(\log(1/\delta_1))(\aleph + 3l)}{\theta\eta^2}, \tag{3.231}$$

*we have $\mathbb{P}(\mathcal{Q}_R) \leq \delta_1$. In other words, after $R$ iterations of PSGD, the defined sequence $(X_i)$ above has either been in situation $\star(1)$ or $\star(4)$. Here, $\theta$ depends polynomially in the other parameters.*

**Proof of Lemma 36**

By Equations (3.210) and (3.226) in Lemmas 33 and 35, there exist a constant $\theta$ depending polynomially on all parameters except $\eta$ such that

$$\mathbb{E}[X_{i+1} - X_i | \ \bar{\mathcal{G}}_i] \leq -\theta(\tau_{i+1} - \tau_i)\eta^2. \tag{3.232}$$

Now for some constant $C$ that we specify later, define the random time $\imath$ as the largest $i$ where $\tau_i \leq C/\eta^2$. Using the fact that $\mathcal{G}_{i-1} \subseteq \mathcal{G}_i$, for every $i$ we have a.s.:

$$X_{i+1}\mathbb{1}\{\bar{\mathcal{G}}_i\} - X_i\mathbb{1}\{\bar{\mathcal{G}}_{i-1}\} = \mathbb{1}\{\mathcal{G}_i - \mathcal{G}_{i-1}\}(-X_i) + (X_{i+1} - X_i)\mathbb{1}\{\bar{\mathcal{G}}_i\}.$$

Now summing this for $i = 1$ to $\imath$, taking expectation from both sides and using (3.232):

$$\begin{aligned}
\mathbb{E}X_{\imath+1}\mathbb{1}\{\bar{\mathcal{G}}_\imath\} - X_0 &= \sum_{i=1}^{\infty} \mathbb{E}\mathbb{1}\{\mathcal{G}_i - \mathcal{G}_{i-1}\}(-X_i)\mathbb{1}\{\imath \geq i\} \\
&+ \sum_{i=1}^{\infty}(X_{i+1} - X_i)\mathbb{1}\{\bar{\mathcal{G}}_i \cap \{\imath \geq i\}\} \\
&\leq \sum_{i=1}^{\infty} \mathbb{E}(\mathbb{1}\{\mathcal{G}_i\} - \mathbb{1}\{\mathcal{G}_{i-1}\})(-X_i) \\
&+ \sum_{i=1}^{\infty} \mathbb{E}(X_{i+1} - X_i \;\Big|\; \bar{\mathcal{G}}_i \cap \{\imath \geq i\})\mathbb{P}(\bar{\mathcal{G}}_i \cap \{\imath \geq i\}) \\
&\leq \sup_i \sup |X_i| \\
&+ \theta \sum_{i=1}^{\infty} \mathbb{E}(-(\tau_{i+1} - \tau_i)\eta^2 \;\Big|\; \bar{\mathcal{G}}_i \cap \{\imath \geq i\})\mathbb{P}(\bar{\mathcal{G}}_i \cap \{\imath \geq i\}) \\
&= \sup_i \sup |X_i| - \eta^2\theta \sum_{i=1}^{\infty} \mathbb{E}(\tau_{i+1} - \tau_i)\mathbb{1}\{\bar{\mathcal{G}}_i \cap \{\imath \geq i\}\}.
\end{aligned}$$

Now using Lemma 34, we know that in except when $\mathfrak{E}_T$ happens (in which we stop the time sequence $\tau_i$), the increments of $X_i$ are at most $l$. Therefore, the value of $X_i$'s always remain bounded by $\aleph + 3l$, hence:

$$LHS \leq \aleph + 3l - \theta\eta^2 \sum_{i=1}^{\infty} \mathbb{E}(\tau_{i+1} - \tau_i)\mathbb{1}\{\bar{\mathcal{G}}_i \cap \{\imath \geq i\}\}.$$

Also, by restricting the integration of the second term to the part $\bigcup_{i=1}^{\infty}\Big(\bar{\mathcal{G}}_i \cap \{\tau_i \geq 2C/\eta^2\}\Big)$ of the sample space, we know that under the event $\{\imath \geq i\}$, $\bar{\mathcal{G}}_i$ automatically happens when $\tau_{i+1} \neq \tau_i$ (it is easy to check). Therefore:

$$\begin{aligned}
LHS &\leq \aleph + 3l - \theta\eta^2 \mathbb{E}\mathbb{1}\{\bigcup_{i=1}^{\infty}\Big(\bar{\mathcal{G}}_i \cap \{\tau_i \geq C/\eta^2\}\Big)\} \sum_{i=1}^{\infty}(\tau_{i+1} - \tau_i)\mathbb{1}\{\bar{\mathcal{G}}_i \cap \{\imath \geq i\}\} \\
&= \aleph + 3l - \theta\eta^2 \mathbb{E}\mathbb{1}\{\bigcup_{i=1}^{\infty}\Big(\bar{\mathcal{G}}_i \cap \{\tau_i \geq C/\eta^2\}\Big)\} \sum_{i=1}^{\infty}(\tau_{i+1} - \tau_i)\mathbb{1}\{\imath \geq i\} \\
&= \aleph + 3l - \theta\eta^2 \mathbb{E}\mathbb{1}\{\bigcup_{i=1}^{\infty}\Big(\bar{\mathcal{G}}_i \cap \{\tau_i \geq C/\eta^2\}\Big)\} \sum_{i=1}^{\imath}(\tau_{i+1} - \tau_i) \\
&= \aleph + 3l - \theta\eta^2 \mathbb{E}\mathbb{1}\{\bigcup_{i=1}^{\infty}\Big(\bar{\mathcal{G}}_i \cap \{\tau_i \geq C/\eta^2\}\Big)\}\tau_{\imath+1}.
\end{aligned}$$

Now by the definition of $\imath$, $\tau_{\imath+1} \geq C/\eta^2$. Hence, we can write

$$\begin{aligned}
LHS \leq \aleph + 3l - \theta\eta^2 \mathbb{E}\mathbb{1}\{\bigcup_{i=1}^{\infty}\Big(\bar{\mathcal{G}}_i \cap \{\tau_i \geq C/\eta^2\}\Big)\}(C/\eta^2) \\
\aleph + 3l - C\theta\mathbb{P}\Big(\bigcup_{i=1}^{\infty}\Big(\bar{\mathcal{G}}_i \cap \{\tau_i \geq C/\eta^2\}\Big)\Big).
\end{aligned}$$

But note that $X_i$'s are a.s. bounded between 0 and $\aleph + 3l$, which implies the LHS above is at least $-(\aleph + 3l)$. Therefore, we finally get:

$$\mathbb{P}(\bigcup_{i=1}^{\infty}\Big(\bar{\mathcal{G}}_i \cap \{\tau_i \geq C/\eta^2\}\Big)) \leq \frac{2\aleph + 6l}{C\theta}.$$

Picking $C = C^* := 2(2\aleph + 6l)/\theta$:

$$\mathbb{P}(\bigcup_{i=1}^{\infty}\Big(\bar{\mathcal{G}}_i \cap \{\tau_i \geq C^*/\eta^2\}\Big)) \leq \frac{1}{2}. \tag{3.233}$$

Note that the differences between $\tau_i$'s is at most $T_{\max} = \tilde{O}(1/\eta)$ Ge et al. [2015a]. Hence, again for $\eta$ polynomially small in other parameters, (3.233) implies that for $R = 2C^*/\eta^2$, there exists $\tilde{R} = poly(.)$ such that after $\tilde{R}$ iterations on the main sequence $(w_t)$, the corresponding sequence $(\tau_i)$ has either been in $\star(1)$ or $\star(4)$ with chance at least 1/2. Repeating this argument $\log(1/\delta_1)$ times (using the markov property of the process) we conclude the proof.

### 3.6.18 Bounding the MGF of $X_i$'s

Next, we want to exploit $X_i$'s to bound the upward deviation of $L^{\Pi}(w_t)$. For a fix $\theta$ the goal here is to bound $\mathbb{E}[\exp\{\theta X_i\}]$ (this is a different $\theta$!). More precisely, let $F_t$ be the sub-sigma field generated by variables $w_t$ from time zero to $t$, and $\mathcal{F}_i := F_{\tau_i}$ be the sigma field of the stop time $\tau_i$. Then, obviously, $X_i$ is measurable w.r.t $\mathcal{F}_i$. We prove the following theorem:

**Theorem 8.** *For any $\theta > 0$, the sequence $(\mathbb{E}e^{\theta(X_i - X_0)})_{i=1}^{\infty}$ is a supermartingale with respect to the filteration $(\mathcal{F}_i)$,*

We proceed inductively by jointly conditioning on the previous $X_i$ and whether $\mathcal{G}_i$ has happened or not, and whether we are in situation $\star(2)$ or $\star(3)$. We have

$$\begin{aligned}
&\mathbb{E}[\exp\{\theta(X_{i+1} - X_0)\}|\ \mathcal{F}_i] \\
&= \mathbb{E}[\exp\{\theta(X_{i+1} - X_i + X_i - X_0)\}\mathbb{1}\{\mathcal{G}_i\}|\mathcal{F}_i] \\
&+ \mathbb{E}[\exp\{\theta(X_{i+1} - X_i + X_i - X_0)\}\mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(2)\}|\mathcal{F}_i] \\
&+ \mathbb{E}[\exp\{\theta(X_{i+1} - X_i + X_i - X_0)\}\mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(3)\}|\mathcal{F}_i].
\end{aligned}$$

Now by the a.s. bounds of Lemmas 33 and 35:

$$\mathbb{E}[X_{i+1} - X_i|\ \bar{\mathcal{G}}_i,\ w_{\tau_i} s.t. \star(2)] = -\mathfrak{R}_1^2\eta^2,$$

$$\mathbb{E}[X_{i+1} - X_i|\ \bar{\mathcal{G}}_i,\ w_{\tau_i} s.t. \star(3)] = -\mathfrak{R}_2^2\eta^2,$$

$$(X_{i+1} - X_i)\mathbb{1}\{\mathcal{G}_i\} = 0. \text{ (a.s.)}$$

Where $\mathfrak{R}_1$ and $\mathfrak{R}_2$ are r.v. defined in Lemmas 33 and 35 and are clearly $\mathcal{F}_i$ measurable. This implies

$$\mathbb{E}[(X_{i+1} - X_i)\mathbb{1}\{\bar{\mathcal{G}}_i,\ w_{\tau_i} s.t. \star(2)\}|\ \mathcal{F}_i] = -\mathfrak{R}_1^2\eta^2\mathbb{1}\{\bar{\mathcal{G}}_i,\ w_{\tau_i} s.t. \star(2)\},$$

$$\mathbb{E}[(X_{i+1} - X_i)\mathbb{1}\{\bar{\mathcal{G}}_i,\ w_{\tau_i} s.t. \star (3)\}|\ \mathcal{F}_i] = -\mathfrak{R}_2^2 \eta^2 \mathbb{1}\{\bar{\mathcal{G}}_i,\ w_{\tau_i} s.t. \star (3)\}.$$

Now we mention the following fact:

**Fact** For a $\sigma$ subGaussian random variable $X$ we have $\mathbb{E}[\exp\{\theta X\}] \leq \exp\{\theta^2\sigma^2\}$.

Using the a.s. bounds of Lemmas 33 and 35, we get that conditioned on $\{\bar{\mathcal{G}}_i,\ w_{\tau_i}\ s.t. \star (2)\}$, $X_{i+1} - \mathbb{E}X_{i+1}$ is a.s. bounded by $2\eta\theta\mathfrak{R}_1/(\rho_2\chi)$, and conditioned on $\{\bar{\mathcal{G}}_i,\ w_{\tau_i}\ s.t. \star (3)\}$, $X_{i+1} - \mathbb{E}X_{i+1}$ is bounded by $2\eta\theta\mathfrak{R}_2/(\rho_2\chi)$. Therefore, using the above fact

$$\mathbb{E}\Big[\exp\{\theta(X_{i+1} - \mathbb{E}(X_{i+1}|\ \bar{\mathcal{G}}_i,\ w_{\tau_i}\ s.t. \star (2)))\}\Big|\ \bar{\mathcal{G}}_i,\ w_{\tau_i}\ s.t. \star (2)\Big] \leq \exp\{4\eta^2\theta^2\mathfrak{R}_1^2/(\rho_2\chi)\}, \tag{3.234}$$

$$\mathbb{E}\Big[\exp\{\theta(X_{i+1} - \mathbb{E}(X_{i+1}|\ \bar{\mathcal{G}}_i,\ w_{\tau_i}\ s.t. \star (3)))\Big|\ \bar{\mathcal{G}}_i,\ w_{\tau_i}\ s.t. \star (3)\}\Big] \leq \exp\{4\eta^2\theta^2\mathfrak{R}_2^2/(\rho_2\chi)\}, \tag{3.235}$$

which implies in the notation of conditional expectation on sigma field:

$$\mathbb{E}\Big[\exp\{\theta(X_{i+1} - \mathbb{E}(X_{i+1}|\mathcal{F}_i))\}\mathbb{1}\{\bar{\mathcal{G}}_i,\ \star(2)\}\Big|\mathcal{F}_i\Big] \leq \exp\{4\eta^2\theta^2\mathfrak{R}_1^2/(\rho_2\chi)\}\mathbb{1}\{\bar{\mathcal{G}}_i,\ \star(2)\}, \tag{3.236}$$

$$\mathbb{E}\Big[\exp\{\theta(X_{i+1} - \mathbb{E}(X_{i+1}|\mathcal{F}_i))\}\mathbb{1}\{\bar{\mathcal{G}}_i,\ \star(3)\}\Big|\mathcal{F}_i\Big] \leq \exp\{4\eta^2\theta^2\mathfrak{R}_2^2/(\rho_2\chi)\}\mathbb{1}\{\bar{\mathcal{G}}_i,\ \star(3)\}. \tag{3.237}$$

Now we write:

$$
\begin{aligned}
&LHS \leq \mathbb{E}[\exp\{\theta(X_{i+1} - X_0)\}\mathbb{1}\{\mathcal{G}_i\}|\ \mathcal{F}_i] \\
&+\mathbb{E}[\exp\{\theta(X_{i+1} - \mathbb{E}[X_{i+1}|\ \mathcal{F}_i])\}\exp\{\theta(\mathbb{E}[X_{i+1}|\ \mathcal{F}_i] - X_i)\}\exp\{\theta(X_i - X_0)\}\mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(2)\}|\ \mathcal{F}_i] \\
&+\mathbb{E}[\exp\{\theta(X_{i+1} - \mathbb{E}[X_{i+1}|\ \mathcal{F}_i])\}\exp\{\theta(\mathbb{E}[X_{i+1}|\ \mathcal{F}_i] - X_i)\}\exp\{\theta(X_i - X_0)\}\mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(3)\}|\mathcal{F}_i] \\
&\qquad \leq \exp\{\theta(X_i - X_0)\}\mathbb{1}\{\mathcal{G}_i\} \\
&\qquad + \exp\{\theta(X_i - X_0)\}\mathbb{E}[\exp\{\theta^2\mathfrak{R}_1^2\eta^2/(\rho_2\chi)\}\exp\{-\theta(\mathfrak{R}_1^2\eta^2)\}\mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(2)\}|\ \mathcal{F}_i] \\
&\qquad + \exp\{\theta(X_i - X_0)\}\mathbb{E}[\exp\{\theta^2\mathfrak{R}_2^2\eta^2/(\rho_2\chi)\}\exp\{-\theta(\mathfrak{R}_2^2\eta^2)\}\mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(3)\}|\ \mathcal{F}_i] \\
&\qquad \leq \exp\{\theta(X_i - X_0)\}\mathbb{E}\Big[\Big(\mathbb{1}\{\mathcal{G}_i\} \\
&\qquad + \exp\{\theta^2\mathfrak{R}_1^2\eta^2/(\rho_2\chi) - \theta(\mathfrak{R}_1^2\eta^2)\}\mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(2)\} \\
&\qquad + \exp\{\theta^2\mathfrak{R}_2^2\eta^2/(\rho_2\chi) - \theta(\mathfrak{R}_2^2\eta^2)\}\mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(3)\}\Big)|\ \mathcal{F}_i\Big].
\end{aligned}
$$

Now setting $\theta \coloneqq 1$ and picking $\chi \geq 1/\rho_2$:

$$
\begin{aligned}
LHS &\leq \exp\{(X_i - X_0)\}\mathbb{E}\Big[\mathbb{1}\{\mathcal{G}_i\} + \mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(2)\} + \mathbb{1}\{\bar{\mathcal{G}}_i \cap \star(3)\}|\ \mathcal{F}_i\Big]. \\
&= \exp\{X_i - X_0\}.
\end{aligned}
$$

Now by hypothesis of Induction we have

$$
\mathbb{E}[\exp\{X_{i+1} - X_0\}] = \mathbb{E}[\mathbb{E}[\exp\{X_{i+1} - X_0\}|\ \mathcal{F}_i]] \leq \mathbb{E}[\exp\{X_i - X_0\}] \leq 1,
$$

which finishes the proof of step of induction.

Now using Doob's Maximal inequality for positive supermartingales and $R$ defined in (3.231):

$$
\begin{aligned}
&\mathbb{P}(\sup_{1\leq i\leq R}(X_i - X_0) \geq z) \\
&= \mathbb{P}(\sup_{1\leq i\leq R}\exp\{X_i - X_0\} \geq \exp\{z\}) \leq \mathbb{E}[\exp\{X_R - X_0\}]/\exp\{z\} \leq e^{-z}. \quad (3.238)
\end{aligned}
$$

### 3.6.19 Proof of Theorem 7

Finally with the developed tools, we are ready to prove Theorem 7.

*of Theorem 7* Starting from $w_0 = 0$ with $L^{\Pi}(w_0) \le \aleph$, we use Equation (3.238) to get $\mathbb{P}(\sup_{1\le i\le R} \exp\{X_i - X_0\} \ge \Omega(\log(1/\delta_1))) \le \delta_1$. Therefore, setting $l = \Theta(\log(1/\delta_1))$ and a union bound implies with probability at least $1 - 2\delta_1$ we should have gotten into situation $\star(1)$ or $\star(4)$ without the value of $X_i$ exceeding $\aleph + 2l$. On the other hand, using Lemma 35 we know that $\mathfrak{E}_T$ happens with probability at least $1 - \tilde{O}(\eta^5)$ for every $1 \le t \le R$ which is equal to $\tau_i$ for some $i$ and when we are in the situation of Lemma 35. As a result, the chance that even one of $\mathfrak{E}_T$'s happen along $R$ iterations is at most

$$R\tilde{O}(\eta^5) = \eta^3 \frac{O(\log(1/\delta_1))(\aleph + 3l)}{\theta}.$$

But picking $\eta$ small enough with respect to $\log(\delta_1)$ and other parameters, we conclude that with probability at least $1 - 3\delta$, after $R$ rounds, we should have gotten into situation $\star(1)$ and not $\star(4)$ and not exceeding $\aleph + 2l$, which means that $X_i = L^{\Pi}(w_{\tau_i})$ has gotten under the threshold $\aleph_\ell$. Note that as soon as that happens, we terminate the algorithm. We elaborate on this more in Section 3.6.10.

### 3.6.20 Gaussian Smoothing

In this section, we describe our smoothing scheme and the approximation that it provides which enables us to keep the signs from the case $\eta = 0$. Recall that we use Gaussian smoothing matrices $V^{\rho}_{j,k} \sim \mathcal{N}(0, \beta_1^2/m_1)$ and $W^{\rho}_{j,k} \sim \mathcal{N}(0, \beta_2^2/m_2)$. Here, we will particularly specify lower bounds for $\beta_1$ and $\beta_2$ in order for our sign approximation to be precise. On the other hand, we normally prefer the smoothing noise to be as low as possible so the primary and smoothed functions are close, so we set $\beta_1, \beta_2$ equal to their lower bounds, and use this setting in the other parts.

To begin fix one of the inputs $x_i$. In order to reduce and simplify the amount of notations, we refer to the sign pattern matrix (diagonal sign matrix) of both the first and second layers by $D$ with the appropriate indices. More specifically, for the first layer, we refer to $\mathrm{Sgn}(W^{(0)} + W' + W^{\rho})x_i$ by $D_{\prime,\rho}$ and $\mathrm{Sgn}(W^{(0)} + (1 - \eta/2)W' + W^{\rho} + \sqrt{\eta}W^{*})x_i$

by $D_{',\rho}$. Similarly, for the second layer, of course depending on the input vector, we refer to the sign matrix with respect to the matrices $V^{(0)} + V' + V^{\rho}$ and $V^{(0)} + (1-\eta/2)V' + V^{\rho} + \sqrt{\eta}V^*$ by by $D_{'.\rho}$ and $D_{',\rho,\eta}$, respectively. We introduce two new notations as well for the output of the first layer with respect to different matrix and sign patterns:

$$x'^{(1)} := W^s D_{',\rho}(W^{(0)} + (1-\eta)W' + W^{\rho} + \sqrt{\eta}V^*)x_i, \tag{3.239}$$

$$x'^{(2)} := W^s D_{',\rho,\eta}(W^{(0)} + (1-\eta)W' + W^{\rho} + \sqrt{\eta}V^*)x_i. \tag{3.240}$$

For further brevity, we sometimes refer to $x'^{(2)}$ by $x'$.

Now we are ready to mention our approximation theorem regarding the smoothing and the sign changes.

