High-Dimensional Statistical Inference from Coarse and Nonlinear Data: Algorithms and Guarantees

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Haoyu Fu, M.S.

Graduate Program in Electrical and Computer Engineering

The Ohio State University

2019

Dissertation Committee:

Yingbin Liang, Advisor

Yuejie Chi, Co-Advisor

Kiryung Lee

Philip Schniter

Abstract

Learning a postulated parametric model from the acquired data to extract useful information is of great importance in modern signal processing, machine learning and statistics. The linear model, where the observed data are assumed to depend linearly on the input data, has been studied extensively and applied successfully to many applications. However, the linear assumption is quite restricted, creating a major roadblock for its accuracy and universality, since the dependency of the data is nonlinear in general, and cannot be approximated by a linear model. The challenges of learning these nonlinear models include high computational cost and susceptibility to local minima in their associated optimization problems. Through case studies, we highlight the statistical and computational issues when learning from high-dimensional coarse and nonlinear data, with the hope of shedding light on resolving these challenges.

In this thesis, we consider data from typical signal processing and machine learning applications. In the context of signal processing, we study the problem of estimating spectrally-sparse signals from their quantized noisy complex-valued random linear measurements, a problem arising naturally from analog-to-digital conversion in sub-Nyquist spectrum sampling. We first study the effects of quantization on estimating the spectrum by characterizing the Cramér-Rao bound under the additive white Gaussian noise. We use the calculated bound to highlight the trade-off between the sample complexity and the bit depth under different signal-to-noise ratios for a

fixed budget of bits. Secondly, we formulate a convex optimization approach based on atomic norm soft thresholding to estimate the spectrum of the signal, which is computationally more efficient than the maximum-likelihood estimator.

Moving to the context of machine learning, we study several one-hidden-layer neural network models for nonlinear regression using both cross-entropy and least-squares loss functions. The neural-network-based models have attracted a significant amount of research interest due to the success of deep learning in practical domains such as computer vision and natural language processing. Learning such neural-network-based models often requires solving a non-convex optimization problem. We propose different strategies to characterize the optimization landscape of the non-convex loss functions and provide guarantees on the statistical and computational efficiency of optimizing these loss functions via gradient descent.

Dedicated to my parents

Acknowledgments

First and foremost, I would like to express my sincere gratitudes to my two advisors, Prof. Yingbin Liang and Prof. Yuejie Chi, for their great guidance and support over the past few years. Prof. Liang and Prof. Chi are both great researchers. Their passion for exploring unknown problems has motivated and inspired me a lot while I develop my own research path. Their valuable feedback and insightful suggestions help me a lot in improving the quality of my research work. Their kindness and wisdom would continue to impact me. I feel so lucky to be their student.

I would like to thank my thesis committee members, Prof. Kiryung Lee and Prof. Philip Schniter, for their insightful comments and invaluable suggestions on my thesis.

I would like to thank Dr. Pu (Perry) Wang who was my mentor when I interned at Mitsubishi Electric Research Laboratories. He gave me plenty guidance and meaningful suggestions on our research project.

I'm also grateful to fellow graduate students in The Ohio State University, including Yuanxin Li, Liming Wang, Ziwei Guan, Yi Zhou and Vince Monardo for their generous help, encouragement and interesting discussions. I want to thank the faculty and staff from The Ohio State University for their selfless dedication to the education career. I benefit tremendously from their solid professionalism.

Lastly, I would like to thank my parents, and my sister for their love, support and encouragement. I thank Yubo for sharing hundreds of delightful stories, and bringing so much happiness to my life. To them I dedicate this thesis.

Vita

2014 ..... B.Eng., The measurement and control technology and instrument,
Wuhan University of Technology,
Wuhan, China

2019 ..... M.S., Electrical and Computer Engineering,
The Ohio State University,
Columbus, USA

Publications

Research Publications

H. Fu, Y. Chi, and Y. Liang, "Guaranteed Recovery of One-Hidden-Layer Neural Networks via Cross Entropy", *submitted*.

H. Fu and Y. Chi, "Quantized Spectral Compressed Sensing: CramerRao Bounds and Recovery Algorithms", *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3268-3279, 2018.

Y. Chi and H. Fu, "Subspace Learning From Bits", *IEEE Transactions on Signal Processing*, vol. 65, no. 12, pp. 4429-4442, 