**Lemma 37.** *Under the conditions $\kappa_1\sqrt{m_3} \gtrsim C_1 + \beta_1\sqrt{m_3}$ and $m_2 \geq m_3\log(m_2)$, then for every $i \in [n]$:*

$$\begin{aligned}
&\Big|\mathbb{E}_{W^{\rho},V^{\rho}}\ a^T D_{',\rho,\eta}(V^{(0)} + (1-\eta)V' + V^{\rho} + \sqrt{\eta}V^*)W^s D_{',\rho,\eta}(W^{(0)} + (1-\eta)W' + W^{\rho} + \sqrt{\eta}V^*)x_i \\
&- a^T D_{',\rho}(V^{(0)} + (1-\eta)V' + V^{\rho} + \sqrt{\eta}V^*)W^s D_{',\rho}(W^{(0)} + (1-\eta)W' + W^{\rho} + \sqrt{\eta}V^*)x_i\Big| \\
&\leq \eta\varrho_2^2\beta_2^{-1}\Big[(C_1 + \sqrt{m_3}\beta_1)^2/(\kappa_1\sqrt{m_3}) + \Big[\sqrt{m_3}m_1\beta_1 + C_1\Big]\exp\{-C_1^2/(8m_3\beta_1^2)\}\Big] \\
&\times\Big[\exp\{-C_2^{4/3}(\sqrt{m_2}\kappa_2)^{2/3}/(8\beta_2^2)\} + \frac{C_2^{2/3}}{(\sqrt{m_2}\kappa_2)^{2/3}}\Big] \\
&+ \eta\Big(\kappa_2\sqrt{m_2} + C_1\Big)\Big(\exp\{-c_2^2/(32\beta_1^2)\} + \frac{c_2}{\kappa_1\sqrt{m_1}}\Big)\frac{\varrho^2 m_3\sqrt{m_3}}{\beta_1} := \eta\Re_8.
\end{aligned} \tag{3.241}$$

**Proof of Lemma 37**

We can bound the Left hand side above as

$$\begin{aligned}
&LHS \leq \\
&\Big|\mathbb{E}_{W^\rho,V^\rho} a^T D_{',\rho,\eta}(V^{(0)} + (1-\eta)V' + V^\rho + \sqrt{\eta}V^*)W^s D_{',\rho,\eta}(W^{(0)} + (1-\eta)W' + W^\rho + \sqrt{\eta}W^*)x_i \\
&- a^T D_{',\rho}(V^{(0)} + (1-\eta)V' + V^\rho + \sqrt{\eta}V^*)W^s D_{',\rho,\eta}(W^{(0)} + (1-\eta)W' + W^\rho + \sqrt{\eta}W^*)x_i\Big| \\
&+ \Big|\mathbb{E}_{W^\rho,V^\rho}\ a^T D_{',\rho}(V^{(0)} + (1-\eta)V' + V^\rho + \sqrt{\eta}V^*)W^s D_{',\rho,\eta}(W^{(0)} + (1-\eta)W' + W^\rho + \sqrt{\eta}W^*)x_i \\
&- a^T D_{',\rho}(V^{(0)} + (1-\eta)V' + V^\rho + \sqrt{\eta}V^*)W^s D_{',\rho}(W^{(0)} + (1-\eta)W' + W^\rho + \sqrt{\eta}W^*)x_i\Big|. \\
&\coloneqq A_1 + A_2. \qquad (3.242)
\end{aligned}$$

We bound $A_1$ and $A_2$ separately. First, we start with $A_1$.

Let $\hat{P}_i$ be the set of indices $j$ for which $\mathbb{1}\{|(V_j^{(0)} + V_j')x'| \leq R^*\kappa_2\|x'\|\}$ happens. Then, from Lemma 40, we have $|\hat{P}_i| \lesssim R^* m_2$. Now for $j \in [m_2]$, we write

$$\mathbb{1}\{\text{sign change in the} j\text{th neuron}\} \times |\text{amount of change}| \qquad (3.243)$$

$$\leq \mathbb{1}\{V_j^\rho x' \in (-V_j^{(0)}x' - V_j'x' + \eta V_j'x' - \sqrt{\eta}V_j^*x', -V_j^{(0)}x' - V_j'x')\} \times \frac{1}{\sqrt{m_2}}(|\eta V_j'x'| + |\sqrt{\eta}V_j^*x'|) \qquad (3.244)$$

Moreover, note that

$$|V_j'x'| = |V_j'(x' - \phi^{(0)}(x_i))|,$$

$$|V_j^*x'| = |V_j^*(x' - \phi^{(0)}(x_i))|.$$

Also, because $\|V_j'\| \leq \|V'\| \leq 2C_2$ plus using Equation (3.129), we can further upper bound the above indicator as:

$$\leq \mathbb{1}\{V_j^\rho x' \in (-V_j^{(0)}x' - V_j'x' - (\eta\|V_j'\| + \sqrt{\eta}\|V_j^*\|)\min\{\|x'\|, \|x' - \phi^{(0)}(x_i)\|\}, -V_j^{(0)}x' - V_j'x')\}$$

$$\times\frac{1}{\sqrt{m_2}}(\eta\|V_j'\| + \sqrt{\eta}\|V_j^*\|)\|x' - \phi^{(0)}(x_i)\|$$

$$\leq \mathbb{1}\{V_j^\rho x' \in (-V_j^{(0)}x' - V_j'x' - (2\eta C_2 + \sqrt{\eta}\varrho_2/\sqrt{m_2})\min\{\|x'\|, \|x' - \phi^{(0)}(x_i)\|\}, -V_j^{(0)}x' - V_j'x')\}$$

$$\times\frac{1}{\sqrt{m}_2}(2\eta C_2+\sqrt{\eta}\varrho_2/\sqrt{m}_2)\|x'-\phi^{(0)}(x_i)\|.$$

Taking $\sqrt{\eta}\le\varrho/(2C_2\sqrt{m}_2)$, we can further upper bound as

$$\lesssim\mathbb{1}\{V_j^{\rho}x'\in(-V_j^{(0)}x'-V_j'x'-2\sqrt{\eta}\varrho_2\min\{\|x'-\phi^{(0)}(x_i)\|,\|x'\|\}/\sqrt{m}_2,-V_j^{(0)}x'-V_j'x')\}$$

$$\times(\sqrt{\eta}\varrho_2/m_2)\|x'-\phi^{(0)}(x_i)\|.$$

Therefore, conditioned on $x'$:

$$\mathbb{E}_{V^{\rho}}[\mathbb{1}\{\text{sign change in the}j\text{th neuron}\}\times|\text{amount of change}|\mid x']\le$$

$$\mathbb{P}(V_j^{\rho}x'\in(-V_j^{(0)}x'-V_j'x'-2\sqrt{\eta}\varrho_2\min\{\|x'-\phi^{(0)}(x_i)\|,\|x'\|\}/\sqrt{m}_2,-V_j^{(0)}x'-V_j'x'))$$

$$\times(\sqrt{\eta}\varrho_2/m_2)\|x'-\phi^{(0)}(x_i)\|.$$

Now notice that for $j\notin\hat{P}_i$, we have

$$|-V_j^{(0)}x'-V_j'x'|\ge R^*\kappa_2\|x'\|.$$

Also, note that the variable $V_j^{\rho}x'$ is gaussian with variance $\|x'\|\beta_2/\sqrt{m}_2$ Therefore, conditioned on $x'$, for $j\notin\hat{P}_i$, we have (note that $x'$ does not depend on the randomness of $V^{\rho}$):

$$\mathbb{P}(V_j^{\rho}x'\in(-V_j^{(0)}x'-V_j'x'-2\sqrt{\eta}\varrho_2\min\{\|x'-\phi^{(0)}(x_i)\|,\|x'\|\}/\sqrt{m}_2,-V_j^{(0)}x'-V_j'x'))$$

$$\lesssim\exp\{\min\{|-V_j^{(0)}x'-V_j'x'-2\sqrt{\eta}\varrho_2\|x'\|/\sqrt{m}_2|,|-V_j^{(0)}x'-V_j'x'|\}/(\sqrt{2}\|x'\|\beta_2/\sqrt{m}_2)\}^2$$

$$\times(\|x'\|\beta_2/\sqrt{m}_2)^{-1}\times(\sqrt{\eta}\varrho_2\min\{\|x'-\phi^{(0)}(x_i)\|,\|x'\|\}/\sqrt{m}_2).$$

This equation follows from the fact that

$$\mathbb{P}(a\le\mathcal{N}\le b)\lesssim\frac{|a-b|}{\sigma}e^{\min^2\{a,b\}/\sigma^2}.\tag{3.245}$$

On the other hand, note that for $\sqrt{\eta} \lesssim \frac{C_2^{2/3}(\sqrt{m}_2\kappa_2)^{1/3}}{\varrho_2}$ we have:

$$R^*\kappa_2\|x'\|/2 = \frac{C_2^{2/3}(\sqrt{m}_2\kappa_2)^{1/3}}{2\sqrt{m}_2}\|x'\| \geq 2\sqrt{\eta}\varrho_2\|x'\|/\sqrt{m}_2$$

which implies

$$\begin{aligned}
&\lesssim \exp\{R^*\kappa_2\|x'\|/(2\sqrt{2}\|x'\|\beta_2/\sqrt{m}_2)\}^2(\sqrt{\eta}\varrho_2/\beta_2)\\
&\leq \exp\{-C_2^{4/3}(\sqrt{m}_2\kappa_2)^{2/3}/(8\beta_2^2)\}(\sqrt{\eta}\varrho_2/\beta_2).
\end{aligned}$$

On the other side, for $j \in \hat{P}_i$, we can write

$$\begin{aligned}
&\mathbb{P}(V_j^\rho x' \in (-V_j^{(0)}x' - V_j'x' - 2\sqrt{\eta}\varrho_2 \min\{\|x' - \phi^{(0)}(x_i)\|, \|x'\|\}/\sqrt{m}_2\ ,\ -V_j^{(0)}x' - V_j'x'))\\
&\lesssim (\|x'\|\beta_2/\sqrt{m}_2)^{-1}(\sqrt{\eta}\varrho_2 \min\{\|x' - \phi^{(0)}(x_i)\|, \|x'\|\}/\sqrt{m}_2) = \min\{\|x' - \phi^{(0)}(x_i)\|/\|x'\|, 1\}\sqrt{\eta}\varrho_2/\beta_2.
\end{aligned}$$

Therefore, overall using the fact that $\|V_j^*\| \leq \varrho_2/\sqrt{m}_2$, we can write

$$\begin{aligned}
A_1 &\lesssim \sum_{j\notin\hat{P}} \exp\{-C_2^{4/3}(\sqrt{m}_2\kappa_2)^{2/3}/(8\beta_2^2)\}(\sqrt{\eta}\varrho_2/\beta_2) \min\{\|x' - \phi^{(0)}(x_i)\|^2/\|x'\|, \|x' - \phi^{(0)}(x_i)\|\}\\
&+ \sum_{j\in\hat{P}}(\sqrt{\eta}\varrho_2/\beta_2) \min\{\|x' - \phi^{(0)}(x_i)\|/\|x'\|, 1\} \times (\sqrt{\eta}\varrho_2/m_2)\|x' - \phi^{(0)}(x_i)\|\\
&\lesssim \Big[m_2 \times \exp\{-C_2^{4/3}(\sqrt{m}_2\kappa_2)^{2/3}/(8\beta_2^2)\}(\sqrt{\eta}\varrho_2/\beta_2) \times (\sqrt{\eta}\varrho_2/m_2)\\
&+ \eta\varrho_2^2\beta_2^{-1}\frac{|\hat{P}_i|}{m_2}\Big] \min\{\|x' - \phi^{(0)}(x_i)\|^2/\|x'\|, \|x' - \phi^{(0)}(x_i)\|\}\\
&\leq \eta\varrho_2^2\beta_2^{-1} \min\{\|x' - \phi^{(0)}(x_i)\|^2/\|x'\|, \|x' - \phi^{(0)}(x_i)\|\}\\
&\times \Big[\exp\{-C_2^{4/3}(\sqrt{m}_2\kappa_2)^{2/3}/(8\beta_2^2)\} + \frac{C_2^{2/3}}{(\sqrt{m}_2\kappa_2)^{2/3}}\Big]. \qquad (3.246)
\end{aligned}$$

Next, we bound $A_2$. First we bound $\mathbb{E}_{W^\rho}\|x'^{(1)} - x'^{(2)}\|$. Recalling the setting $c_2 = 2\sqrt{nm_3}C_1/\sqrt{\lambda_0}$ and the definition of in $P$ from Lemma 10, we obtain that for

$j \notin P$, we have for all $i \in [n]$:

$$|W_j^{(0)} x_i| \geq c_2/\sqrt{m}_1,$$
$$|W_j' x_i| \leq c_2/(2\sqrt{m}_1),$$

which means for $j \notin P$:

$$|(W_j^{(0)} + W_j')x_i| \geq c_2/(2\sqrt{m}_1). \tag{3.247}$$

Also, we have

$$|P| \leq c_2\sqrt{m}_1/\kappa_1. \tag{3.248}$$

Now using Equation (3.105) in Lemma 14, we can write for every $i \in [n]$:

$$\begin{aligned}
&val_j := \mathbb{1}\{\text{sign change in the} j\text{th neuron}\} \times \Big|\text{amount of change}\Big| \\
&\leq \mathbb{1}\{W_j^\rho x_i \in (-W_j^{(0)} x - W_j' x_i + \eta W_j' x_i - \sqrt{\eta} W_j^* x_i, -W_j^{(0)} x_i - W_j' x_i)\} \\
&\times \frac{1}{\sqrt{m}_1}(|\sqrt{\eta} W_j^* x_i + \eta W_j' x_i|).
\end{aligned}$$

Using the fact that $\|W_j'\| \leq \|W'\|_F \leq C_1$, and Equation (3.105) ($\|W_j^*\| \leq \varrho\sqrt{\frac{m_3}{m_1}}$) and picking $\sqrt{\eta} \leq \frac{\varrho\sqrt{m}_3}{C_1\sqrt{m}_1}$, we obtain

$$\begin{aligned}
&\leq \mathbb{1}\{W_j^\rho x_i \in (-W_j^{(0)} x_i - W_j' x_i - \eta\|W_j'\| - \sqrt{\eta}\|W_j^*\|, -W_j^{(0)} x_i - W_j' x_i)\} \\
&\times \frac{1}{\sqrt{m}_1}(\sqrt{\eta}\|W_j^*\| + \eta\|W_j'\|) \\
&\leq \mathbb{1}\{W_j^\rho x_i \in (-W_j^{(0)} x_i - W_j' x_i - 2\sqrt{\eta}\varrho\frac{\sqrt{m}_3}{\sqrt{m}_1}, -W_j^{(0)} x_i - W_j' x_i)\} \\
&\times \frac{2}{\sqrt{m}_1}(\sqrt{\eta}\varrho\sqrt{\frac{m_3}{m_1}}).
\end{aligned}$$

Now for $j \notin P$, because $W^{\rho}_j x_i$ is Gaussian with std $\frac{\beta_1}{\sqrt{m_1}}$:

$$\begin{aligned}\mathbb{E}_{W^{\rho}}[val_j] &\leq \mathbb{P}(W^{\rho}_j x_i \in (-W^{(0)}_j x_i - W'_j x_i - 2\sqrt{\eta}\varrho\frac{\sqrt{m_3}}{\sqrt{m_1}}, -W^{(0)}_j x_i - W'_j x_i)) \times \frac{1}{\sqrt{m_1}}\sqrt{\eta}\varrho\frac{\sqrt{m_3}}{\sqrt{m_1}} \\ &\lesssim \exp -\{\min\{|-W^{(0)}_j x_i - W'_j x_i - 2\sqrt{\eta}\varrho\frac{\sqrt{m_3}}{\sqrt{m_1}}|, |-W^{(0)}_j x' - W'_j x'|\}/(\sqrt{2}\beta_1/\sqrt{m_1})\}^2 \\ &\times (\beta_1/\sqrt{m_1})^{-1} \times (\sqrt{\eta}\varrho\frac{\sqrt{m_3}}{\sqrt{m_1}}) \times \frac{1}{\sqrt{m_1}}\sqrt{\eta}\varrho\frac{\sqrt{m_3}}{\sqrt{m_1}}.\end{aligned}$$

Now from Equation (3.247) and by picking $\sqrt{\eta} \lesssim \frac{c_2}{\varrho\sqrt{m_3}}$ so that

$$2\sqrt{\eta}\varrho\frac{\sqrt{m_3}}{\sqrt{m_1}} \leq c_2/(4\sqrt{m_1}),$$

then

$$LHS \lesssim \exp\{-c_2^2/(32\beta_1^2)\}\eta\frac{\varrho^2 m_3}{\beta_1 m_1}. \tag{3.249}$$

On the other hand, for $j \in P$ we have

$$\begin{aligned}\mathbb{E}val_j &\leq \mathbb{P}(W^{\rho}_j x_i \in (-W^{(0)}_j x_i - W'_j x_i - 2\sqrt{\eta}\varrho\frac{\sqrt{m_3}}{\sqrt{m_1}}, -W^{(0)}_j x_i - W'_j x_i)) \times \frac{1}{\sqrt{m_1}}\sqrt{\eta}\varrho\frac{\sqrt{m_3}}{\sqrt{m_1}} \\ &\lesssim (\beta_1/\sqrt{m_1})^{-1}\sqrt{\eta}\varrho\frac{\sqrt{m_3}}{\sqrt{m_1}} \times \sqrt{\eta}\varrho\frac{\sqrt{m_3}}{m_1} = \eta\frac{\varrho^2 m_3}{\beta_1 m_1}. \qquad (3.250)\end{aligned}$$

Now define the following random variable with respect to the randomness of $W^{\rho}$:

$$Val \coloneqq \sum_{j=1}^{m_1} \mathbb{1}\{\text{sign change in the} j\text{th neuron}\} \times |\text{amount of change}|$$

then for every $k \in [m_3]$, we have

$$|x'^{(1)}_k - x'^{(2)}_k| \leq Val,$$

which implies

$$\|x'^{(1)} - x'^{(2)}\| \leq \sqrt{m_3} Val.$$

But Combining Equations (3.249) and  (3.250):

$$\begin{aligned}\mathbb{E}Val &\leq \Big( \exp\{-c_2^2/(32\beta_1^2)\} + \frac{|P|}{m_1}\Big)\eta\frac{\varrho^2 m_3}{\beta_1}\\ &\leq \Big( \exp\{-c_2^2/(32\beta_1^2)\} + \frac{c_2}{\kappa_1\sqrt{m_1}}\Big)\eta\frac{\varrho^2 m_3}{\beta_1},\end{aligned}$$

which implies

$$\mathbb{E}_{W^\rho}\|x'^{(1)} - x'^{(2)}\| \leq \Big( \exp\{-c_2^2/(32\beta_1^2)\} + \frac{c_2}{\kappa_1\sqrt{m_1}}\Big)\eta\frac{\varrho^2 m_3\sqrt{m_3}}{\beta_1}. \tag{3.251}$$

Now we can write

$$\begin{aligned}&\Big|a^T D_{',\rho}(V^{(0)} + (1-\eta)V' + V^\rho + \sqrt{\eta}V^*)x'^{(2)} - a^T D_{',\rho}(V^{(0)} + (1-\eta)V' + V^\rho + \sqrt{\eta}V^*)x'^{(1)}\Big|\\ &\leq \frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}|(V_j^{(0)} + (1-\eta)V_j' + V_j^\rho + \sqrt{\eta}V_j^*)(x'^{(2)} - x'^{(1)})|\\ &\leq \frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}|V_j^{(0)}(x'^{(2)} - x'^{(1)})| + (1-\eta)|V_j'(x'^{(2)} - x'^{(1)})| + |V_j^\rho(x'^{(2)} - x'^{(1)})| + \sqrt{\eta}|V_j^*(x'^{(2)} - x'^{(1)})|\\ &= \frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}|V_j^{(0)}(x'^{(2)} - x'^{(1)})| + |V_j^\rho(x'^{(2)} - x'^{(1)})| + \frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}((1-\eta)\|V_j'\| + \sqrt{\eta}\|V_j^*\|)\|x'^{(2)} - x'^{(1)}\|\\ &\leq \frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}|V_j^{(0)}(x'^{(2)} - x'^{(1)})| + |V_j^\rho(x'^{(2)} - x'^{(1)})| + \Big((1-\eta)\|V'\|_F + \sqrt{\eta}\|V^*\|_F\Big)\|x'^{(2)} - x'^{(1)}\|.\end{aligned}$$

Now by Equation (3.127) in Lemma 19 (i.e. $\|V^*\|_F \lesssim \sqrt{\zeta}_2$) and the fact that $\|V'\|_F \leq C_2$, and by taking

$$\sqrt{\eta} \leq \frac{C_2}{\sqrt{\zeta}_2},$$

we have

$$\Big((1-\eta)\|V'\|_F + \sqrt{\eta}\|V^*\|_F\Big) \lesssim C_1, \tag{3.252}$$

so we can bound the above as

$$LHS \lesssim \frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}\Big(|V_j^{(0)}(x'^{(2)} - x'^{(1)})| + |V_j^\rho(x'^{(2)} - x'^{(1)})|\Big) + C_1\|x'^{(2)} - x'^{(1)}\|.$$

Furthermore, using Lemma 41 and noting the fact that the entries of $V^{(0)}$ are normal with standard deviation $\kappa_2$, we get with high probability over the randomness of $V^{(0)}$:

$$\lesssim \frac{1}{\sqrt{m}_2}\sum_{j=1}^{m_2}|V_j^{\rho}(x'^{(2)}-x'^{(1)})|+\Big(\kappa_2\sqrt{m_2}+\kappa_2\sqrt{m_3(\log(m_3)+\log(\log(m_2)))}+C_1\Big)\|x'^{(2)}-x'^{(1)}\|.$$

Now note that $V_j^{\rho}(x'^{(2)}-x'^{(1)})$ is normal with standard deviation $\frac{\beta_2}{\sqrt{m}_2}\|x'^{(2)}-x'^{(1)}\|$. Hence, taking expectation with respect to $V^{\rho}$:

$$\begin{aligned}\mathbb{E}_{V^{\rho}}\Big|a^T D_{',\rho,\eta}(V^{(0)}+(1-\eta)V'+V^{\rho}+\sqrt{\eta}V^*)x'^{(2)}-a^T D_{',\rho,\eta}(V^{(0)}+(1-\eta)V'+V^{\rho}+\sqrt{\eta}V^*)x'^{(1)}\Big|\\ \lesssim \Big(\kappa_2\sqrt{m_2}+\kappa_2\sqrt{m_3(\log(m_3)+\log(\log(m_2)))}+C_1\Big)\|x'^{(2)}-x'^{(1)}\|.\end{aligned} \tag{3.253}$$

Finally, Combining Equation (3.246) and (3.253) and applying it to Equation (3.242) implies with high probability over the random initialization:

$$\begin{aligned}&\Big|\mathbb{E}_{W^{\rho},V^{\rho}}\ a^T D_{',\rho,\eta}(V^{(0)}+(1-\eta)V'+V^{\rho}+\sqrt{\eta}V^*)W^s x'^{(2)}\\ &-a^T D_{',\rho}(V^{(0)}+(1-\eta)V'+V^{\rho}+\sqrt{\eta}V^*)W^s x'^{(1)}\Big|\\ &\leq A_1+A_2\\ &\lesssim \mathbb{E}_{W^{\rho}}\eta\varrho_2^2\beta_2^{-1}\min\{\|x'-\phi^{(0)}(x_i)\|^2/\|x'\|,\|x'-\phi^{(0)}(x_i)\|\}\\ &\times\Big[\exp\{-C_2^{4/3}(\sqrt{m}_2\kappa_2)^{2/3}/(8\beta_2^2)\}+\frac{C_2^{2/3}}{(\sqrt{m}_2\kappa_2)^{2/3}}\Big]\\ &+\mathbb{E}_{W^{\rho}}\Big(\kappa_2\sqrt{m_2}+\kappa_2\sqrt{m_3(\log(m_3)+\log(\log(m_2)))}+C_1\Big)\|x'^{(2)}-x'^{(1)}\|\end{aligned}$$

Now notice that under $E^c$, using the assumption $\kappa_1\sqrt{m}_3\gtrsim C_1+\sqrt{m}_3\beta_1$ and Lemma 42 we have

$$\begin{aligned}\|x'\|&\geq\|\phi^{(0)}(x_i)\|-\|x'-\phi^{(0)}(x_i)\|\\ &\geq\kappa_1\sqrt{m}_3-(C_1+\sqrt{m}_3\beta_1)\\ &\gtrsim\kappa_1\sqrt{m}_3,\end{aligned}$$

and

$$\|x' - \phi^{(0)}(x_i)\|^2 \leq (C_1 + \sqrt{m}_3\beta_1)^2, \tag{3.254}$$

which implies:

$$\begin{aligned}
&\mathbb{E}_{W^\rho} \min\{\|x' - \phi^{(0)}(x_i)\|^2/\|x'\|, \|x' - \phi^{(0)}(x_i)\|\} \\
&\leq \mathbb{E}_{W^\rho} \mathbb{1}\{E^c\}\|x' - \phi^{(0)}(x_i)\|^2/\|x'\| + \mathbb{1}\{E\}\|x' - \phi^{(0)}(x_i)\| \\
&\lesssim (C_1 + \sqrt{m}_3\beta_1)^2/(\kappa_1\sqrt{m}_3) + \mathbb{E}_{W^\rho}\mathbb{1}\{E\}\|x' - \phi^{(0)}(x_i)\| \\
&\lesssim (C_1 + \sqrt{m}_3\beta_1)^2/(\kappa_1\sqrt{m}_3) + \Big[\sqrt{m}_3 m_1\beta_1 + C_1\Big]\exp\{-C_1^2/(8m_3\beta_1^2)\}.
\end{aligned}$$

Substituting this above and further applying the result of Lemma 42 and Equation (3.251) and the assumption that $m_2 \geq m_3 \log(m_2)$:

$$\begin{aligned}
&A_1 + A_2 \lesssim \eta\varrho_2^2\beta_2^{-1}\Big[(C_1 + \sqrt{m}_3\beta_1)^2/(\kappa_1\sqrt{m}_3) + \Big[\sqrt{m}_3 m_1\beta_1 + C_1\Big]\exp\{-C_1^2/(8m_3\beta_1^2)\}\Big] \\
&\times \Big[\exp\{-C_2^{4/3}(\sqrt{m}_2\kappa_2)^{2/3}/(8\beta_2^2)\} + \frac{C_2^{2/3}}{(\sqrt{m}_2\kappa_2)^{2/3}}\Big] \\
&+ \eta\Big(\kappa_2\sqrt{m_2} + C_1\Big)\Big(\exp\{-c_2^2/(32\beta_1^2)\} + \frac{c_2}{\kappa_1\sqrt{m_1}}\Big)\frac{\varrho^2 m_3\sqrt{m_3}}{\beta_1},
\end{aligned}$$

which completes the proof.

**Setting $\beta_1$ and $\beta_2$**

As we mentioned, to minimize the amount of deviation of the smoothed function compared to the original one, we prefer to choose $\beta_1, \beta_2$ as small as possible. (The benefit of such choice, indeed, can be observed more explicitly in other parts of the proof, e.g. Section 3.6.14.) Observing the bound in Equation (3.241) and noting that we can easily make the exponential terms orders of magnitude smaller than the poly terms, it is easy to find the following optimal setting for the smoothing parameters:

$$\begin{aligned}
\beta_2 &\coloneqq \Theta_p\Big((\kappa_1\sqrt{m}_3)^{-1}(\sqrt{m}_2\kappa_2)^{-\frac{2}{3}}\Big),\\
\beta_1 &\coloneqq \Theta_p\Big(m_3\sqrt{m}_3/(\kappa_1\sqrt{m}_1)\Big).
\end{aligned}$$

Using this setting, we still can make $\Re_8$ arbitrarily small. Here, we remind the reader that $O_p$ only cares about the non-logarithmic dependencies on the overparameterization, i.e. $m_1, m_2, m_3, \kappa_1, \kappa_2$.

### 3.6.21 Basic Tools

In this section, we introduce and prove some lemmas that we use in our analysis as basic tools.

**Lemma 38.** *Suppose $V^{(0)} \in \mathbb{R}^{m_2 \times m_3}$ has standard normal entries and $a$ is a random sign vector. Suppose theta $> 1, R < 1$ are given thresholds, such that*

$$m_2 R \gtrsim m_3(\log(1/R) + \log(m_3) + \log(\log(m_2)),$$

$$e^{-\theta^2/8} \lesssim m_3/m_2.$$

*Then, for the following quantities:*

$$N_R^1(x) = \#\Big(j \in [m] : \ |V_j^{(0)} x| \leq R\Big)$$
$$N_\theta^2(x) = \#\Big(j \in [m] : \ |V_j^{(0)} x| \geq \theta\Big),$$

*with high probability we have*

$$\sup_{\|x'\|=1} N_R^1(x') \lesssim m_2 R,$$
$$\sup_{\|x'\|=1} N_\theta^2(x') \lesssim m_3(\log(m_3) + \log(\log(m_2))).$$

**Proof of Lemma 38**

Suppose $B_1(\epsilon)$ is a cover for the Euclidean ball in $\mathbb{R}^{m_3}$ with precision $\epsilon$. We know

$$|B_1(\epsilon)| \lesssim (1/\epsilon)^{m_3}.$$

Now for a fixed $\|x\| = 1$, we have

$$\mathbb{P}(W_j^{(0)} x \leq 2R) \lesssim R.$$

Therefore, using Bernstein, with high probability we have

$$\#\Big(j \in [m_2] :\ |V_j^{(0)} x| \le 2R\Big) \lesssim m_2 R + \sqrt{m_2 R} + 1.$$

Hence, using union bound, we have with high probability

$$\sup_{x \in B_1(\epsilon)} \#\Big(j \in [m] :\ |V_j^{(0)} x| \le 2R\Big) \lesssim m_2 R + \sqrt{\log |B_1(\epsilon)|}\sqrt{m_2 R} + \log |B_1(\epsilon)|$$

$$= m_2 R + \sqrt{m_2 R m_3 \log(1/\epsilon)} + m_3 \log(1/\epsilon).$$

By picking

$$\epsilon \lesssim R/(\sqrt{m_3 \log(m_2 m_3)}),$$

The assumption implies $m_2 R \ge m_3 \log(1/\epsilon)$, which implies

$$LHS \lesssim m_2 R.$$

On the other hand, note that with high probability we have

$$\sup_{j \in [m_2], k \in [m_3]} |V_{j,k}^{(0)}| \le \sqrt{\log(m_2 m_3)}. \tag{3.255}$$

Now for $\|x'\| = 1$ which is not in the cover, if $x$ is the closest point to it in the cover, i.e. $x \in B_1(\epsilon)$ and $\|x - x'\| \le \epsilon$, then for every $j \in [m_2]$ we have

$$||V_j^{(0)} x| - |V_j^{(0)} x'|| \le \|V_j^{(0)}\| \|x - x'\| \le \sqrt{m_3 \log(m_2 m_3)}\epsilon \le R,$$

by picking a small enough constant. Therefore, for a $j$ that $|V_j^{(0)} x| \ge 2R$, then

$$|V_j^{(0)} x'| \ge 2R - R = R.$$

Therefore, we get that with high probability, for every $\|x'\| = 1$:

$$\sup_{\|x'\|=1} \#\Big(j \in [m_2] : \ |V_j^{(0)} x'| \leq R\Big) \lesssim m_2 R.$$

For the second part, note that for $\|x\| = 1$, by the tail bound for normal vars:

$$\mathbb{P}(W_j^{(0)} x \geq \theta/2) \lesssim e^{-\theta^2/8}.$$

Hence, again using Bernstein, we have with high probability

$$\sup_{x \in B_1(\epsilon)} \#\Big(j \in [m_2] : \ |V_j^{(0)} x| \geq \theta/2\Big) \lesssim m_2 e^{-\theta^2/8} + \sqrt{\log |B_1(\epsilon)|}\sqrt{m_2 e^{-\theta^2/8}} + \log |B_1(\epsilon)|$$

$$\lesssim m_2 e^{-\theta^2/8} + \sqrt{m_3 \log(1/\epsilon)}\sqrt{m_2 e^{-\theta^2/8}} + m_3 \log(1/\epsilon).$$

By picking

$$\epsilon \lesssim 1/(\sqrt{m_3 \log(m_2 m_3)}),$$

and using the assumption $m_2 e^{-\theta^2/8} \lesssim m_3$, all terms are dominated by the third term so we can bound the above as

$$\sup_{x \in B_1(\epsilon)} \#\Big(j \in [m_2] : \ |V_j^{(0)} x| \geq \theta/2\Big) \lesssim m_3(\log(m_3) + \log(\log(m_2))).$$

Now for $\|x'\| = 1$ not in the cover, for the new $\epsilon$ we can write

$$||V_j^{(0)} x| - |V_j^{(0)} x'|| \leq \|V_j^{(0)}\| \|x - x'\| \leq \sqrt{m_3 \log(m_2 m_3)}\epsilon \leq 1/2 \leq \theta/2.$$

Hence, with high prob.

$$\sup_{\|x'\|=1} \#\Big(j \in [m_2] : \ |V_j^{(0)} x| \geq \theta\Big) \lesssim m_3(\log(m_3) + \log(\log(m_2))).$$

**Lemma 39.** *For $x \in \mathbb{R}^d$ and $W^{(0)} \in \mathbb{R}^{m \times d}$ which has standard normal entries (and a is a random sign vector), we have with high probability:*

$$\sup_{\|x\|=1} f(x) := \frac{1}{\sqrt{m}} a^T \sigma(W^{(0)} x) \leq \sqrt{d}.$$

**Proof of Lemma 39**

For the first part, we first compute an upper bound on

$$\mathbb{E} \sup_{\|x\|=1} \frac{1}{\sqrt{m}} a^T \sigma(W^{(0)} x).$$

To do so, we use Dudley's chaining. Note that the for $x_1, x_2 \in \mathbb{R}^d$, the variable $\sigma(W_j^{(0)} x_1) - \sigma(W_j^{(0)} x_2)$ is subGaussian with parameter $\|x_1 - x_2\|$, so the variable $f(x_1) - f(x_2)$ is also subGaussian with parameter $\|x_1 - x_2\|$. Hence, by Dudley's integral:

$$\mathbb{E} \sup_{\|x\|=1} \frac{1}{\sqrt{m}} a^T \sigma(W^{(0)} x) \leq \int_0^1 \sqrt{\log(\mathcal{N}(\mathcal{B}_1^d, \epsilon))} \lesssim \sqrt{d}.$$

Now for a fixed $x$, note that

$$\frac{1}{\sqrt{m}} a^T \sigma(W_1 x) - \frac{1}{\sqrt{m}} a^T \sigma(W_2 x) \leq \frac{1}{\sqrt{m}} \sum_{j=1}^m \|W_{1j} - W_{2j}\| \leq \|W_1 - W_2\|_F.$$

Hence, the function $f(x)$ is 1-lipchitz with respect to $W$ and $l2$ norm, so is the function $\sup f(x)$. Hence, by Gaussian concentration, $\sup f(x)$ is 1-subGaussian around its mean, so we finally get with high probability

$$\sup f(x) \lesssim \sqrt{d} + 1 \lesssim \sqrt{d}.$$

**Lemma 40.** *For*

$$R^* := \frac{C_2^{2/3}}{(\sqrt{m}_2 \kappa_2)^{2/3}},$$

*we have with high probability over the randomness of $V^{(0)}$:*

$$\sup_{x',V':\ \|V'\|\le C_2} \#\Big(j\in[m_2]:\ |(V_j^{(0)}+V_j')x'|\le R^*\kappa_2\|x'\|\Big)\lesssim R^* m_2.$$

**Proof of Lemma 40**

Note that obviously the condition of Lemma 38 is satisfied with this choice of $R = R^*$. Therefore, with high probability we have for an arbitrary $x'$:

$$\#\Big(|V_j^{(0)}x'|\le 2R^*\kappa_2\|x'\|\Big)\le m_2R^*.$$

On the other hand, note that for $j\in[m_2]$ such that $|V_j'x'|\ge R\kappa_2\|x'\|$, we have

$$\|V_j'\|\|x'\|\ge|V_j'x'|\ge R\kappa_2\|x'\|,$$

which implies

$$\|V_j'\|\ge R\kappa_2.$$

Therefore, there are at most $\frac{C_2^2}{R^2\kappa_2^2}$. Therefore, setting aside $m_2R+\frac{C_2^2}{R^2\kappa_2^2}$ of $j$'s, for the rest we have

$$|(V_j^{(0)}+V_j')x'|\ge|V_j^{(0)}x'|-|V_j'x'|\ge 2R\kappa_2\|x'\|-R\kappa_2\|x'\|=R\kappa_2\|x'\|.$$

Setting $R^*$ as defined above balances the terms $m_2R$ and $\frac{C_2^2}{R^2\kappa_2^2}$, which completes the proof.

**Lemma 41.** *If $V^{(0)}\in\mathbb{R}^{m_2\times m_3}$ is a matrix with standard normal entries, then with high probability*

$$\sup_{\|x'\|=1}\frac{1}{\sqrt{m}_2}\sum_{j=1}^{m_2}|V_j^{(0)}x'|\lesssim\sqrt{m_2}+\sqrt{m_3(\log(m_3)+\log(\log(m_2)))}.$$

**Proof of Lemma 41**

Let $B_1(\epsilon)$ be a cover for the unit Euclidean ball with precision $\epsilon$, for which we have $|B_1(\epsilon)| \lesssim (\frac{1}{\epsilon})^{m_3}$. Now for a fixed $x \in B_1(\epsilon)$, note that because $V_j^{(0)}x$ is a standard normal variable, the random variable $|V_j^{(0)}x'| - \mathbb{E}|V_j^{(0)}x'|$ is $O(1)$-subGaussian, which means $\frac{1}{\sqrt{m_2}}\sum_{j=1}^{m_2}(|V_j^{(0)}x'| - \mathbb{E}|V_j^{(0)}x'|)$ is also $O(1)-$subGaussian. Now from the tail of maximum of subGaussian variables:

$$\sup_{x\in B_1(\epsilon)} \frac{1}{\sqrt{m}_2}\sum_{j=1}^{m_2}(|V_j^{(0)}x| - \mathbb{E}|V_j^{(0)}x|) \lesssim \sqrt{\log(|B_1(\epsilon)|)} = \sqrt{m_3\log(1/\epsilon)}.$$

On the other hand, note that $\mathbb{E}|V_j^{(0)}x'|) = O(1)$, which implies w.h.p:

$$\sup_{x\in B_1(\epsilon)} \frac{1}{\sqrt{m}_2}\sum_{j=1}^{m_2}|V_j^{(0)}x| \lesssim \sqrt{m_2} + \sqrt{m_3\log(1/\epsilon)}.$$

Moreover, note that again by the tail of subGaussian variables, we have w.h.p:

$$\max_{j\in[m_2],k\in[m_3]} |V_{j,k}^{(0)}| \lesssim \sqrt{\log(m_2m_3)},$$

which implies with high prob for every $j \in [m_2]$:

$$\|V_j^{(0)}\| \lesssim \sqrt{m_3\log(m_2m_3)}.$$

Now by picking

$$\epsilon := \left(\sqrt{m_3\log(m_2m_3)}\right)^{-1},$$

we get with high probability

$$\sup_{x\in B_1(\epsilon)} \frac{1}{\sqrt{m}_2}\sum_{j=1}^{m_2}|V_j^{(0)}x| \lesssim \sqrt{m_2} + \sqrt{m_3(\log(m_3) + \log(\log(m_2)))}. \tag{3.256}$$

On the other hand, for an arbitrary $x'$ with $\|x'\| = 1$, if $x \in B_1(\epsilon)$ is the representative of $x'$, we have by definition $\|x' - x\| \leq \epsilon$, which combined with (3.6.21) implies

$$\begin{aligned}\Big||V_j^{(0)}x'| - |V_j^{(0)}x|\Big| &\leq |V_j^{(0)}(x' - x)| \leq \|V_j^{(0)}\|\|x' - x\|, \\ &\lesssim \sqrt{m_3 \log(m_2 m_3)}\Big(\sqrt{m_3 \log(m_2 m_3)}\Big)^{-1} \leq 1.\end{aligned}$$

Therefore

$$\Big|\frac{1}{\sqrt{m}_2}\sum_{j=1}^{m_2}|V_j^{(0)}x'| - \frac{1}{\sqrt{m}_2}\sum_{j=1}^{m_2}|V_j^{(0)}x|\Big| \leq \sqrt{m}_2. \tag{3.257}$$

Combining Equations (3.256) and (3.257), we conclude the result.

### Defining the rare events $E_j$

**Lemma 42.** *For $x'^{(2)}$ defined in Equation (3.240) we have*

$$\mathbb{E}_{W^\rho}\|x'^{(2)} - \phi^{(0)}(x_i)\|, \mathbb{E}_{W^\rho}\|\phi^{(2)}(x_i)\| \leq C_1 + \sqrt{m}_3\beta_1,$$
$$\mathbb{E}_{W^\rho}\|\phi^{(2)}(x_i)\|^2 \lesssim C_1^2 + m_3\beta_1^2.$$

*Moreover, for the events*

$$E_j = \{|W_j^\rho x_i| \geq C_1/\sqrt{m_3 m_1}\},\ E = \cup_j E_j,$$

*we have under $E^c$:*

$$\|x'^{(2)} - \phi^{(0)}(x_i)\|, \|\phi^{(2)}(x_i)\| \lesssim C_1.$$

*Furthermore, $E$ happens rarely:*

$$\mathbb{P}(E) \lesssim m_1 \exp\{-C_1^2/(8m_3\beta_1^2)\},$$
$$\mathbb{E}_{W^\rho}\mathbb{1}\{E\}\|\phi^{(2)}(x_i)\| \leq \left[\sqrt{m}_3 m_1\beta_1 + C_1\right]\exp\{-C_1^2/(8m_3\beta_1^2)\}.$$
$$\mathbb{E}_{W^\rho}\mathbb{1}\{E\}\|x'^{(2)} - \phi^{(0)}(x_i)\| \leq \left[\sqrt{m}_3 m_1\beta_1 + C_1\right]\exp\{-C_1^2/(8m_3\beta_1^2)\}.$$

*Finally, we have the following almost surely bound:*

$$\|\phi^{(2)}(x_i)\| \leq C_1 + \sqrt{m}_3 \sum_{j=1}^{m_1} \frac{1}{\sqrt{m}_1}|W_j^\rho x_i|.$$

#### Proof of Lemma 42

We start by writing

$$|x_k'^{(2)} - \frac{1}{\sqrt{m}_1} W_k^s \sigma(W^{(0)} + (1-\eta)W' + \sqrt{\eta}W^*)x_i| \leq \sum_{j=1}^{m_1}\frac{1}{\sqrt{m}_1}|W_j^\rho x_i|. \qquad (3.258)$$

Now notice that by Lemma 10, we know for every $j \notin \tilde{P}$:

$$|W_j^{(0)} x_i| \geq c_2/\sqrt{m_2}, \tag{3.259}$$

$$|(1-\eta)W_j' x_i| \leq c_2/(2\sqrt{m_1}). \tag{3.260}$$

In addition, by Equations in (3.105) from Lemma 14, for every $j \in [m_1]$:

$$\|W_j^*\| \leq \varrho_1 \sqrt{\frac{m_3}{m_1}},$$

so by picking

$$\eta \leq c_2/(4\varrho\sqrt{m_3})$$

we obtain

$$\eta |W_j^* x_i| \leq \frac{c_2}{4\sqrt{m_1}}. \tag{3.261}$$

Combining this with Equations in (3.260), we see that the signs of $(W_j^{(0)} + (1-\eta)W_j' + \sqrt{\eta}W_j^*)x_i$ and $W_j^{(0)}x_i$ are the same for $j \notin \tilde{P}$.

Moreover, the matrix $(1-\eta)W' + \sqrt{\eta}W^*$ satisfies

$$\|(1-\eta)W' + \sqrt{\eta}W^*\| \leq (1-\eta)C_1 + \sqrt{\eta}\sqrt{2\zeta_2} \leq C_1,$$

by picking $\sqrt{\eta} \leq C_1/\sqrt{\zeta}_2$. Hence, the conditions of Lemma 15 are satisfied and we get:

$$\|\frac{1}{\sqrt{m_1}} W^s \sigma(W^{(0)} + (1-\eta)W' + \sqrt{\eta}W^*)x_i - \phi^{(0)}(x_i)\| \leq C_1. \tag{3.262}$$

Combining Equations (3.258) and (3.262):

$$\|x'^{(2)} - \phi^{(0)}(x_i)\| \leq C_1 + \sqrt{m_3} \sum_{j=1}^{m_1} \frac{1}{\sqrt{m_1}} |W_j^\rho x_i|. \tag{3.263}$$

In exactly similar fashion, one can derive

$$\|\phi^{(2)}(x_i)\| \leq C_1 + \sqrt{m}_3 \sum_{j=1}^{m_1} \frac{1}{\sqrt{m}_1} |W_j^\rho x_i|. \tag{3.264}$$

Now first of all, note

$$\mathbb{E}_{W^\rho} \sum_{j=1}^{m_1} \frac{1}{\sqrt{m}_1} |W_j^\rho x_i| \leq \beta_1,$$

which proves the first part of the claims. For the second part, note that by the Gaussian tail bound

$$\mathbb{P}(|W_j^\rho x_i| \geq C_1/\sqrt{m_3 m_1}) \lesssim \exp\{-C_1^2/(8m_3\beta_1^2)\}.$$

Therefore,

$$\mathbb{P}(E) \leq \sum_j \mathbb{P}(E_j) \leq m_1 \exp\{-C_1^2/(8m_3\beta_1^2)\}.$$

Moreover

$$\begin{aligned}
\mathbb{E}_{W^\rho} \mathbb{1}\{E\} \sum_{j=1}^{m_1} \frac{1}{\sqrt{m}_1} |W_j^\rho x_i| &\leq \mathbb{E}_{W^\rho} \frac{1}{\sqrt{m}_1} \sum_{j_2} \sum_{j \neq j_2} \mathbb{1}\{E_{j_2}\} |W_j^\rho x_i| + \frac{1}{\sqrt{m}_1} \sum_j \mathbb{E}_{W^\rho} \mathbb{1}\{E_j\} |W_j^\rho x_i|, \\
&= \Big[\frac{1}{\sqrt{m}_1} \sum_{j_2} \sum_{j \neq j_2} \mathbb{E}_{W^\rho} |W_j^\rho x_i| + \frac{1}{\sqrt{m}_1} \sum_j E_{W^\rho}[|W_j^\rho x_i| \Big| \; E_j]\Big] \mathbb{P}(E_j) \\
&\lesssim \Big[m_1\beta_1 + C_1/\sqrt{m}_3\Big] \exp\{-C_1^2/(8m_3\beta_1^2)\}.
\end{aligned}$$

Plugging this into (3.263) finishes the proof. Also, under $E^c$ by Equation (3.263) we have

$$\|x'^{(2)} - \phi^{(0)}(x_i)\|, \|\phi^{(2)}(x_i)\| \lesssim C_1.$$

Finally, exploiting Equation (3.264):

$$\begin{aligned}
&\mathbb{E}_{W^\rho}\|\phi^{(2)}(x_i)\|^2 \\
&\lesssim C_1^2 + m_3\frac{1}{m_1}\mathbb{E}_{W^\rho}(\sum_j |W_j^\rho x_i|)^2 \leq C_1^2 + m_3\mathbb{E}_{W^\rho}\sum_j |W_j^\rho x_i|^2 \\
&\leq C_1^2 + m_3\beta_1^2.
\end{aligned}$$

### Bounding the value of $f'$

The following Lemma provides a reasonable bound on the value of the smoothed function.

**Lemma 43.** *We have the following general bound on the values of the smoothed function: With high probability over the initialization, for $\|W'\| \leq C_1, \|V'\| \leq C_2$ and $\forall i \in [n]$ (having small enough choices of $\beta_1, \beta_2$ described in Section 3.6.20):*

$$|f'_{W',V'}(x_i)| \lesssim (\kappa_2\sqrt{m}_3 + \beta_2)\Big(\sqrt{m}_3\kappa_1 + C_1 + \sqrt{m}_3\beta_1\Big) + C_2(C_1 + \sqrt{m}_3\beta_1),$$

*which is $O(C_1C_2)$ for large enough overparameterization as described in Section3.6.3. Moreover, we have the following almost surely bound (with respect to the randomness of $W^\rho$ and $V^\rho$):*

$$|f_{W'+W^\rho,V'+V^\rho}(x_i)| \\ \lesssim (\kappa_2\sqrt{m}_3 + \|V^\rho\|_F)\Big(\sqrt{m}_3\kappa_1 + C_1 + \sqrt{m_3}(\frac{1}{\sqrt{m}_1}\sum_j |W^\rho_j x_i|)\Big) + C_2(C_1 + \sqrt{m_3}(\frac{1}{\sqrt{m}_1}\sum_j |W^\rho_j x_i|)).$$

*Notably, with slightly higher overparameterization, the high probability bound in (43) holds even if we take supremum over $x$.*

### Proof of Lemma 43

Using Lemmas 39 and 42 and using the fact that $\phi^{(0)}(x_i)$ is orthogonal to the rows of $V'$ (recall $x'_i = \phi^{(0)}(x_i) + \phi^{(2)}(x_i)$):

$$
\begin{aligned}
&|f_{W'+W^\rho,V'+V^\rho}(x_i)| \\
&\leq \frac{1}{\sqrt{m}_2} a^T \sigma(V^{(0)} x'_i) + \frac{1}{\sqrt{m}_2} \sum_j |(V^\rho_j + V'_j) x'_i| \\
&\leq (\kappa_2 \sqrt{m}_3) \|x'_i\| + \frac{1}{\sqrt{m}_2} \sum_j |V^\rho_j x'_i| + C_2 \|\phi^{(2)}(x_i)\| \\
&\leq (\kappa_2 \sqrt{m}_3 + \|V^\rho\|_F) \|x'_i\| + C_2 \|\phi^{(2)}(x_i)\| \\
&\lesssim (\kappa_2 \sqrt{m}_3 + \|V^\rho\|_F) \Big( \sqrt{m}_3 \kappa_1 + C_1 + \sqrt{m}_3 ( \frac{1}{\sqrt{m}_1} \sum_j |W^\rho_j x_i| ) \Big) + C_2 (C_1 + \sqrt{m}_3 ( \frac{1}{\sqrt{m}_1} \sum_j |W^\rho_j x_i| )).
\end{aligned}
\tag{3.265}
$$

Note that above, if we apply the stronger worst-case norm bound of the first layer's output presented in Lemma 55, we would get $\sup_{x, \|x\|=1} |f_{W'+W^\rho,V'+V^\rho}(x)|$ is bounded by the RHS, which in turn proves a stronger uniform bound on $f'$.

Similarly, this time by taking expectation with respect to $W^\rho$ and $V^\rho$:

$$
\begin{aligned}
|f'_{W',V'}(x_i)| &= |\mathbb{E}_{W^\rho,V^\rho} f_{W',V'}(x_i)| \\
&\leq \mathbb{E}_{W^\rho,V^\rho} |f_{W',V'}(x_i)| \\
&= \mathbb{E}_{W^\rho} (\kappa_2 \sqrt{m}_3 + \beta_2) \|x'_i\| + C_2 \|x'_i - \phi^{(0)}(x_i)\| \\
&\lesssim (\kappa_2 \sqrt{m}_3 + \beta_2) \Big( \sqrt{m}_3 \kappa_1 + C_1 + \sqrt{m}_3 \beta_1 \Big) + C_2 (C_1 + \sqrt{m}_3 \beta_1).
\end{aligned}
$$

**Corollary 8.1.** *If we set $C_1 = C_2 = 0$ above, we get*

$$
|f'_{0,0}(x_i)| \leq (\kappa_2 \sqrt{m_3} + \beta_2) \Big( \sqrt{m}_3 \kappa_1 + \sqrt{m}_3 \beta_1 \Big),
$$

*the point being these terms go to zero by an order of $O((\sqrt{m}_2 \kappa_2)^{-\frac{2}{3}})$. Therefore, taking $(\sqrt{m}_2 \kappa_2)^{-\frac{2}{3}} << B$, we make sure that $|f'_{0,0}| < B$, so by the $1$ smoothness of $\ell$ and $B$ boundedness of the labels we get $\ell(f'_{0,0}(x_i), y_i) < 4B^2$.*

**Bounding the difference between Original and Smoothed Functions**

The following Lemma bounds the difference between the smoothed function and original function of the network.

**Lemma 44.** *Bound on the smoothing change under the assumption $m_2 \geq m_3 \log(m_2)$: with high probability over the initialization, for any $(W', V')$ with $\|W'\| \leq C_1, \|V'\| \leq C_2$:*

$$\begin{aligned}
&|f_{W',V'}(x_i) - f'_{W',V'}(x_i)| \\
&\leq \beta_2(\kappa_1\sqrt{m}_3 + C_1 + \sqrt{m}_3\beta_1) + \Big(C_2 + \kappa_2\sqrt{m_2}\Big)\sqrt{m}_3\beta_1.
\end{aligned}$$

### Proof of Lemma 44

We write

$$\begin{aligned}
&|f_{W',V'}(x_i) - f'_{W',V'}(x_i)| = |f_{W',V'}(x_i) - \mathbb{E}_{W^\rho,V^\rho} f_{W'+W^\rho,V'+V^\rho}(x_i)| \\
&= \Big|\mathbb{E}_{W^\rho,V^\rho}\Big(f_{W',V'}(x_i) - f_{W'+W^\rho,V'+V^\rho}(x_i)\Big)\Big| \\
&\leq \mathbb{E}_{W^\rho,V^\rho}\Big|f_{W',V'}(x_i) - f_{W'+W^\rho,V'+V^\rho}(x_i)\Big|.
\end{aligned}$$

In the following, $\sigma$ means we apply Relu activation to the vector in front of it (entrywise):

$$\begin{aligned}
LHS \leq \mathbb{E}_{W^\rho,V^\rho}\Big| &\frac{1}{\sqrt{m}_2} a^T\sigma(V^{(0)} + V' + V^\rho)\frac{1}{\sqrt{m}_1} W^s\sigma(W^{(0)} + W' + W^\rho)x_i \\
&- \frac{1}{\sqrt{m}_2} a^T\sigma(V^{(0)} + V')\frac{1}{\sqrt{m}_1} W^s\sigma(W^{(0)} + W' + W^\rho)x_i\Big| \\
+\mathbb{E}_{W^\rho,V^\rho}\Big| &\frac{1}{\sqrt{m}_2} a^T\sigma(V^{(0)} + V')\frac{1}{\sqrt{m}_1} W^s\sigma(W^{(0)} + W' + W^\rho)x_i \\
&- \frac{1}{\sqrt{m}_2} a^T\sigma(V^{(0)} + V')\frac{1}{\sqrt{m}_1} W^s\sigma(W^{(0)} + W')x_i\Big|.
\end{aligned}$$

Now for the first term above, using the previous notation of $x_i'$ representing the output of the first layer and using Lemma 42:

$$\begin{aligned}
\mathbb{E}_{W^\rho, V^\rho}\Big| &\frac{1}{\sqrt{m_2}} a^T \sigma(V^{(0)} + V' + V^\rho) \frac{1}{\sqrt{m_1}} W^s \sigma(W^{(0)} + W' + W^\rho) x_i \\
&- \frac{1}{\sqrt{m_2}} a^T \sigma(V^{(0)} + V') \frac{1}{\sqrt{m_1}} W^s \sigma(W^{(0)} + W' + W^\rho) x_i \Big| \\
&\leq \mathbb{E}_{W^\rho, V^\rho} \frac{1}{\sqrt{m_2}} \sum_j |V_j^\rho x_i'| \\
&\lesssim \beta_2 \mathbb{E}_{W^\rho} \|x_i'\| \\
&\leq \beta_2 \mathbb{E}_{W^\rho, V^\rho} (\|\phi^{(0)}(x_i)\| + \|\phi^{(2)}(x_i)\|) \\
&\lesssim \beta_2 (\kappa_1 \sqrt{m_3} + C_1 + \sqrt{m_3} \beta_1).
\end{aligned} \tag{3.266}$$

For the second term, by starting off with a simple triangle inequality:

$$\begin{aligned}
\mathbb{E}_{W^\rho, V^\rho}\Big| &\frac{1}{\sqrt{m_2}} a^T \sigma(V^{(0)} + V') x_i' - \frac{1}{\sqrt{m_2}} a^T \sigma(V^{(0)} + V')(\phi^{(0)}(x_i) + \phi^{(2)\not\rho}(x_i)) \Big| \\
&\leq \mathbb{E}_{W^\rho, V^\rho} \frac{1}{\sqrt{m_2}} \sum_{j=1}^{m_2} \Big| (V_j^{(0)} + V_j')(x_i' - \phi^{(0)}(x_i) - \phi^{(2)\not\rho}(x_i)) \Big| \\
&\leq \mathbb{E}_{W^\rho, V^\rho} \frac{1}{\sqrt{m_2}} \sum_{j=1}^{m_2} \Big| V_j^{(0)}(x_i' - \phi^{(0)}(x_i) - \phi^{(2)\not\rho}(x_i)) \Big| + \Big| V_j'(x_i' - \phi^{(0)}(x_i) - \phi^{(2)\not\rho}(x_i)) | \Big| \\
&\leq C_2 E_{W^\rho} \|x_i' - \phi^{(0)}(x_i) - \phi^{(2)\not\rho}(x_i)\| + \mathbb{E}_{W^\rho} \frac{1}{\sqrt{m_2}} \sum_{j=1}^{m_2} \Big| V_j^{(0)}(x_i' - \phi^{(0)}(x_i) - \phi^{(2)\not\rho}(x_i)) \Big|.
\end{aligned}$$

Now using Lemma 41:

$$\begin{aligned}
&\lesssim \Big(C_2 + \kappa_2 \sqrt{m_2}\Big) \mathbb{E}_{W^\rho} \|x_i' - \phi^{(0)}(x_i) - \phi^{(2)\not\rho}(x_i)\| \\
&\lesssim \Big(C_2 + \kappa_2 \sqrt{m_2}\Big) \sqrt{m_3} E_{W^\rho} \sum_j \frac{1}{\sqrt{m_1}} |W_j^\rho x_i| \\
&\lesssim \Big(C_2 + \kappa_2 \sqrt{m_2}\Big) \sqrt{m_3} \beta_1.
\end{aligned} \tag{3.267}$$

Combining Equations (3.266) and (3.267) we conclude the proof.

## 3.7 Appendix

### 3.7.1 Smoothness coefficients

Recall that for a function $f \in \mathcal{C}^3$ on $\mathbb{R}^d$, we say it is $\mu_1$ Lipschitz, $\mu_2$ gradient Lipschitz, and $\mu_3$ hessian Lipschitz at point $x$ if for every unit direction $v$, $|\frac{d}{d\lambda} f(x + \lambda v)| \leq \mu_1$, $|\frac{d^2}{d\lambda^2} f(x + \lambda v)| \leq \mu_2$, and $|\frac{d^2}{d\lambda^2} f(x + \lambda v)| \leq \mu_3$.

The aim of this section is to bound the Lipschitz coefficients of the loss $\ell(, y)$ and objective $L(W', V')$ in a bounded domain $\|W'\| \leq C_1, \|V'\| \leq C_2$. The following is our main Theorem in this regard:

**Theorem 9.** *For given values $C_1, C_2 > 0$, in the domain $\|W'\| \leq C_1, \|V'\| \leq C_2$, for any label $|y| \leq B$, the loss function $\ell(., y)$ is $O((C_1C_2 + B^2))$-Lipschitz (having enough overparameterization) and $1$ gradient-Lipschitz $x = f'_{W',V'}$. Moreover, the loss function $L(W', V')$ is $(O(C_1C_2) + B)\Psi_1 + 2(C_1 + C_2)$ Lipschitz, $\Psi_1^2 + (O(C_1C_2) + B)\Psi_2 + 4$ gradient Lipschitz, and $3\Psi_2\Psi_1 + (O(C_1C_2) + B)\Psi_3$ hessian Lipschitz, where $\Psi_1, \Psi_2, \Psi_3$ are defined in Lemma 45.*

**Proof of Lemma 9**

As in the proof of Lemma 45, let $(\tilde{W}, \tilde{V})$ be a unit direction, i.e. $\|\tilde{W}\|^2 + \|\tilde{V}\|^2 = 1$. Then, using Lemma 43, we know that for every $i \in [n]$: $|f'_{W',V'}(x_i)| = O(C_1C_2)$, so by 1-smoothness of the loss and $B$-boundedness of the labels, we get that $\ell(., y)$ is $(O(C_1C_2) + B)$ lipshcitz at point $f'_{W',V'}$. The gradient smoothness parameter of the square loss $\ell$ is bounded by 1 and its third derivative is zero. Now using these coefficients, we can easily compute the coefficients for $L$ as well by simple differentiation:

$$
\begin{aligned}
&|\frac{d}{d\lambda}\ell(f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i),y_i)| = |\dot{\ell}(f',y_i)\frac{d}{d\lambda}f'| \leq (O(C_1C_2)+B)\Psi_1.\\
&|\frac{d}{d\lambda^2}\ell(f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i),y_i)| = |\ddot{\ell}(f',y_i)(\frac{d}{d\lambda}f')^2 + \dot{\ell}(f',y_i)\frac{d^2}{d\lambda^2}f'| \leq \Psi_1^2 + (O(C_1C_2)+B)\Psi_2.\\
&|\frac{d}{d\lambda^3}\ell(f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i),y_i)| = |\dddot{\ell}(f',y_i)(\frac{d}{d\lambda}f')^3 + 3\ddot{\ell}(f',y_i)\frac{d^2}{d\lambda^2}f'\frac{d}{d\lambda}f' + \dot{\ell}(f',y_i)\frac{d^3}{d\lambda^3}f'|\\
&\leq \Psi_1^3 + 3\Psi_2\Psi_1 + (O(C_1C_2)+B)\Psi_3.
\end{aligned}
$$

Moreover, note that

$$
\begin{aligned}
&\frac{d}{d\lambda}\|W'+\lambda\tilde{W}\|^2 = 2\langle W'+\lambda\tilde{W},\tilde{W}\rangle\Big|_{\lambda=0} = 2\langle W',\tilde{W}\rangle \leq 2\|W'\| = 2C_1,\\
&\frac{d^2}{d\lambda^2}\|W'+\lambda\tilde{W}\|^2 = \langle\tilde{W},\tilde{W}\rangle = 2,\\
&\frac{d^3}{d\lambda^3}\|W'+\lambda\tilde{W}\|^2 = 0,
\end{aligned}
$$

and similarly for $\|V'+\lambda\tilde{V}\|^2$. Combining these results finishes the proof.

Above, we used parameters $\Psi_1,\Psi_2,\Psi_3$, the Lipschitz coefficients of $f'$ in domain $\|W'\| \leq C_1, \|V'\| \leq C_2$, which we bound in Lemma 45 below.

### Computing the Lipschitz Coefficients of $f'_{W',V'}$

In this section, we bound the Lipschitz coefficients of $f'_{W',V'}$ in the domain $\|W'\| \leq C_1, \|V'\| \leq C_2$ by $poly(m_1, m_2, m_3, \beta_1, \beta_2)$ functions.

**Lemma 45.** *For every point $(W', V')$ in the domain $\|W'\| \leq C_1, \|V'\| \leq C_2$, we have the following bounds on the Lipschitz coefficients of $f'_{W',V'}$ ($(\tilde{W}, \tilde{V})$ is a unit direction with $\|\tilde{W}\|^2 + \|\tilde{V}\|^2 = 1$):*

$$
\begin{aligned}
&\Big|\frac{d}{d\lambda} f'_{W'+\lambda\tilde{W}, V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0}\Big| \\
&\lesssim \frac{m_2}{\beta_2^2}\Big(\frac{\beta_2}{\sqrt{m_2}}\sqrt{m_3}(\kappa_1 + C_1 + \beta_1)\Big(\kappa_2\sqrt{m_3} + C_2\Big) + \frac{\beta_2^2}{\sqrt{m_2}}\sqrt{m_3}\Big(\kappa_1 + C_1 + \beta_1\Big)\Big) \\
+&\frac{m_1}{\beta_1^2}\Big(\sqrt{m_3}\Big(\frac{\beta_1}{\sqrt{m_1}}\Big(\kappa_1 + C_1\Big) + \frac{\beta_1^2}{\sqrt{m_1}}\Big)\Big(\kappa_2\sqrt{m_3} + C_2\Big) + \beta_2\sqrt{m_3}\Big(\frac{\beta_1}{\sqrt{m_1}}\Big(\kappa_1 + C_1\Big) + \frac{\beta_1^2}{\sqrt{m_1}}\Big)\Big) := \Psi_1, \\
&\frac{d^2}{d\lambda^2} f'_{W'+\lambda\tilde{W}, V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0} \\
&\lesssim \Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2 + \frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)\sqrt{m_3\Big((\kappa_2\sqrt{m_3} + C_2)^2 + \beta_2^2\Big)\Big[(\kappa_1 + C_1)^2 + \beta_1^2\Big]} := \Psi_2 \\
&\Big|\frac{d^3}{d\lambda^3} f'_{W'+\lambda\tilde{W}, V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0}\Big| \\
&\lesssim \Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2 + \frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)^{3/2}\sqrt{m_3\Big((\kappa_2\sqrt{m_3} + C_2)^2 + \beta_2^2\Big)\Big[(\kappa_1 + C_1)^2 + \beta_1^2\Big]} := \Psi_3
\end{aligned}
\tag{3.268}
$$

#### Proof of Lemma 45

Let

$$
\rho(W^\rho, V^\rho) := \frac{1}{(\sqrt{2\pi})^{m_2 m_3 + m_1 d}(\beta_1/\sqrt{m_1})^{m_1 d}(\beta_2/\sqrt{m_2})^{m_2 m_3}} \exp\{-\frac{\|W^\rho\|^2}{2\beta_1^2/m_1} - \frac{\|V^\rho\|^2}{2\beta_2^2/m_2}\},
$$

be the density function of the law of $W^\rho$ and $V^\rho$ which is a joint Gaussian. Then to compute the derivative and second derivative of the function in the unit direction $(\tilde{W}, \tilde{V})$, s.t. $\|\tilde{W}\|_F^2 + \|\tilde{V}\|_F^2$, we can write the value of the smoothed function as an integration with density $\rho$, change variable, and then take derivatives:

$$
\begin{aligned}
&\frac{d}{d\lambda} f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0} \\
&= \frac{d}{d\lambda}\mathbb{E}_{W^\rho,V^\rho} f_{W'+\lambda\tilde{W}+W^\rho,V'+\lambda\tilde{V}+V^\rho}(x_i) \\
&= \frac{d}{d\lambda}\int f_{W'+\lambda\tilde{W}+W^\rho,V'+\lambda\tilde{V}+V^\rho}(x_i)\rho(W^\rho,V^\rho)d(W^\rho,V^\rho) \\
&= \frac{d}{d\lambda}\int f_{W^+,V^+}(x_i)\rho(W^+-(W'+\lambda\tilde{W}),V^+-(V'+\lambda\tilde{V}))d(W^+,V^+).
\end{aligned}
$$

But one can easily see that for fixed $V'$ and $\tilde{V}$, the set of functions $f_{W^+,V^+}(x_i)\rho(W^+-(W'+\lambda\tilde{W}),V^+-(V'+\lambda\tilde{V}))$ for a small neighborhood of $\lambda$ can simultaneously be upper bounded by an integrable function. Hence, the Leibnitz rule holds here because of dominated convergence theorem, and we can change the order of integration and derivation:

$$
\begin{aligned}
&= \int f_{W^+,V^+}(x_i)\frac{d}{d\lambda}\rho(W^+-(W'+\lambda\tilde{W}),V^+-(V'+\lambda\tilde{V}))d(W^+,V^+) \\
&= -\int f_{W^+,V^+}(x_i)\Big\langle(\tilde{W},\tilde{V}),((\frac{\beta_1^2}{m_1})^{-1}\Big(W^+-(W'+\lambda\tilde{W})\Big),(\frac{\beta_2^2}{m_2})^{-1}\Big(V^+-(V'+\lambda\tilde{V})\Big))\Big\rangle \\
&\rho(W^+-(W'+\lambda\tilde{W}),V^+-(V'+\lambda\tilde{V}))d(W^+,V^+) \\
&= \int f_{W^+,V^+}(x_i)\Big\langle(\tilde{W},\tilde{V}),(\frac{m_1}{\beta_1^2}W^\rho,\frac{m_2}{\beta_2^2}V^\rho)\Big\rangle\rho(W^\rho,V^\rho)d(W^\rho,V^\rho) \\
&= \mathbb{E}_{W^\rho,V^\rho}\Big(\frac{m_1}{\beta_1^2}\Big\langle\tilde{W},W^\rho\Big\rangle+\frac{m_2}{\beta_2^2}\Big\langle\tilde{V},V^\rho\Big\rangle\Big)f_{W'+\lambda\tilde{W}+W^\rho,V'+\lambda\tilde{V}+V^\rho(x_i)}(x_i)\Big|_{\lambda=0} \\
&= \mathbb{E}_{W^\rho,V^\rho}\Big(\frac{m_1}{\beta_1^2}\Big\langle\tilde{W},W^\rho\Big\rangle+\frac{m_2}{\beta_2^2}\Big\langle\tilde{V},V^\rho\Big\rangle\Big)f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i). \qquad (3.269)
\end{aligned}
$$

Similarly we can compute the second derivative:

$$\frac{d^2}{d\lambda^2} f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0}$$

$$= \frac{d}{d\lambda}\int f_{W^+,V^+}(x_i)\Big\langle(\tilde{W},\tilde{V}),((\frac{\beta_1^2}{m_1})^{-1}\Big(W^+-(W'+\lambda\tilde{W})\Big),(\frac{\beta_2^2}{m_2})^{-1}\Big(V^+-(V'+\lambda\tilde{V})\Big))\Big\rangle$$

$$\rho(W^+-(W'+\lambda\tilde{W}),V^+-(V'+\lambda\tilde{V}))d(W^+,V^+)$$

$$= \int f_{W^+,V^+}(x_i)\Big[\Big\langle(\tilde{W},\tilde{V}),((\frac{\beta_1^2}{m_1})^{-1}\Big(W^+-(W'+\lambda\tilde{W})\Big),(\frac{\beta_2^2}{m_2})^{-1}\Big(V^+-(V'+\lambda\tilde{V})\Big))\Big\rangle^2-$$

$$\Big\langle(\tilde{W},\tilde{V}),((\frac{\beta_1^2}{m_1})^{-1}\tilde{W},(\frac{\beta_2^2}{m_2})^{-1}\tilde{V})\Big\rangle\Big]\rho(W^+-(W'+\lambda\tilde{W}),V^+-(V'+\lambda\tilde{V}))d(W^+,V^+)$$

$$= \int f_{W^+,V^+}(x_i)\Big[\Big\langle(\tilde{W},\tilde{V}),(\frac{m_1}{\beta_1^2}W^\rho,\frac{m_2}{\beta_2^2}V^\rho)\Big\rangle^2+\Big\langle(\tilde{W},\tilde{V}),(\frac{m_1}{\beta_1^2}\tilde{W},\frac{m_2}{\beta_2^2}\tilde{V})\Big\rangle\Big]\rho(W^\rho,V^\rho)d(W^\rho,V^\rho)$$

$$= \mathbb{E}_{W^\rho,V^\rho}\Big[\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)^2-\Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)\Big]f_{W'+\lambda\tilde{W}+W^\rho,V'+\lambda\tilde{V}+V^\rho(x_i)}(x_i)\Big|_{\lambda=0}$$

$$= \mathbb{E}_{W^\rho,V^\rho}\Big[\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)^2-\Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)\Big]f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i). \tag{3.270}$$

Similarly for the third derivative:

$$\frac{d^3}{d\lambda^3} f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0}$$

$$\frac{d}{d\lambda}\int f_{W^+,V^+}(x_i)\Big[\Big\langle(\tilde{W},\tilde{V}),((\frac{\beta_1^2}{m_1})^{-1}\Big(W^+-(W'+\lambda\tilde{W})\Big),(\frac{\beta_2^2}{m_2})^{-1}\Big(V^+-(V'+\lambda\tilde{V})\Big))\Big\rangle^2-$$

$$\Big\langle(\tilde{W},\tilde{V}),((\frac{\beta_1^2}{m_1})^{-1}\tilde{W},(\frac{\beta_2^2}{m_2})^{-1}\tilde{V})\Big\rangle\Big]\rho(W^+-(W'+\lambda\tilde{W}),V^+-(V'+\lambda\tilde{V}))d(W^+,V^+)$$

$$= \int f_{W^+,V^+}(x_i)\Big[\Big\langle(\tilde{W},\tilde{V}),((\frac{\beta_1^2}{m_1})^{-1}\Big(W^+-(W'+\lambda\tilde{W})\Big),(\frac{\beta_2^2}{m_2})^{-1}\Big(V^+-(V'+\lambda\tilde{V})\Big))\Big\rangle^3$$

$$-3\Big\langle(\tilde{W},\tilde{V}),((\frac{\beta_1^2}{m_1})^{-1}\Big(W^+-(W'+\lambda\tilde{W})\Big),(\frac{\beta_2^2}{m_2})^{-1}\Big(V^+-(V'+\lambda\tilde{V})\Big))\Big\rangle$$

$$\times\Big\langle(\tilde{W},\tilde{V}),\Big((\frac{\beta_1^2}{m_1})^{-1}\tilde{W},(\frac{\beta_2^2}{m_2})^{-1}\tilde{V}\Big)\Big\rangle\Big]\rho(W^+-(W'+\lambda\tilde{W}),V^+-(V'+\lambda\tilde{V}))d(W^+,V^+)$$

$$= \mathbb{E}_{W^\rho,V^\rho}\Big[\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)^3-3\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)\Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)$$

$$f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\Big].$$

Now for first derivative, exactly similar to the derivation in (3.265), we can write

$$\begin{aligned}
&\Big|\mathbb{E}_{W^\rho,V^\rho}\Big(\frac{m_1}{\beta_1^2}\Big\langle\tilde{W},W^\rho\Big\rangle+\frac{m_2}{\beta_2^2}\Big\langle\tilde{V},V^\rho\Big\rangle\Big)f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\Big|\\
&\leq \mathbb{E}_{W^\rho,V^\rho}\Big|\frac{m_1}{\beta_1^2}\Big\langle\tilde{W},W^\rho\Big\rangle+\frac{m_2}{\beta_2^2}\Big\langle\tilde{V},V^\rho\Big\rangle\Big|\Big|f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\Big|\\
&\leq \mathbb{E}_{W^\rho,V^\rho}\Big|\frac{m_1}{\beta_1^2}\Big\langle\tilde{W},W^\rho\Big\rangle+\frac{m_2}{\beta_2^2}\Big\langle\tilde{V},V^\rho\Big\rangle\Big|\Big(\kappa_2\sqrt{m}_3\|x_i'^{(2)}\|+\|V'\|_F\|x_i'^{(2)}\|+\mathbb{E}_{V^\rho}\frac{1}{\sqrt{m}_2}\sum_j|V_j^\rho x_i'^{(2)}|\Big)\\
&\leq \mathbb{E}_{W^\rho,V^\rho}\Big(\Big|\frac{m_1}{\beta_1^2}\Big\langle\tilde{W},W^\rho\Big\rangle\Big|+\Big|\frac{m_2}{\beta_2^2}\Big\langle\tilde{V},V^\rho\Big\rangle\Big|\Big)\Big(\kappa_2\sqrt{m}_3\|x_i'^{(2)}\|+\|V'\|_F\|x_i'^{(2)}\|+\frac{1}{\sqrt{m}_2}\sum_j|V_j^\rho x_i'^{(2)}|\Big)\\
&=\frac{m_2}{\beta_2^2}\Big(E_{V^\rho}\Big|\Big\langle\tilde{V},V^\rho\Big\rangle\Big|E_{W^\rho}\|x_i'^{(2)}\|\Big(\kappa_2\sqrt{m}_3+\|V'\|_F\Big)+\mathbb{E}_{W^\rho}\mathbb{E}_{V^\rho}\frac{1}{\sqrt{m}_2}\Big|\Big\langle\tilde{V},V^\rho\Big\rangle\Big|\sum_j|V_j^\rho x_i'^{(2)}|\Big)\\
&+\frac{m_1}{\beta_1^2}\Big(\Big(\mathbb{E}_{W^\rho}\|x_i'^{(2)}\|\Big|\Big\langle\tilde{W},W^\rho\Big\rangle\Big|\Big)\Big(\kappa_2\sqrt{m}_3+\|V'\|_F\Big)+\mathbb{E}_{W^\rho}\Big|\Big\langle\tilde{W},W^\rho\Big\rangle\Big|\mathbb{E}_{V^\rho}\frac{1}{\sqrt{m}_2}\sum_j|V_j^\rho x_i'^{(2)}|\Big)\\
&\lesssim\frac{m_2}{\beta_2^2}\Big(E_{V^\rho}\Big|\Big\langle\tilde{V},V^\rho\Big\rangle\Big|E_{W^\rho}\|x_i'^{(2)}\|\Big(\kappa_2\sqrt{m}_3+C_2\Big)+\mathbb{E}_{W^\rho}\mathbb{E}_{V^\rho}\frac{1}{\sqrt{m}_2}\Big|\Big\langle\tilde{V},V^\rho\Big\rangle\Big|\sum_j|V_j^\rho x_i'^{(2)}|\Big)\\
&+\frac{m_1}{\beta_1^2}\Big(\Big(\mathbb{E}_{W^\rho}\|x_i'^{(2)}\|\Big|\Big\langle\tilde{W},W^\rho\Big\rangle\Big|\Big)\Big(\kappa_2\sqrt{m}_3+C_2\Big)+\mathbb{E}_{W^\rho}\Big|\Big\langle\tilde{W},W^\rho\Big\rangle\Big|\mathbb{E}_{V^\rho}\frac{1}{\sqrt{m}_2}\sum_j|V_j^\rho x_i'^{(2)}|\Big),
\end{aligned}$$

where the last line follows because $\|V'\|\lesssim C_2$. But notice that because $\|\tilde{V}\|_F\leq 1$, $\|\tilde{W}\|_F\leq 1$, then $\Big\langle\tilde{V},V^\rho\Big\rangle$ and $\Big\langle\tilde{W},W^\rho\Big\rangle$ are Gaussian variables with variances at most $\beta_2^2/m_2$ and $\beta_1^2/m_1$. Hence

$$E_{V^\rho}\Big|\Big\langle\tilde{V},V^\rho\Big\rangle\Big|\lesssim\beta_2/\sqrt{m}_2, \tag{3.271}$$

$$E_{W^\rho}\Big|\Big\langle\tilde{W},W^\rho\Big\rangle\Big|\lesssim\beta_1/\sqrt{m}_1. \tag{3.272}$$

Similarly, using the same derivation as in (42), one can also get the following a.s. bound (over the randomness of $W^\rho$):

$$\|x_i'^{(2)}\|\lesssim\sqrt{m}_3\Big(\kappa_1+C_1+\frac{1}{\sqrt{m}_1}\sum_j|W_j^\rho x_i|\Big), \tag{3.273}$$

therefore

$$
\begin{aligned}
E_{W^\rho}\|x_i'^{(2)}\| &\lesssim \sqrt{m}_3\Big(\kappa_1 + C_1 + E_{W^\rho}\frac{1}{\sqrt{m}_1}\sum_j |W_j^\rho x_i|\Big) && (3.274)\\
&\lesssim \sqrt{m}_3\Big(\kappa_1 + C_1 + \beta_1\Big). && (3.275)
\end{aligned}
$$

Moreover, for every $j \in [m_2]$:

$$
\begin{aligned}
E_{V^\rho}\Big|\Big\langle \tilde{V}, V^\rho\Big\rangle\Big||V_j^\rho x_i'^{(2)}| &\le \sqrt{E_{V^\rho}\Big|\Big\langle \tilde{V}, V^\rho\Big\rangle\Big|^2}\sqrt{E_{V^\rho}|V_j^\rho x_i'^{(2)}|^2}\\
&= \frac{\beta_2}{\sqrt{m}_2}\frac{\beta_2}{\sqrt{m}_2}\|x_i'^{(2)}\| = \frac{\beta_2^2}{m_2}\|x_i'^{(2)}\|,\\
E_{W^\rho}\Big|\Big\langle \tilde{W}, W^\rho\Big\rangle\Big||W_j^\rho x_i| &\le \frac{\beta_1^2}{m_1}. && (3.276)
\end{aligned}
$$

Similarly, using Equation (3.273) we bound

$$
\begin{aligned}
\mathbb{E}_{W^\rho}\|x_i'^{(2)}\|\Big|\Big\langle \tilde{W}, W^\rho\Big\rangle\Big| &\lesssim \sqrt{m}_3\Big(\mathbb{E}_{W^\rho}\Big|\Big\langle \tilde{W}, W^\rho\Big\rangle\Big|\Big(\kappa_1 + C_1\Big) + \mathbb{E}_{W^\rho}\frac{1}{\sqrt{m}_1}\sum_j \Big|\Big\langle \tilde{W}, W^\rho\Big\rangle\Big||W_j^\rho x_i|\Big)\\
&\le \sqrt{m}_3\Big(\frac{\beta_1}{\sqrt{m}_1}\Big(\kappa_1 + C_1\Big) + \frac{\beta_1^2}{\sqrt{m}_1}\Big). && (3.277)
\end{aligned}
$$

Now applying these bounds (3.272), (3.275), (3.276), and (3.277) to (3.271) and using the fact that

$$
\mathbb{E}_{V^\rho}\frac{1}{\sqrt{m}_2}\sum_j |V_j^\rho x_i'^{(2)}| \le \beta_2\|x_i'^{(2)}\|,
$$

we get

$$
\begin{aligned}
LHS \lesssim &\frac{m_2}{\beta_2^2}\Big(\frac{\beta_2}{\sqrt{m}_2}\sqrt{m_3}(\kappa_1 + C_1 + \beta_1)\Big(\kappa_2\sqrt{m}_3 + C_2\Big) + \mathbb{E}_{W^\rho}\frac{\beta_2^2}{\sqrt{m}_2}\|x_i'^{(2)}\|\Big)\\
&+\frac{m_1}{\beta_1^2}\Big(\sqrt{m}_3\Big(\frac{\beta_1}{\sqrt{m}_1}\Big(\kappa_1 + C_1\Big) + \frac{\beta_1^2}{\sqrt{m}_1}\Big)\Big(\kappa_2\sqrt{m}_3 + C_2\Big) + \mathbb{E}_{W^\rho}\Big|\Big\langle \tilde{W}, W^\rho\Big\rangle\Big|\beta_2\|x_i'^{(2)}\|\Big)\\
\lesssim &\frac{m_2}{\beta_2^2}\Big(\frac{\beta_2}{\sqrt{m}_2}\sqrt{m_3}(\kappa_1 + C_1 + \beta_1)\Big(\kappa_2\sqrt{m}_3 + C_2\Big) + \frac{\beta_2^2}{\sqrt{m}_2}\sqrt{m}_3\Big(\kappa_1 + C_1 + \beta_1\Big)\Big)\\
&+\frac{m_1}{\beta_1^2}\Big(\sqrt{m}_3\Big(\frac{\beta_1}{\sqrt{m}_1}\Big(\kappa_1 + C_1\Big) + \frac{\beta_1^2}{\sqrt{m}_1}\Big)\Big(\kappa_2\sqrt{m}_3 + C_2\Big) + \beta_2\sqrt{m}_3\Big(\frac{\beta_1}{\sqrt{m}_1}\Big(\kappa_1 + C_1\Big) + \frac{\beta_1^2}{\sqrt{m}_1}\Big)\Big).
\end{aligned}
$$

To make it easier for handling the second and third derivatives, we first bound the expectations of $f^2_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)$ which enables us to use Cauchy-schwartz. Again using similar derivation as in (3.265) and Equation (3.273):

$$
\begin{aligned}
&\mathbb{E}_{W^\rho,V^\rho} f^2_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\\
&\le \mathbb{E}_{W^\rho,V^\rho}\Big(\kappa_2\sqrt{m}_3\|x_i'^{(2)}\| + C_2\|x_i'^{(2)}\| + \frac{1}{\sqrt{m}_2}\sum_j |V_j^\rho x_i'^{(2)}|\Big)^2\\
&\lesssim (\kappa_2\sqrt{m}_3 + C_2)^2\mathbb{E}_{W^\rho}\|x_i'^{(2)}\|^2 + E_{W^\rho}E_{V^\rho}\frac{1}{m_2}\Big(\sum_j |V_j^\rho x_i'^{(2)}|\Big)^2\\
&\lesssim (\kappa_2\sqrt{m}_3 + C_2)^2\mathbb{E}_{W^\rho}\|x_i'^{(2)}\|^2 + E_{W^\rho}\frac{1}{m_2}\sum_j E_{V_j^\rho}|V_j^\rho x_i'^{(2)}|^2 + E_{W^\rho}\frac{1}{m_2}\sum_{j_1\neq j_2} E_{V_{j_1}^\rho}|V_{j_1}^\rho x_i'^{(2)}|E_{V_{j_2}^\rho}|V_{j_2}^\rho x_i'^{(2)}|\\
&\lesssim (\kappa_2\sqrt{m}_3 + C_2)^2\mathbb{E}_{W^\rho}\|x_i'^{(2)}\|^2 + E_{W^\rho}\frac{\beta_2^2}{m_2}\|x_i'^{(2)}\|^2 + E_{W^\rho}\beta_2^2\|x_i'^{(2)}\|^2\frac{m(m-1)}{m^2}\\
&\lesssim \Big((\kappa_2\sqrt{m}_3 + C_2)^2 + \beta_2^2\Big)\mathbb{E}_{W^\rho}\|x_i'^{(2)}\|^2\\
&\lesssim \Big((\kappa_2\sqrt{m}_3 + C_2)^2 + \beta_2^2\Big)E_{W^\rho}m_3\Big(\kappa_1 + C_1 + \frac{1}{\sqrt{m}_1}\sum_j |W_j^\rho x_i|\Big)^2\\
&\lesssim \Big((\kappa_2\sqrt{m}_3 + C_2)^2 + \beta_2^2\Big)m_3\Big[(\kappa_1 + C_1)^2 + \frac{1}{m_1}\sum_j \mathbb{E}_{W^\rho}|W_j^\rho x_i|^2 + \frac{1}{m_1}\sum_{j_1\neq j_2}\mathbb{E}_{W_{j_1}^\rho}|W_{j_1}^\rho x_i|\mathbb{E}_{W_{j_1}^\rho}|W_{j_2}^\rho x_i|\Big]\\
&\lesssim \Big((\kappa_2\sqrt{m}_3 + C_2)^2 + \beta_2^2\Big)m_3\Big[(\kappa_1 + C_1)^2 + \frac{\beta_1^2}{m_1} + \beta_1^2\frac{m(m-1)}{m^2}\Big]\\
&= m_3\Big((\kappa_2\sqrt{m}_3 + C_2)^2 + \beta_2^2\Big)\Big[(\kappa_1 + C_1)^2 + \beta_1^2\Big]. \qquad (3.278)
\end{aligned}
$$

Now for the second derivative, we can proceed by applying Cauchy-Swartz:

$$
\begin{aligned}
&\frac{d^2}{d\lambda^2} f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0} \\
&\leq \mathbb{E}_{W^\rho,V^\rho}\Big|\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle + \frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)^2 - \Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2 + \frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)\Big|\Big|f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\Big| \\
&\leq \mathbb{E}_{W^\rho,V^\rho}\Big|\frac{m_1^2}{\beta_1^4}\langle\tilde{W},W^\rho\rangle^2 - \frac{m_1}{\beta_1^2}\|\tilde{W}\|^2 + \frac{m_2^2}{\beta_2^4}\langle\tilde{V},V^\rho\rangle^2 - \frac{m_2}{\beta_2^2}\|\tilde{V}\|^2 + \frac{2m_1m_2}{\beta_1^2\beta_2^2}\langle\tilde{W},W^\rho\rangle\langle\tilde{V},V^\rho\rangle\Big| \\
&\Big|f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\Big| \\
&\leq \sqrt{\mathbb{E}_{W^\rho,V^\rho}\Big|\frac{m_1^2}{\beta_1^4}\langle\tilde{W},W^\rho\rangle^2 - \frac{m_1}{\beta_1^2}\|\tilde{W}\|^2 + \frac{m_2^2}{\beta_2^4}\langle\tilde{V},V^\rho\rangle^2 - \frac{m_2}{\beta_2^2}\|\tilde{V}\|^2 + \frac{2m_1m_2}{\beta_1^2\beta_2^2}\langle\tilde{W},W^\rho\rangle\langle\tilde{V},V^\rho\rangle\Big|^2} \\
&\sqrt{\mathbb{E}_{W^\rho,V^\rho}\Big|f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\Big|^2}.
\end{aligned}
$$

Now note that the cross terms have expectation zero, so we get

$$
\begin{aligned}
&=\sqrt{\mathbb{E}_{W^\rho,V^\rho}\Big(\frac{m_1^2}{\beta_1^4}\langle\tilde{W},W^\rho\rangle^2 - \frac{m_1}{\beta_1^2}\|\tilde{W}\|^2\Big)^2 + \Big(\frac{m_2^2}{\beta_2^4}\langle\tilde{V},V^\rho\rangle^2 - \frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)^2 + \frac{4m_1^2m_2^2}{\beta_1^4\beta_2^4}\mathbb{E}_{W^\rho,V^\rho}\langle\tilde{W},W^\rho\rangle^2\langle\tilde{V},V^\rho\rangle^2} \\
&\sqrt{\mathbb{E}_{W^\rho,V^\rho}\Big|f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\Big|^2} \\
&\lesssim \sqrt{\frac{m_1^2}{\beta_1^4}\|\tilde{W}\|^4 + \frac{m_2^2}{\beta_2^4}\|\tilde{V}\|^4 + \frac{4m_1m_2}{\beta_1^2\beta_2^2}\|\tilde{W}\|^2\|\tilde{V}\|^2}\sqrt{\mathbb{E}_{W^\rho,V^\rho}\Big|f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\Big|^2} \\
&\lesssim \Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2 + \frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)\sqrt{\mathbb{E}_{W^\rho,V^\rho}\Big|f_{W'+W^\rho,V'+V^\rho(x_i)}(x_i)\Big|^2}.
\end{aligned}
$$

Now applying Cauchy-shwartz and Equation (3.278) to above and combining it with Equations (3.270):

$$
\begin{aligned}
&\frac{d^2}{d\lambda^2} f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0} \\
&\lesssim \Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2 + \frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)\sqrt{m_3\Big((\kappa_2\sqrt{m}_3 + C_2)^2 + \beta_2^2\Big)\Big[(\kappa_1 + C_1)^2 + \beta_1^2\Big]}.
\end{aligned}
$$

Similarly for the third derivative:

$$\begin{aligned}&\Big|\frac{d^3}{d\lambda^3}f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0}\Big|\\&=\mathbb{E}_{W^\rho,V^\rho}\Big|\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)^3-3\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)\Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)\Big|\\&\Big|f_{W'+W^\rho,V'+V^\rho}(x_i)\Big|\\&\leq\sqrt{\mathbb{E}_{W^\rho,V^\rho}\Big[\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)^3-3\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)\Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)\Big]^2}\\&\sqrt{\mathbb{E}_{W^\rho,V^\rho}\Big|f_{W'+W^\rho,V'+V^\rho}(x_i)\Big|^2}.\end{aligned}\tag{3.279}$$

But note that

$$\begin{aligned}&\mathbb{E}_{W^\rho,V^\rho}\Big[\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)^3-3\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)\Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)\Big]^2\\&\leq 2\mathbb{E}_{W^\rho,V^\rho}\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)^6\\&+18\Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)^2\mathbb{E}_{W^\rho,V^\rho}\Big(\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle\Big)^2\end{aligned}$$

Now note that $\frac{m_1}{\beta_1^2}\langle\tilde{W},W^\rho\rangle+\frac{m_2}{\beta_2^2}\langle\tilde{V},V^\rho\rangle$ is a normal variable with variance $\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2$. Therefore, by the bound on the moments of normal random variables:

$$LHS\lesssim\Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)^3.\tag{3.280}$$

Plugging this into Equation (3.279) and also using Equation (3.278):

$$\begin{aligned}&\Big|\frac{d^3}{d\lambda^3}f'_{W'+\lambda\tilde{W},V'+\lambda\tilde{V}}(x_i)\Big|_{\lambda=0}\Big|\\&\lesssim\Big(\frac{m_1}{\beta_1^2}\|\tilde{W}\|^2+\frac{m_2}{\beta_2^2}\|\tilde{V}\|^2\Big)^{3/2}\sqrt{m_3\Big((\kappa_2\sqrt{m_3}+C_2)^2+\beta_2^2\Big)\Big[(\kappa_1+C_1)^2+\beta_1^2\Big]}.\end{aligned}$$

### 3.7.2 Representation Lemmas

In this section, we prove lemmas mostly related to the representation power of the network, which we mainly use in Section 3.6.12.

#### Representation Toolbox

**Lemma 46.** *Recall the definitions of $W_k^+$, and $\bar{x}_{i,k}$ from Equations (3.65), and (3.61). For all $k, i \in [n]$:*

$$|\bar{x}_{i,k} - \mathit{trace}(W_k^+ Z_k^i)| \leq \sqrt{n/(m_1 \lambda_0)} \|\mathcal{V}_k\|_{H^\infty}.$$

**Proof of Lemma 46**

$$\operatorname{trace}(W_k^+ Z_k^i) = \operatorname{trace}(Z_k^i \sum_{j=1}^{n} \mathcal{V}_{k,j} Z_k^j) = \sum_{i'=1}^{n} \mathcal{V}_{k,i'} \langle Z_k^i, Z_k^{i'} \rangle. \tag{3.281}$$

But note

$$\begin{aligned} \langle Z_k^i, Z_k^{i'} \rangle &= 1/m_1 \sum_{j=1}^{m_1} {W'_{k,j}}^2 \mathbb{1}\{W_j^{(0)T} x_i\} \mathbb{1}\{W_j^{(0)T} x_{i'}\} \langle x_{i'}, x_i \rangle \\ &= \frac{\langle x_{i'}, x_i \rangle}{m_1} \sum_{j=1}^{m_1} \mathbb{1}\{W_j^{(0)T} x_i\} \mathbb{1}\{W_j^{(0)T} x_{i'}\}. \end{aligned}$$

Now note that $\langle x_{i'}, x_i \rangle \leq 1$, and $\mathbb{1}\{W_j^{(0)T} x_i\} \mathbb{1}\{W_j^{(0)T} x_{i'}\}$ is a Bernoulli with

$$\mathbb{E} \mathbb{1}\{W_j^{(0)T} x_i\} \mathbb{1}\{W_j^{(0)T} x_{i'}\} = 1/4 + \arcsin(\langle x, y \rangle)/2\pi.$$

Therefore, by Hoeffding inequality we get

$$|\langle Z_k^i, Z_k^{i'} \rangle - H_{i,i'}^\infty| = O(1/\sqrt{m_1}).$$

Hence, because obviously $\|H^\infty\|_2 \leq 1$, we get

$$\operatorname{trace}(W_k^+ Z_k^i) = \sum_{i'=1}^n \mathcal{V}_{k,i'} H^\infty_{i,i'} + O(1/\sqrt{m}_1) \sum_{i'=1}^n \mathcal{V}_{k,i'} = \bar{x}_{i,k} + O(1/\sqrt{m}_1) \sum_{i'=1}^n \mathcal{V}_{k,i'},$$

which implies

$$\begin{aligned} |\operatorname{trace}(W_k^+ Z_k^i) - \bar{x}_{i,k}| &\leq O(1/\sqrt{m}_1)\sqrt{n}\|\mathcal{V}_k\|_2 \\ &\lesssim O(\sqrt{n}/(\sqrt{m}_1\sqrt{\lambda}_0))\|\mathcal{V}_k\|_{H^\infty} \\ &\leq \sqrt{n/(m_1\lambda_0)}\|\mathcal{V}_k\|_{H^\infty}. \end{aligned}$$

**Lemma 47.** *(Bounding the rows norm) For every $1 \leq j \leq m_1$, we have*

$$\|W_j^+\| \leq \sqrt{nm_3}/(\sqrt{m_1\lambda_0})\sqrt{\sum_k \|\mathcal{V}_k\|^2_{H^\infty}}.$$

*Furthermore, for every $k \in [m_3]$, we have*

$$\|W_j^{k+}\| \leq \sqrt{n}/\sqrt{\lambda_0 m_1}\|\mathcal{V}_k\|_{H^\infty}. \tag{3.282}$$

For the ease of notation, because here we want to work with row sub indices of the matrix $W_k^+$, we refer to it by $W^{k+}$. **Proof of Lemma 47**

For a fixed $1 \leq j \leq m_1$ we have with high probability over the randomness of the sign matrix $W^s$:

$$\begin{aligned}
\|W_j^+\| &= \|\sum_{k=1}^{m_3} W_j^{k+}\| = \|\sum_{k=1}^{m_3} W_{k,j}^s 1/\sqrt{m_1}\sum_{i=1}^{n} \mathcal{V}_{k,i} x_i \mathbb{1}\{W_j^{(0)} x_i \geq 0\}\| \\
&\leq \sqrt{m_3}/\sqrt{m_1}\sqrt{\sum_{k=1}^{m_3}\|\sum_{i=1}^{n}\mathcal{V}_{k,i} x_i \mathbb{1}\{W_j^{(0)} x_i \geq 0\}\|^2} \\
&\leq \sqrt{m_3}/\sqrt{m_1}\sqrt{\sum_k \Big(\sum_i |\mathcal{V}_{k,i}|\Big)^2} \\
&\leq \sqrt{m_3}/\sqrt{m_1}\sqrt{n\sum_k \|\mathcal{V}_k\|^2} \\
&\leq \sqrt{m_3 n}/(\sqrt{m_1\lambda_0})\sqrt{\sum_k \|\mathcal{V}_k\|_{H^\infty}^2}.
\end{aligned}$$

Furthermore, for every $k \in [m_3]$, we have

$$\|W_j^{k+}\| \leq 1/\sqrt{m_1}\sum_i |\mathcal{V}_{k,i}| \leq (\sqrt{n}/\sqrt{m_1})\|\mathcal{V}_k\|_2 \leq (\sqrt{n}/\sqrt{\lambda_0 m_1})\|\mathcal{V}_k\|_{H^\infty}.$$

**Lemma 48.** *With high probability over the initialization, we have*

$$\|W^{k+}\|_F^2 \leq (1 \pm O(n/(\lambda_0\sqrt{m_1})))\|\mathcal{V}_k\|_{H^\infty}^2.$$

**Proof of Lemma 48**

Recall from the definition of $W^{k+}$ in Equation (3.65):

$$\begin{aligned}
\|W^{k+}\|_F^2 &= \|\sum_{i=1}^{n}\mathcal{V}_{k,i} Z_k^i\|_F^2 = \sum_{i=1}^{n}\sum_{i'=1}^{n}\mathcal{V}_{k,i}\mathcal{V}_{k,i'}\langle Z_k^i, Z_k^{i'}\rangle \\
&= \sum_{i=1}^{n}\sum_{i'=1}^{n}\mathcal{V}_{k,i}\mathcal{V}_{k,i'}(H_{i,i'}^\infty \pm O(1/\sqrt{m_1})) \lesssim \mathcal{V}_k^T H^\infty \mathcal{V}_k \pm O((\|\mathcal{V}_k\|_1)^2/\sqrt{m_1}) \\
&= \|\mathcal{V}_k\|_{H^\infty}^2 \pm \|\mathcal{V}_k\|_{H^\infty}^2 O(n/(\lambda_0\sqrt{m_1})) = (1 \pm O(n/(\lambda_0\sqrt{m_1})))\|\mathcal{V}_k\|_{H^\infty}^2
\end{aligned}$$

**Some Linear Algebra**

**Lemma 49.** *For $n \le s$, let $r_1, \dots, r_n$ be $s$-dimensional vectors that are approximately normalized and orthogonal to one another, i.e. given some $\delta > 0$, for every $1 \le i \neq j \le n$:*

$$-\delta \le \langle r_i, r_j \rangle \le \delta,\ \|r_i\|^2 \le 1 + \delta.$$

*Then, for any vector $v$ we have*

$$\sum_{i=1}^{n} \langle v, r_i \rangle^2 \le (1 + \delta + n(n-1)\delta(1+\delta)^2)\|v\|^2.$$

**Proof of Lemma 49**

Define

$$v_1 = \sum_{i=1}^{n} \langle v, r_i \rangle r_i,\ v_2 = v - v_1.$$

First, note that

$$\begin{aligned}\sum_{i=1}^{n} \langle v, r_i \rangle^2 &\le (1+\delta) \sum_{i=1}^{n} \langle v, r_i \rangle^2 \|r_i\|^2 \\ &= (1+\delta)\|\sum_{i} \langle v, r_i \rangle r_i\|^2 - 2(1+\delta) \sum_{1 \le i \neq j \le n} \langle v, r_i \rangle \langle v, r_i \rangle \langle r_i, r_j \rangle \\ &= (1+\delta)\|v_1\|^2 - 2(1+\delta) \sum_{1 \le i \neq j \le n} \langle v, r_i \rangle \langle v, r_i \rangle \langle r_i, r_j \rangle.\end{aligned}$$

Next, we write

$$\begin{aligned}\langle v_1, v - v_1 \rangle &= \langle v, \sum_{i=1}^{n} \langle v, r_i \rangle r_i \rangle - \langle \sum_{i=1}^{n} \langle v, r_i \rangle r_i, \sum_{i=1}^{n} \langle v, r_i \rangle r_i \rangle \\ &= \sum_{i} \langle v, r_i \rangle^2 - \sum_{i} \langle v, r_i \rangle^2 - 2 \sum_{i \neq j} \langle v, r_i \rangle \langle v, r_j \rangle \langle r_i, r_j \rangle \\ &= -2 \sum_{i \neq j} \langle v, r_i \rangle \langle v, r_j \rangle \langle r_i, r_j \rangle.\end{aligned}$$

Therefore

$$\begin{aligned}\sum_{i=1}^{n} \langle v, r_i \rangle^2 &\leq (1+\delta)(\|v_1\|^2 + \|v - v_1\|^2) - 2(1+\delta) \sum_{1 \leq i \neq j \leq n} \langle v, r_i \rangle \langle v, r_i \rangle \langle r_i, r_j \rangle. \\ &\leq (1+\delta)(\|v_1\|^2 + \|v - v_1\|^2 + 2\langle v_1, v - v_1 \rangle) \\ &+ 2(1+\delta) \sum_{1 \leq i \neq j \leq n} \langle v, r_i \rangle \langle v, r_i \rangle \langle r_i, r_j \rangle \\ &\leq (1+\delta)\|v\|^2 + 2(1+\delta) \sum_{1 \leq i \neq j \leq n} \|v\|^2 \|r_i\| \|r_j\| \delta \\ &\leq (1+\delta)\|v\|^2 + 2(1+\delta) \sum_{1 \leq i \neq j \leq n} \|v\|^2 (1+\delta)\delta \\ &= (1+\delta)\|v\|^2 + n(n-1)\delta(1+\delta)^2 \|v\|^2 \\ &= (1 + \delta + n(n-1)\delta(1+\delta)^2)\|v\|^2,\end{aligned}$$

which completes the proof.

In the following lemma, we state a trivial bound on the norm of $\bar{x}_i$ based on $\zeta_1$.

### Bound on the norm of $\bar{x}_i$'s

**Lemma 50.** *For every $i \in [n]$, we have*

$$\|\bar{x}_i\| \leq \sqrt{\sum_k \|\mathcal{V}_k\|^2_{H^\infty}} = \sqrt{\zeta_1}.$$

**Proof of Lemma 50**

By definition:

$$\bar{x}_i^T = \left(H^\infty_{i,}\mathcal{V}_k\right)_{k=1}^{m_3}.$$

Now consider the Cholskey factorization $H^\infty = KK^T$. Because of the assumption $\|x_i\| = 1$, we know that the diagonal of $H^\infty$ is all $1/2$. Hence, for the $i$th row of $K$ we have $\|K_i\| = 1/2$. Now by Cauchy-Swartz, we have

$$x_{i_1}^2 = \langle \sum_i \mathcal{V}_{k,i} K_i, K_{i_1} \rangle^2 \leq \| \sum_i \mathcal{V}_{k,i} K_i \|^2 \|K_{i_1}\|^2 = 1/2 \|\mathcal{V}_k\|^2_{H^\infty}.$$

Summing over $i$ and noting Equation (3.60) completes the proof.

**Lemma 51.** *In the context of Lemma 19, for $\zeta_2 \leq 2nB^2$, one can substitute $f^*$ by $\bar{f}^*$ such that*

$$R_n(\bar{f}^*) \leq 2R_n(f^*) + \frac{B^2}{n},$$
$$\bar{f}^{*T} A^{-1} \bar{f}^* \leq f^{*T} A^{-1} f^*,$$

*and furthermore, $\bar{f}^*$ is in the subspace of eigenvectors of $A$ with eigenvaue larger than $\Omega(\frac{1}{n^2})$. Moreover, the constant 2 is arbitrary and can be changed to any constant more than one, with the cost of an additional constant behind the second term.*

### Proof of Lemma 51

For an arbitrary $i \in [n]$ and some given vector $\bar{f}^*$ (we will specify later), we define

$$\delta = |f_i^* - \bar{f}_i^*|,$$

and suppose the slope of $\ell(., y_i)$ at point $f_i^*$ is equal to $c$. Then, using the convexity, the fact that $\ell(y_i, y_i) = 0$, and the 1-smoothness of $\ell(., y_i)$, it is not hard to see the following poincare inequality between the value and derivative of $\ell(., y_i)$ at point $f_i^*$:

$$c \leq \sqrt{2\ell(f_i^*, y_i)} \coloneqq 2\ell. \tag{3.283}$$

where from now on, for brevity, we refer to $\ell(f_i^*, y_i)$ by $\ell$. Also, from the definition of $\delta$ and again using 1 smoothness property, it is easy to see that

$$\ell(\bar{f}_i^*, y_i) \leq (c + \delta)\delta + \ell(f_i^*, y_i) = (c + \delta)\delta + \ell, \tag{3.284}$$

Plugging Equation (3.283) into (3.284) and using AM-GM inequality:

$$\begin{aligned}\ell(\bar{f}_i^*, y_i) &\leq \delta^2 + c\delta + \ell \leq \delta^2 + \sqrt{2\ell}\delta + \ell \\ &\leq \delta^2 + \ell + \delta^2/2 + \ell \\ &\leq 2\ell + 3\delta^2/2.\end{aligned}$$

Summing above for $i \in [n]$, we obtain

$$R_n(\bar{f}^*) \leq 2R_n(f^*) + 3\|f^* - \bar{f}^*\|_2^2/2. \tag{3.285}$$

Now we write an eigendecomposition for $A$ as $A = \sum_{i=1}^n \lambda_i u_i u_i^T$ for orthonormal basis $\{u_i\}$, and let $f^* = \sum_i \gamma_i u_i$ be the representation of $f^*$ in this basis. Then, from our assumption, for arbitrary $\omega > 0$

$$\sum_i \gamma_i^2 \lambda_i^{-1} = f^{*T} A^{-1} f^* \leq 4nB^2,$$

which implies

$$\omega^{-1} \sum_{i:\ \lambda_i \leq \omega} \gamma_i^2 \leq 4nB^2,$$

or equivalently

$$\sum_{i:\ \lambda_i \leq \omega} \gamma_i^2 \leq 4nB^2\omega, \tag{3.286}$$

where notice that $\sum_{i:\ \lambda_i \leq \omega} \gamma_i^2$ is the squared norm of the projection of $f^*$ onto the directions whose eigenvalue is at most $\omega$. Now taking $\omega = \frac{1}{12n^2}$ and defining $\bar{f}^*$ by keeping only the directions in the expansion of $f^*$ in the eigenbasis of $A$, for which $\lambda_i > \omega$, completes the proof.

### 3.7.3 Coupling for $\hat{\nabla}_W, \hat{\nabla}_V$

In general, because the gaussian smoothing matrices $(W^\rho, V^\rho)$ can become unbounded, the gradient estimates $(\hat{\nabla}_W, \hat{\nabla}_V) = \nabla_{W,V}\ell(f_{W'+W^\rho,V'+V^\rho}(x_i), y_i)$ also become unbounded. However, in analyzing the stochastic behavior of SGD and showing that it can escape saddle points, it is convenient to assume the gradient's noise vector is almost surely bounded. The goal of this section is to introduce a coupling between $(W^\rho, V^\rho)$ and another random variable that is a.s. bounded polynomially in other parameters. As that the coupled random variables take different values is exponentially small while the number of iterations in our algorithm is only polynomially large, without any concern we instead work with this new random varaible, and with an overload of notation we also denote it by $(W^\rho, V^\rho)$.

**Lemma 52.** *For an arbitrary parameter $\chi >> 1$, On any pair for $(W', V')$ with $\|W'\| \leq C_1$, $\|V'\| \leq C_2$, there exist a mean zero random vector $\bar{\Lambda}$ with respect to the randomness of the uniformly picked data point $(x_i, y_i)$ and the smoothing matrices $W^{\rho,1}$, $W^{\rho,2}$, $V^{\rho,1}$, and $V^{\rho,2}$ which define $\hat{\nabla}_{W',V'}$ (meaning it is a function of those variables), such that with probability at least*

$$1 - 2\exp\{-(\chi^2 - 1)dm_1/4\} - 2\exp\{-(\chi^2 - 1)m_3m_2/4\} := 1 - \delta_1,$$

*we have*

$$\hat{\nabla}_{W',V'} = \nabla_{W',V'}L(W', V') + \bar{\Lambda}, \tag{3.287}$$

*and finally $\bar{\Lambda}$ is a.s. polynomially bounded, i.e. almost surely we have*

$$\|\bar{\Lambda}\| \leq poly(m_1, m_2, m_3, C_1, C_2, B, \chi).$$

**Proof of Lemma 52**

Remember that $x_i'$ was the output of the first layer (by considering the smoothing matrix $W^\rho$). Now with high probability over the initialization,

$$\begin{aligned}
&\|\nabla_{W'} f_{W'+W^\rho,V'+V^\rho}(x_i)\|_F = \|\nabla_{x_i'} f_{W'+W^\rho,V'+V^\rho}(x_i)^T \frac{D(x_i')}{dW'}\| \\
&\|\nabla_{x_i'} f_{W'+W^\rho,V'+V^\rho}(x_i)\|\|\frac{D(x_i')}{dW'}\| \\
&\|\frac{1}{\sqrt{m_2}} a^T D_{V'+V^\rho}(V^{(0)} + V' + V^\rho)\|\|\frac{1}{\sqrt{m_1}} \sum_{k=1}^{m_3} \mathrm{diag}(W_k^s) D_{W'+W^\rho,x_i} x_i^T\|_F \\
&\leq (\|V^{(0)}\|_F + \|V'\|_F + \|V^\rho\|_F)\Big(\frac{1}{\sqrt{m_1}} \sum_k \|\mathrm{diag}(W_k^s) D_{W'+W^\rho,x_i} x_i^T\|_F\Big) \\
&\leq (\kappa_2\sqrt{m_2 m_3} + C_2 + \|V^\rho\|_F)\Big(\frac{1}{\sqrt{m_1}} \sum_k \|\mathrm{diag}(W_k^s) D_{W'+W^\rho,x_i} x_i^T\|_F\Big) \\
&\leq (\kappa_2\sqrt{m_2 m_3} + C_2 + \|V^\rho\|_F). \qquad (3.288)
\end{aligned}$$

On the other hand, using the final bound in Lemma 42:

$$\begin{aligned}
&\|\nabla_{V'} f_{W'+W^\rho,V'+V^\rho}(x_i)\|_F = \|\frac{1}{\sqrt{m_2}} \mathrm{diag}(a) D_{V'+V^\rho} x_i'^T\|_F \leq \|x_i'\| \\
&\leq \kappa_1\sqrt{m_3} + C_1 + \sqrt{m_3} \sum_j \frac{1}{\sqrt{m_1}} |W_j^\rho x_i| \\
&\leq \kappa_1\sqrt{m_3} + C_1 + \sqrt{m_3}\|W^\rho\|_F. \qquad (3.289)
\end{aligned}$$

Denoting $\dot{\ell}(f_{W'+W^{\rho,1},V'+V^{\rho,1}}, y_i)\nabla_{W',V'}\ell(f_{W'+W^{\rho,2},V'+V^{\rho,2}}(x_i), y_i)$ by $\tilde{\nabla}_{W',V'}$, then combining Equations (3.288) and (3.289) and using the 1 gradient lipschitzness property of the square loss,

$$\begin{aligned}
&\|\tilde{\nabla}_{W',V'}\|_F \\
&= |\frac{d(\ell(f, y_i))}{df}| \sqrt{\|\nabla_{W'} f_{W'+W^{\rho,2},V'+V^{\rho,2}}(x_i)\|_F^2 + \|\nabla_{V'} f_{W'+W^{\rho,2},V'+V^{\rho,2}}(x_i)\|_F^2} \\
&\leq \Big(|f_{W'+W^{\rho 1},V'+V^{\rho 1}}(x_i)| + |B|\Big)\Big(\kappa_1\sqrt{m_3} + C_1 + \sqrt{m_3}\|W^{\rho,2}\|_F + \kappa_2\sqrt{m_2 m_3} + C_2 + \|V^{\rho,2}\|_F\Big). \\
&\qquad (3.290)
\end{aligned}$$

Finally, applying Cauchy-Swartz to the second a.s. bound in Lemma 43 we have:

$$\begin{aligned}
&\Big|f_{W'+W^\rho,V'+V^\rho}(x_i)\Big| \\
&\le (\kappa_2\sqrt{m}_3 + \|V^\rho\|_F)\Big(\sqrt{m}_3\kappa_1 + C_1 + \sqrt{m_3}\|W^\rho\|\Big) + C_2(C_1 + \sqrt{m_3}\|W^\rho\|).
\end{aligned}$$

Combining this with (3.290):

$$\begin{aligned}
&\|\tilde{\nabla}_{W',V'}\|_F \\
&\le \Big[B + (\kappa_2\sqrt{m}_3 + \|V^{\rho,1}\|_F)\Big(\sqrt{m}_3\kappa_1 + C_1 + \sqrt{m_3}\|W^{\rho,1}\|\Big) + C_2(C_1 + \sqrt{m_3}\|W^{\rho,1}\|)\Big] \\
&\times \Big(\kappa_1\sqrt{m}_3 + C_1 + \sqrt{m_3}\|W^{\rho,2}\|_F + \kappa_2\sqrt{m_2 m_3} + C_2 + \|V^{\rho,2}\|_F\Big).
\end{aligned}$$

Therefore, using the Lipschitz bound in Theorem 9:

$$\begin{aligned}
&\|\tilde{\nabla}_{W',V'} - \nabla_{W',V'}\mathbb{E}_{(x_i,y_i)\sim\mathcal{Z}}\ell(f'_{W',V'}(x_i), y_i)\|_F \\
&\le \|\hat{\nabla}_{W',V'}\|_F + \|\mathbb{E}_{(x_i,y_i)\sim\mathcal{Z}}\nabla_{W',V'}\ell(f_{W'+W^\rho,V'+V^\rho}(x_i), y_i)\|_F \\
&\le \Big[B + (\kappa_2\sqrt{m}_3 + \|V^{\rho,1}\|_F)\Big(\sqrt{m}_3\kappa_1 + C_1 + \sqrt{m_3}\|W^{\rho,1}\|\Big) + C_2(C_1 + \sqrt{m_3}\|W^{\rho,1}\|)\Big] \\
&\times \Big(\kappa_1\sqrt{m}_3 + C_1 + \sqrt{m_3}\|W^{\rho,2}\|_F + \kappa_2\sqrt{m_2 m_3} + C_2 + \|V^{\rho,2}\|_F\Big) + (O(C_1C_2) + B)\Psi_1.
\end{aligned} \tag{3.291}$$

Now we define the following events

$$\begin{aligned}
\Xi_1 &:= \{\|W^{\rho_1}\|_F \ge \chi\sqrt{d}\beta_1 \vee \|W^{\rho_2}\|_F \ge \chi\sqrt{d}\beta_1\}, \\
\Xi_2 &:= \{\|V^{\rho_1}\|_F \ge \chi\sqrt{m}_3\beta_2 \vee \|V^{\rho_2}\|_F \ge \chi\sqrt{m}_3\beta_2\},
\end{aligned}$$

where recall we assume $\chi >> 1$. Then, as we know the variable $\|W^\rho\|_F^2$ has mean $d\beta_1^2$ and is subexponential with parameters $(d\beta_1^4/m_1, \beta_1^2/m_1)$. Hence, by a union bound and Bernstein (Note that $W^{\rho,1}, W^{\rho,2}$ are independent):

$$
\begin{aligned}
&\mathbb{P}(\Xi_1)\\
&\le 2\mathbb{P}(\|W^\rho\|_F \ge \chi\beta_1\sqrt{d})\\
&= 2\mathbb{P}(\|W^\rho\|_F^2 \ge \chi^2\beta_1^2 d)\\
&\le 4\max\Big(\exp\{-(\chi^2-1)^2\beta_1^4 d^2/(4d\beta_1^4/m_1)\},\ \exp\{-(\chi^2-1)\beta_1^2 d/(4\beta_1^2/m_1)\}\Big)\\
&= 4\max\Big(\exp\{-(\chi^2-1)^2 dm_1/4\},\ 2\exp\{-(\chi^2-1)dm_1/4\}\Big) \le 2\exp\{-(\chi^2-1)dm_1/4\}.
\end{aligned}
$$

Similarly for $\|V^\rho\|_F$:

$$\mathbb{P}(\Xi_2) = \mathbb{P}(\|V^\rho\|_F \ge \chi\beta_2\sqrt{m_3}) \le 4\exp\{-(\chi^2-1)m_3m_2/4\}.$$

Moreover, because of the subexponential tails of $\|W^\rho\|_F^2$ and $\|V^\rho\|_F^2$, for each of $W^{\rho,1}$ or $W^{\rho_2}$, $V^{\rho_1}$, or $V^{\rho_2}$:

$$
\begin{aligned}
&\mathbb{E}(\|W^\rho\|_F|\ \Xi_1) \lesssim \chi\sqrt{d}\beta_1,\ \mathbb{E}(\|W^\rho\|_F^2|\ \Xi_1) \lesssim \chi^2 d\beta_1^2.\\
&\mathbb{E}(\|V^\rho\|_F|\ \Xi_2) \lesssim \chi\sqrt{m_3}\beta_2,\ \mathbb{E}(\|V^\rho\|_F^2|\ \Xi_2) \lesssim \chi^2 m_3\beta_2^2.
\end{aligned}
$$

Now Defining $\Xi = \Xi_1 \cup \Xi_2$ and combining the above equations:

$$
\begin{aligned}
&\mathbb{E}(\|W^\rho\|\mathbb{1}\{\Xi\}) \le \mathbb{E}\|W^\rho\|(\mathbb{1}\{\Xi_1\} + \mathbb{1}\{\Xi_2\}) = \mathbb{E}(\|W^\rho\||\ \Xi_1)\mathbb{P}(\Xi_1) + \mathbb{E}(\|W^\rho\|)\mathbb{P}(\Xi_2)\\
&\lesssim \chi\sqrt{d}\beta_1 2\exp\{-(\chi^2-1)dm_1/4\} + \sqrt{d}\beta_1 2\exp\{-(\chi^2-1)m_3m_2/4\}\\
&= 2\sqrt{d}\beta_1(\chi\exp\{-(\chi^2-1)dm_1/4\} + \exp\{-(\chi^2-1)m_3m_2/4\}),
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\|W^\rho\|^2\mathbb{1}\{\Xi_1\} &= \mathbb{E}(\|W^\rho\|^2|\ \Xi_1)\mathbb{P}(\Xi_1)\\
&\le 2d\beta_1^2\chi^2\exp\{-(\chi^2-1)dm_1/4\}.
\end{aligned}
$$

Similarly

$$\mathbb{E}(\|V^\rho\|\mathbb{1}\{\Xi\}) \leq 2\sqrt{m}_3\beta_2(\exp\{-(\chi^2-1)dm_1/4\} + \chi\exp\{-(\chi^2-1)m_3m_2/4\}),$$

and

$$\mathbb{E}(\|V^\rho\|^2\mathbb{1}\{\Xi_2\}) \leq 2m_3\beta_2^2\chi^2\exp\{-(\chi^2-1)m_3m_2/4\}.$$

Applying these equations to (3.291) with Cauchy-Schwartz to get the upper bounds $\mathbb{E}\mathbb{1}\{\Xi_1\}\|W^{\rho,1}\|\|W^{\rho,2}\| \leq \mathbb{E}\mathbb{1}\{\Xi_1\}\|W^\rho\|^2$ and $(\mathbb{E}_{W^\rho}\|W^\rho\|)^2 \leq \mathbb{E}_{W^\rho}\|W^\rho\|^2$ (for terms with only one $W^{\rho,i}$ or $V^{\rho,i}$, we simply write them as $W^\rho$ and $V^\rho$):

$$\begin{aligned}
&\mathbb{E}_{W^\rho,V^\rho}\|\tilde{\nabla}_{W',V'} - \nabla_{W',V'}\mathbb{E}_{(x_i,y_i)\sim\mathcal{Z}}\ell(f'_{W',V'}(x_i),y_i)\|_F\mathbb{1}\{\Xi\} \\
&\leq \mathbb{E}_{W^\rho,V^\rho,(x_i,y_i)}\Big[B + (\kappa_2\sqrt{m}_3 + \|V^{\rho,1}\|_F)\Big(\sqrt{m}_3\kappa_1 + C_1 + \sqrt{m_3}\|W^\rho\|\Big) + C_2(C_1 + \sqrt{m_3}\|W^\rho\|)\Big] \\
&\times\Big(\kappa_1\sqrt{m}_3 + C_1 + \sqrt{m}_3\|W^\rho\|_F + \kappa_2\sqrt{m_2m_3} + C_2 + \|V^\rho\|_F\Big) + \alpha(O(C_1C_2) + B)\Psi_1 \\
&= \Big[B + (\kappa_2\sqrt{m}_3)(\sqrt{m}_3\kappa_1 + C_1) + C_1C_2\Big] \\
&\times\Big(\kappa_1\sqrt{m}_3 + C_1 + \kappa_2\sqrt{m_2m_3} + C_2 + \sqrt{m}_3\mathbb{E}_{W^\rho}\mathbb{1}\{\Xi\}\|W^\rho\|_F + \mathbb{E}_{V^\rho}\mathbb{1}\{\Xi\}\|V^\rho\|_F\Big) \\
&+ \Big(B + (\sqrt{m}_3\kappa_1 + C_1)\sqrt{m}_3 + (\kappa_2m_3 + C_2\sqrt{m}_3) + \sqrt{m}_3(\kappa_1\sqrt{m}_3 + C_1 + \kappa_2\sqrt{m_2m_3} + C_2)\Big) \\
&\times(\mathbb{E}_{W^\rho}\mathbb{1}\{\Xi_1\}\|W^\rho\|_F\mathbb{E}_{V^\rho}\|V^\rho\|_F + \mathbb{E}_{W^\rho}\|W^\rho\|_F\mathbb{E}_{V^\rho}\mathbb{1}\{\Xi_2\}\|V^\rho\|_F) \\
&+ (B + \sqrt{m}_3\kappa_1 + C_1)(E_{V^\rho}\mathbb{1}\{\Xi_2\}\|V^\rho\|^2 + \mathbb{P}(\Xi_1)\mathbb{E}\|V^\rho\|^2) \\
&+ C_2m_3(\mathbb{E}_{W^\rho}\mathbb{1}\{\Xi_1\}\|W^\rho\|^2 + \mathbb{P}(\Xi_2)\mathbb{E}\|W^\rho\|^2) \\
&+ m_3(\mathbb{E}_{W^\rho}\mathbb{1}\{\Xi_1\}\|W^\rho\|^2\mathbb{E}_{V^\rho}\|V^\rho\| + \mathbb{E}_{W^\rho}\|W^\rho\|^2\mathbb{E}_{V^\rho}\mathbb{1}\{\Xi_2\}\|V^\rho\|) \\
&+ \sqrt{m}_3(\mathbb{E}_{V^\rho}\mathbb{1}\{\Xi_2\}\|V^\rho\|^2\mathbb{E}_{W^\rho}\|W^\rho\| + \mathbb{E}_{V^\rho}\|V^\rho\|^2\mathbb{E}_{W^\rho}\mathbb{1}\{\Xi_1\}\|W^\rho\|) \\
&+ (O(C_1C_2) + B)\Psi_1\mathbb{P}(\Xi) \\
&\leq (\exp\{-(\chi^2-1)dm_1/4\} + \chi\exp\{-(\chi^2-1)m_3m_2/4\})\text{poly}(m_1,m_2,m_3) = \text{negligible.}
\end{aligned} \tag{3.292}$$

But note that

$$\begin{aligned}
&\hat{\nabla}_{W',V'} = \tilde{\nabla}_{W',V'} + \nabla_{W',V'}(\psi_1\|W'\|^2 + \psi_2\|V'\|^2),\\
&\nabla_{W',V'}L(W',V') = \nabla_{W',V'}\mathbb{E}_{(x_i,y_i)\sim\mathcal{Z}}\ell(f'_{W',V'}(x_i),y_i) + \nabla_{W',V'}(\psi_1\|W'\|^2 + \psi_2\|V'\|^2).
\end{aligned}$$

Applying this to Equation (3.292), we get that if we define

$$\Lambda \coloneqq \hat{\nabla}_{W',V'} - \nabla_{W',V'}L(W',V'),$$

then

$$\mathbb{E}\|\Lambda\mathbb{1}\{\Xi\}\| \leq (\exp\{-(\chi^2-1)dm_1/4\} + \chi\exp\{-(\chi^2-1)m_3m_2/4\})\mathrm{poly}(m_1,m_2,m_3).$$

On the other hand, note that using again Equation (3.291), we have the following a.s. bound:

$$\|\Lambda\mathbb{1}\{\Xi^c\}\| = \mathrm{poly}(m_1,m_2,m_3,C_1,C_2,\chi).$$

Defining

$$\begin{aligned}
\Lambda_1 &= \Lambda\mathbb{1}\{\Xi^c\},\\
\Lambda_2 &= \mathbb{1}\{\Xi\}\mathbb{E}(\Lambda|\Xi),\\
\bar{\Lambda} &= \Lambda_1 + \Lambda_2,
\end{aligned}$$

we get that with probability at least $1-\mathbb{P}(\Xi)$:

$$\hat{\nabla}_{W',V'} = \nabla_{W',V'}L(W',V') + \bar{\Lambda}$$

and also note that

$$\mathbb{E}\bar{\Lambda} = \mathbb{E}\Lambda = 0.$$

Finally by the a.s. bound for $\Lambda_1$, we have a.s.:

$$\begin{aligned}&\|\bar{\Lambda}\| \leq \mathbb{E}_{W^\rho,V^\rho,(x_i,y_i)}\|\Lambda\mathbb{1}\{\Xi\}\| + \|\Lambda\mathbb{1}\{\Xi^c\}\| \\ &\leq (\exp\{-(\chi^2-1)dm_1/4\} + \chi\exp\{-(\chi^2-1)m_3m_2/4\})\text{poly}(m_1,m_2,m_3) + \text{poly}(m_1,m_2,m_3) \\ &= \text{poly}(m_1,m_2,m_3,C_1,C_2,B,\chi),\end{aligned}$$

which completes the proof.

**Corollary 9.1.** *It is easy to check that running* **PSGD** *with unbiased gradient estimate $\hat{\nabla}_{W',V'}$ is equivalent to running SGD after our change of coordinates, with unbiased gradient estimate $\hat{\nabla}_{w',v'} := \bar{\Upsilon}\nabla_{W',V'}$, where $\bar{\Upsilon}$ is the matrix for our change of coordinate, which is equal to $\Upsilon$ defined in Section 3.6.10 for the coordinates in $V'$ and simply identity for the coordinates in $W'$. Therefore, projecting both sides in Equation (3.293) of Lemma 52 onto $\Phi^\perp$ by multiplying $\Upsilon$ implies that with high probability for all iterations of the algorithm*

$$\begin{aligned}\hat{\nabla}_{w',v'} &= \bar{\Upsilon}\nabla_{W',V'}L^\Pi(W',V') + \bar{\Upsilon}\bar{\Lambda} \\ &= \nabla_{w',v'}L^\Pi(w',v') + \bar{\Upsilon}\bar{\Lambda}, \end{aligned} \tag{3.293}$$

*where $\bar{\pounds} := \bar{\Upsilon}\bar{\Lambda}$ (using the properties of $\bar{\Lambda}$ in Lemma 52) is a mean zero noise vector with almost surely bounded norm, i.e. $\|\bar{\pounds}\| \leq Q'$ for some $Q' = poly(m_1,m_2,m_3,C_1,C_2)$. (we dropped the $\chi$ parameters by considering constant high probability argument).*

*Finally, note that injecting noise $(\Xi_1/\|\Xi_1\|, \Xi_2/(\sqrt{m_1}\|\Xi_2\|))$ by* **PSGD** *results in adding an extra zero mean noise $(\tilde{\Xi}_1,\tilde{\Xi}_2) := (\bar{\Upsilon}\Xi_1/\|\Xi_1\|, \bar{\Upsilon}\Xi_2/(\sqrt{m_1}\|\Xi_2\|))$ to the gradient $\nabla_{w',v'}L^\Pi(w',v')$. Therefore, overall running SGD on $L^\Pi$ (which is equivalent to* **PSGD** *on L) observe an unbiased noise vector defined as $\pounds := \bar{\pounds} + (\tilde{\Xi}_1,\tilde{\Xi}_2)$. Now it is easy to check that the moment matrix of $\tilde{\Xi}_1$ and $\tilde{\Xi}_2$ are ${\sigma'_1}^2 I$ and ${\sigma'_2}^2 I$ for*

$${\sigma'_1}^2 := \frac{1}{m_1^2 d}, \tag{3.294}$$

$${\sigma'_2}^2 := \frac{m_2(m_3-n)}{{m_2^m}_3,} \tag{3.295}$$

*which implies the moment matrix of $\pounds$ is upper bounded by*

$$\sigma_2^2 I \coloneqq (Q'/(m_2 m_3 + m_1 d) + \max\{{\sigma'_1}^2, {\sigma'_2}^2\})I,$$

*and lower bounded by*

$$\sigma_2^2 I \coloneqq \min\{{\sigma'_1}^2, {\sigma'_2}^2\}I,$$

*i.e.*

$$\sigma_1^2 I \mathbb{E}\pounds\pounds^T \leq \sigma_2^2 I.$$

*(Note that we look at the new coordinates $(w', v')$ as a vectors, so the term $\mathbb{E}\pounds\pounds^T$ makes sense.)*

*Moreover, $\|\tilde{\Xi}_1\| = 1/\sqrt{m_1}, \|\tilde{\Xi}_2\| \leq 1$ almost surely, which implies the following almost surely bound for $\pounds$:*

$$\|\pounds\| \leq Q \coloneqq Q' + 1 + 1/\sqrt{m_1}.$$

**Lemma 53.** *Let $g(x)$ be a second order differentiable function over $\mathbb{R}^N$ such that at point $x$, there exist a random direction $y$ and deterministic direction $z$ and fixed positive real $r$ with:*

$$\begin{aligned}&\mathbb{E}y = 0,\\ &\mathbb{E}_y g(x + \eta z + \sqrt{\eta} y) \leq g(x) - \eta r.\end{aligned}$$

*Then, for the gradient and Hessian at point $x$, we have either*

$$\|\nabla g(x)\| \geq \frac{r}{4\|z\|},$$

*or*

$$\lambda_{min}\Big(\nabla^2 g(x)\Big) \leq -\frac{r}{2\|y\|^2}.$$

**Proof of Lemma 53**

We write the second order tailor approximation of $g$ around $x$:

$$g(x+w) = g(x) + \nabla g(x)^T w + \frac{1}{2} w^T \nabla^2 g(x) w + o(\|w\|^2).$$

Now substituting $w$ with $\eta z + \sqrt{\eta} y$ and taking expectation with respect to $y$, as we send $\eta \to 0$ and using the fact that $\mathbb{E}y = 0$:

$$\begin{aligned}&\mathbb{E}_y g(x + \eta z + \sqrt{\eta} y) = g(x) + \mathbb{E}_y \nabla g(x)^T (\eta z + \sqrt{\eta} y) + \frac{1}{2}(\eta z + \sqrt{\eta} y)^T \nabla^2 g(x) (\eta z + \sqrt{\eta} y)\\ &+ o(\|\eta z + \sqrt{\eta} y\|^2)\\ &= \mathbb{E}_y g(x) + \eta \nabla g(x)^T z + \frac{1}{2}\eta^2 z^T \nabla^2 g(x) z + \eta \frac{1}{2} y^T \nabla^2 g(x) y + o(\eta \|y\|^2)\\ &= \mathbb{E}_y g(x) + \eta \nabla g(x)^T z + \eta \frac{1}{2} y^T \nabla^2 g(x) y + o(\eta).\end{aligned}$$

Combining the assumption with the above Equation, we get that for small enough $\eta$, we have

$$\eta \nabla g(x)^T z + \eta \frac{1}{2} \mathbb{E}_y y^T \nabla^2 g(x) y \leq -\eta r/2,$$

i.e.

$$\nabla g(x)^T z + \frac{1}{2}\mathbb{E}_y y^T \nabla^2 g(x) y \leq -r/2,$$

which means we should either have

$$\nabla g(x)^T z \leq -r/4,$$

which implies

$$\|\nabla g(x)\| \geq \frac{r}{4\|z\|},$$

or

$$\mathbb{E}_y y^T \nabla^2 g(x) y \leq -r/2,$$

which implies

$$\lambda_{min}\Big(\nabla^2 g(x)\Big) \leq -\frac{r}{2\max_{\tilde{y}\in \text{support}(y)} \|\tilde{y}\|^2}.$$

### 3.7.4 Handling the Injected Noise by PSGD

In this section, we prove that having SGD injecting noise into our gradient estimates mostly does not change the sign pattern of the first layer, namely among the set of rows in $P$ defined in Lemma 10.

**Lemma 54.** *Having enough overparameterization, with high probability, at every iteration of the* PSGD *for* ${W'}^{(2)}$ *defined in the proof of Lemma 10, we have for every* $j \in [m_1]$*:*

$$\|{W'}^{(2)}_j\| \leq c_2/(4\sqrt{m_1}).$$

**Proof of Lemma 54**

Let $\Phi'$ be the subspace of the first layer weight matrices which is zero in rows $j \in P$ ($P$ is defined in Lemma 10), while in other rows it is the span of $Z_i^k$'s, i.e. using our notation $\tilde{Z}_k^i$ introduced in the proof of Lemma 10, we can write $\Phi'$ is $\text{span}(Z_i^k)_{i,k}$.

Recall from Lemma 10 that we decompose the first layer weight $W'$ as ${W'}^{(1)} + {W'}^{(2)}$, namely the parts in the subspace $\Phi'$ and subspace ${\Phi'}^{\perp}$ respectively. Moreover, let $\Xi_1/(\sqrt{m_1}\|\Xi_1\|) = \Xi^{(1)} + \Xi^{(2)}$ be the decomposition of the injected noise at some iteration of PSGD into subspaces $\Phi'$ and ${\Phi'}^{\perp}$ respectively.

Now recall that the current $W'$ is the value of the previous iteration moved by the gradient plus the injected noise:

$$\begin{aligned} W' &= W' - \eta(\hat{\nabla}_{W'} + \Xi^{(1)} + \Xi^{(2)}) \\ &= W' - \eta\Big(\tilde{\nabla}_{W'} + 2\psi_1 {W'}^{-,(1)} + 2\psi_1 {W'}^{-,(2)} + \Xi^{(1)} + \Xi^{(2)}\Big), \end{aligned}$$

where $W'$ is the weight of the previous iteration and ${W'}^{-,(1)}, {W'}^{-,(2)}$ are again its decomposition to $\Phi'$ and ${\Phi'}^{\perp}$, where $\tilde{\nabla}_{W',V'}$ is defined in Lemma 52. Applying Lemma 33 for the previous iteration of the algorithm, we get $\tilde{\nabla}_{W'} \in \Phi'$ since the bad events $E$ defined in Lemma 42 occurs only with probability exponentially small (hence union bound across all the iterations rules it out). Hence, the decomposition for the current iteration becomes

$$W'^{(1)} = W'^{-,(1)} - \eta(\hat{\nabla}_{W'} + 2\psi_1 W'^{-,(1)} + \Xi^{(1)}), \tag{3.296}$$

$$W'^{(2)} = (1 - 2\eta\psi_1)W'^{-,(2)} + \eta\Xi^{(2)}. \tag{3.297}$$

We handle the $W'^{(1)}$ part in Lemma 10 and prove that as long as $\|W'^{(1)}\|^2 \leq \|W'\|^2$ remains bounded by $C_1^2$, then the sign pattern of the first layer, when only considering the $W'^{(1)}$ part, is specified by the initialization except within set $P$; here we handle the $W'^{(2)}$ part as well.

Note that for every row $j \in [m_1]$, the variable $\|\left(\Xi_1/(\sqrt{m_1}\|\Xi_1\|)\right)_j\|^2$ is $(O(1/(m_1^4 d)), O(1/(m_1^2 d)))$-subexponential with mean $1/m_1$. Therefore, with probability that is exponentially small in $m_1$, $\|\left(\Xi_1/(\sqrt{m_1}\|\Xi_1\|)\right)_j\|$ is bounded by $O(1/m_1)$. It is not hard to see the same argument holds for the projection of $\Xi_1/(\sqrt{m_1}\|\Xi_1\|)$ onto $\Phi^\perp$, i.e. $\Xi^{(2)}$. Applying a union bound for all iterations, again using the fact that we run `PSGD` for poly iterations while the chance of error is exponentially small in $m_1$, we can then argue that with high probability over the noise of gradients, at every iteration and for every $j \in [m_1]$:

$$\|\Xi_j^{(2)}\| = \tilde{O}(1/m_1). \tag{3.298}$$

But applying trinagle inequality to Equation (3.297) and writing it in a telescope form, particularly for the $j$th row, and further using the assumption in 3.298, we get that $\|W'^{(2)}_j\|$ grows at most to $O(1/(m_1\psi_1))$; as we set $1/\psi_1 = O(poly(n))$, assuming polynomially large enough $m_1$ concludes the claim.

### Bounding the Norm of the first Layer's Output in the Worst Case

**Lemma 55.** *Suppose $W'$ satisfies the assumption of Lemma 10, i.e. $\|W'\| \leq C_1$, and $\|W'_j\| \leq c_2/(2\sqrt{m_1})$ except possible for indices in $P$, also defined in 10. Then, with* *high probability over initialization*

$$\sup_{x,\|x\|=1} \|\phi'(x)\| \lesssim (1 + O(m_3^2 d^2 \log(m_1)^2/m_1))C_1 + \sqrt{m}_3 \frac{C_1^{3/2}}{\sqrt{\kappa}_1} (\frac{n^3 m_3}{m_1 \lambda_0})^{1/4},$$

*which is $O(C_1)$ for large enough overparameterization.*

**Proof of Lemma 55**

Note the because the VC-dimension of the class of binary functions with respect to halfspaces in $\mathbb{R}^d$ is $d+1$, the number of different sign patterns $D_{W^{(0)},x}$ for different $x$ can be at most $m_1^{d+1}$. Now similar to Equation (3.299), for $k \in [m_3]$ define

$$Z_k(x) = 1/\sqrt{m_1}\Big(W_{k,j}^s \mathbb{1}\{W_j^{(0)T} x\} x\Big)_{j=1}^{m_1}. \tag{3.299}$$

Then, for $k_1 \neq k_2$, as $\|x\| = 1$:

$$\begin{aligned}\langle Z_{k_1}(x), Z_{k_2}(x)\rangle &= \frac{1}{m_1} \sum_{j=1}^{m_1} W_{k_1,j}^s W_{k_2,j}^s \mathbb{1}\{W_j^{(0)T} x\} \\ &\leq \frac{1}{m_1} \sup_x \sum_{j=1}^{m_1} D_{W^{(0)},xj,j} W_{k_1,j}^s W_{k_2,j}^s.\end{aligned}$$

But for each fixed $D_{W^{(0)},x}$, using Hoeffding bound, we have with probability $1-\delta$:

$$\frac{1}{\sqrt{m}_1} \sum_{j=1}^{m_1} D_{W^{(0)},xj,j} W_{k_1,j}^s W_{k_2,j}^s \lesssim \sqrt{\frac{\log(1/\delta)}{m_1}}.$$

Applying the above for all possible sign patterns with $\delta < O(1/{m_1}^{d+1})$ and a union bound, we have with high probability

$$\sup_{x,\|x\|=1} \langle Z_{k_1}(x), Z_{k_2}(x)\rangle \leq \frac{1}{m_1} \sup_x \sum_{j=1}^{m_1} D_{W^{(0)},xj,j} W_{k_1,j}^s W_{k_2,j}^s \lesssim d\log(m_1)/\sqrt{m_1}.$$

We can even state the following stronger bound with respect to two adversarially picked vectors $x, x'$:

$$\sup_{\|x\|=1, \|x'\|=1} \langle Z_{k_1}(x), Z_{k_2}(x') \rangle \leq \frac{1}{m_1} \sup_x \sum_{j=1}^{m_1} D_{W^{(0)},xj,j} D_{W^{(0)},x'j,j} W^s_{k_1,j} W^s_{k_2,j} \lesssim d \log(m_1)/\sqrt{m_1}, \tag{3.300}$$

because each $D_{W^{(0)},x'j,j}$ has at most $m_1^{d+1}$ cases as we discussed above, then $D_{W^{(0)},xj,j} D_{W^{(0)},x'j,j}$ has at most $m_1^{d+1}$ possible cases, and applying a similar Hoeffding bound for each of them and a union bound as we did will imply (3.300). We will use this generalized version in another section.

Now combining Equation (3.300) with the fact that $\|W'\| \leq C_1$ and applying Lemma 49:

$$\sup_{x, \|x\|=1} \sum_{k=1}^{m_3} \langle W', Z_k(x) \rangle^2 \leq (1 + O(m_3^2 d^2 \log(m_1)^2/m_1)) C_1^2. \tag{3.301}$$

On the other hand, setting $m_2 = m_1$, $m_3 = d$, and $R = c_2/(2\sqrt{m_1}\kappa_1)$ in Lemma 38, we get with high probability

$$\#\Big(j \in [m] : \ |V_j^{(0)} x| \leq c_2/(2\sqrt{m_1})\Big) \leq m_1 c_2/(2\sqrt{m_1}) = \sqrt{m_1} c_2/(2\kappa_1).$$

Noting that $\|W'_j\| \leq c_2/(2\sqrt{m_1})$ for $j \notin P$, we conclude that with high probability, for any $x$, $\mathbb{1}\{(W^{(0)} + W')_j^T x \geq 0\}$ and $\mathbb{1}\{W_j^{(0)T} x \geq 0\}$ can be different in at most $\sqrt{m_1} c_2/(2\kappa_1)$ of the $j$'s outside of $[m_1] \setminus P$. Therefore, as we have also $|P| \lesssim n c_2 \sqrt{m_1}/\kappa_1$ from Lemma 10, we conclude that with high probability, for any $x$, there are at most $O(nc_2\sqrt{m}_1/\kappa_1)$ sign changes by adding $W'$ to $W^{(0)}$. This further implies:

$$\begin{aligned}|\phi_k'(x) - \langle W', Z_k(x)\rangle| &\leq 2/\sqrt{m_1} \sum_{j:\ \mathrm{Sgn}(W_j^{(0)T}x)\neq \mathrm{Sgn}((W^{(0)}+W')_j^T x)} |W_j'x| \\ &\leq \|W'\|2\sqrt{\left|\{j|\ \mathrm{Sgn}(W_j^{(0)T}x) \neq \mathrm{Sgn}((W^{(0)}+W')_j^T x)\}\right|}/\sqrt{m_1} \\ &\lesssim C_1\sqrt{nc_2\sqrt{m}_1/\kappa_1}/\sqrt{m}_1 \\ &= \frac{C_1^{3/2}}{\sqrt{\kappa}_1}(\frac{n^3 m_3}{m_1\lambda_0})^{1/4}.\end{aligned}$$

Combining this with (3.301), we conclude with high probability:

$$\sup_{x,\|x\|=1} \|\phi'(x)\| \lesssim (1 + O(m_3^2 d^2 \log(m_1)^2/m_1))C_1 + \sqrt{m}_3 \frac{C_1^{3/2}}{\sqrt{\kappa}_1}(\frac{n^3 m_3}{m_1\lambda_0})^{1/4},$$

which completes the proof.

# Chapter 4

# Conclusion

In this thesis we studied two closely related problems under a high dimensional setting: testing and learning.

For testing, we settled the sample complexity of testing the important class of DPP distributions; we showed that the exponential dependence in the sample complexity, which is due to the exponential size of the support, is unavoidable. However, this does not rule out the opportunity of adding further constaints to the class of DPPs in the hope of breaking the exponential barrier in the compelxity. As an example, a natural assumption to investigate is if one assumes a low rank structure in the kernel of the DPP distribution. In this regard, an interesting question is also designing computationally efficient algorithms for the task of testing.

For learning with a deep network, we investigated a particular regime regarding a three-layer network model which goes beyond the NTK approximation of the network, and show (1) convergence of a deliberately-chosen variant of SGD in training, and (2) generalization of the trained network with respect to a new data-dependent complexity measure which gneneralizes the NTK-based compelxity measure proposed in Arora et al. [2018a]. An interesting question for future is to try to generalize and strengthen these ideas for multi-layer networks to achieve algorithmic depth-separation results, in regimes that go beyond NTK regarding function classes that are hard to learn with shallower networks.

# Bibliography

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3591–3599, 2015.

Raja Hafiz Affandi, Emily Fox, Ryan Adams, and Ben Taskar. Learning the parameters of determinantal point process kernels. In *Int. Conference on Machine Learning (ICML)*, pages 1224–1232, 2014.

Maryam Aliakbarpour, Themis Gouleakis, John Peebles, Ronitt Rubinfeld, and Anak Yodpinyanee. Towards testing monotonicity of distributions over general posets. In *Conference on Learning Theory (COLT)*, pages 34–82, 2019.

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019a.

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. PMLR, 2019b.

N. Anari and S. Oveis Gharan. The Kadison-Singer problem for strongly Rayleigh measures and applications to asymmetric TSP. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2015.

Nima Anari, Shayan Oveis Gharan, and Alireza Rezaei. Monte Carlo Markov Chain algorithms for sampling strongly Rayleigh distributions and determinantal point processes. In *Conference on Learning Theory (COLT)*, 2016.

Nima Anari, Shayan Oveis Gharan, and Cynthia Vinzant. Log-concave polynomials, entropy, and a deterministic approximation algorithm for counting bases of matroids. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2018.

Nima Anari, Kuikui Liu, Shayan Oveis Gharan, and Cynthia Vinzant. Log-concave polynomials II: High-dimensional walks and an FPRAS for counting bases of a matroid. In *Symposium on Theory of Computing (STOC)*, 2019.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018a.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018b.

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, 2019a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019b.

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

Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory*, pages 195–268. PMLR, 2019.

Rémi Bardenet and Michalis Titsias RC AUEB. Inference for determinantal point processes without spectral knowledge. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3393–3401, 2015.

Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1706.08498*, 2017.

Peter Bartlett, Dave Helmbold, and Philip Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International conference on machine learning*, pages 521–530. PMLR, 2018.

Nematollah Kayhan Batmanghelich, Gerald Quon, Alex Kulesza, Manolis Kellis, Polina Golland, and Luke Bornn. Diversifying sparsity using variational determinantal point processes. *arXiv preprint arXiv:1411.6307*, 2014.

Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 442–451, 2001.

Tugkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Symposium on Theory of Computing (STOC)*, pages 381–390, 2004.

Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):4:1–4:25, 2013.

Eric Blais and Abhinav Bommireddi. Testing submodularity and other properties of valuation functions. *arXiv preprint arXiv:1611.07879*, 2016.

J. Borcea, P. Bränden, and T.M. Liggett. Negative dependence and the geometry of polynomials. *Journal of American Mathematical Society*, 22:521–567, 2009.

Victor-Emmanuel Brunel. Learning signed determinantal point processes through the principal minor assignment problem. In *Advances in Neural Information Processing Systems (NIPS)*, pages 7365–7374, 2018.

Victor-Emmanuel Brunel, Ankur Moitra, Philippe Rigollet, and John Urschel. Rates of estimation for determinantal point processes. In *Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pages 343–345. PMLR, 2017.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *International conference on machine learning*, pages 605–614. PMLR, 2017.

Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Deeparnab Chakrabarty and Zhiyi Huang. Testing coverage functions. In *International Colloquium on Automata, Languages, and Programming*, pages 170–181. Springer, 2012.

Siu-on Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SIAM-ACM Symposium on Discrete Algorithms (SODA)*, pages 1193–1203, 2014.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, pages 690–696. PMLR, 2017.

Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning $k$-modal distributions via testing. *Theory of Computing*, 10(20):535–570, 2014. doi: 10.4086/toc.2014.v010a020. URL `http://www.theoryofcomputing.org/articles/v010a020`.

Michał Dereziński and Michael W. Mahoney. Determinantal Point Processes in Randomized Numerical Linear Algebra. *arXiv e-prints*, art. arXiv:2005.03185, May 2020.

A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. *Symposium on Theory of Computing (STOC)*, 2:225–247, 2006.

Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694, 2016.

Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *International Colloquium on Automata, Languages, and Programming*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Josip Djolonga and Andreas Krause. From MAP to marginals: Variational inference in Bayesian submodular models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 244–252, 2014.

Josip Djolonga, Stefanie Jegelka, and Andreas Krause. Provable variational inference for constrained log-submodular models. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Xialiang Dou and Tengyuan Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *Journal of the American Statistical Association*, pages 1–14, 2020.

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

Lixin Duan, Ivor W Tsang, and Dong Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479, 2012.

D Dubhashi and Alessandro Panconesi. Concentration of measure for the analysis of randomised algorithms. *Draft Manuscript, http://www.brics.dk/ale/papers. html*, 1998.

Christophe Dupuy and Francis Bach. Learning determinantal point processes in sublinear time. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 244–257, 2018.

Freeman J Dyson. Statistical theory of the energy levels of complex systems. i. *Journal of Mathematical Physics*, 3(1):140–156, 1962.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Vitaly Feldman and Jan Vondrak. Optimal bounds on approximation of submodular and xos functions by juntas. *SIAM Journal on Computing*, 45(3):1129–1170, 2016.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

A. Frieze, N. Goyal, L. Rademacher, and S. Vempala. Expanders via random spanning trees. *SIAM Journal on Computing*, 43(2):497–513, 2014.

Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Bayesian low-rank determinantal point processes. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 349–356. ACM, 2016.

Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Low-rank factorization of determinantal point processes. In *Proc. AAAI Conference on Artificial Intelligence*, 2017.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pages 797–842. PMLR, 2015a.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 797–842, Paris, France, 03–06 Jul 2015b. PMLR. URL `http://proceedings.mlr.press/v40/Ge15.html`.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

S. Oveis Gharan, A. Saberi, and M. Singh. A randomized rounding approach to the traveling salesman problem. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 550–559, 2011.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural networks. *arXiv preprint arXiv:1906.08899*, 2019.

Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. Expectation-maximization for learning determinantal point processes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3149–3157, 2014.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2069–2077. Curran Associates, Inc., 2014a.

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2069–2077, 2014b.

Alkis Gotovos, S. Hamed Hassani, and Andreas Krause. Sampling from probabilistic submodular models. In *Advances in Neural Information Processing Systems (NIPS)*, December 2015.

Alkis Gotovos, Hamed Hassani, Andreas Krause, and Stefanie Jegelka. Discrete sampling using semigradient-based product mixtures. In *Uncertainty in Artificial Intelligence (UAI)*, August 2018.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

J. Ben Hough, Manjunath Krishnapur, Yuval Peres, and Balint Virag. Determinantal processes and independence. *Probability Surveys*, 3:206–229, 2006.

Ilse CF Ipsen and Rizwana Rehman. Perturbation bounds for determinants and characteristic polynomials. *SIAM Journal on Matrix Analysis and Applications*, 30 (2):762–776, 2008.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

A. Javanmard, M. Mondelli, and A. Montanari. Analysis of a two-layer neural network via displacement convexity. *ArXiv*, abs/1901.01375, 2019.

Adel Javanmard, Marco Mondelli, and Andrea Montanari. Analysis of a two-layer neural network via displacement convexity. *The Annals of Statistics*, 48(6):3619–3642, 2020.

Kenji Kawaguchi. Deep learning without poor local minima. In *nips*, 2016a.

Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016b.

Pitas Konstantinos, Mike Davies, and Pierre Vandergheynst. Pac-bayesian margin bounds for convolutional neural networks-technical report. *arXiv preprint arXiv:1801.00171*, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Alex Kulesza and Ben Taskar. Learning determinantal point processes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 419–427, Arlington, Virginia, USA, 2011a. AUAI Press. ISBN 9780974903972.

Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *Int. Conference on Machine Learning (ICML)*, pages 1193–1200, 2011b.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

John A. Kulesza. *Learning with Determinantal Point Processes*. PhD thesis, University of Pennsylvania, 2012.

Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):853–877, 2015.

Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast DPP sampling for Nyström with application to kernel methods. In *Int. Conference on Machine Learning (ICML)*, 2016a.

Chengtao Li, Suvrit Sra, and Stefanie Jegelka. Fast mixing markov chains for Strongly Rayleigh measures, DPPs, and constrained sampling. In *Advances in Neural Information Processing Systems (NIPS)*, 2016b.

Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Polynomial time algorithms for dual volume sampling. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *arXiv preprint arXiv:1808.01204*, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. *arXiv preprint arXiv:1705.09886*, 2017.

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

Hui Lin and Jeff Bilmes. Learning mixtures of submodular shells with application to document summarization. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 479–490. AUAI Press, 2012. ISBN 9780974903989.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, , and Lexing Ying. A mean-field analysis of deep resnet and beyond: Towards provable optimization via overparameterization from depth. In *icml*, 2020.

Odile Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

Zelda Mariet and Suvrit Sra. Fixed-point algorithms for learning determinantal point processes. In *Int. Conference on Machine Learning (ICML)*, pages 2389–2397, 2015.

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.

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.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231 (694-706):289–337, 1933.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International conference on machine learning*, pages 2603–2612. PMLR, 2017.

Liam Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE TOIT*, 54:4750–4755, 2008.

Prasad Raghavendra, Nick Ryder, and Nikhil Srivastava. Real stability testing. In *Innovations in Theoretical Computer Science*, 2017.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2847–2854. PMLR, 06–11 Aug 2017. URL `http://proceedings.mlr.press/v70/raghu17a.html`.

Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pages 2595–2623. PMLR, 2019.

Alain Rakotomamonjy, Francis Bach, Stéphane Canu, and Yves Grandvalet. More efficiency in multiple kernel learning. In *Proceedings of the 24th international conference on Machine learning*, pages 775–782, 2007.

Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

Ronitt Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

Itay Safran and Ohad Shamir. Depth separation in relu networks for approximating smooth non-linear functions. *arXiv preprint arXiv:1610.09887*, 14, 2016.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pages 4433–4441. PMLR, 2018.

Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: what is actually being separated? In *Conference on Learning Theory*, pages 2664–2666. PMLR, 2019.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Comandur Seshadhri and Jan Vondrák. Is submodularity testable? *Algorithmica*, 69 (1):1–25, 2014.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Jasper Snoek, Richard Zemel, and Ryan P Adams. A determinantal point process latent variable model for inhibition in neural spiking data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1932–1940, 2013.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7: 1531–1565, 2006.

Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

D.A. Spielman and N. Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL `https://proceedings.neurips.cc/paper/2010/file/76cf99d3614e23eabab16fb27e944bf9-Paper.pdf`.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Yuandong Tian. Symmetry-breaking convergence analysis of certain two-layered neural networks with relu nonlinearity. *Network*, 100(1):1–1, 2016.

Sebastian Tschiatschek, Josip Djolonga, and Andreas Krause. Learning probabilistic submodular diversity models via noise contrastive estimation. In *Proc. Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

John Urschel, Victor-Emmanuel Brunel, Ankur Moitra, and Philippe Rigollet. Learning determinantal point processes with moments and cycles. In *Int. Conference on Machine Learning (ICML)*, pages 3511–3520. JMLR. org, 2017.

Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *JACM*, 64(6):37:1–37:41, 2017.

Colin Wei, Jason Lee, Qiang Liu, and Tengyu Ma. On the margin theory of feedforward neural networks. 2018.

Colin Wei, Jason Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. 2019.

Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. Practical diversified recommendations on YouTube with Determinantal Point Processes. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2018.

Lei Wu, Chao Ma, and Weinan E. A priori estimates of the generalization error for two-layer neural networks. 2018.

Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62 (6):3702–3720, 2016.

Yihong Wu, Pengkun Yang, et al. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019.

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

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International conference on machine learning*, pages 4140–4149. PMLR, 2017.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. *arXiv preprint arXiv:1804.05862*, 2018.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. arxiv e-prints, art. *arXiv preprint arXiv:1811.08888*, 2018.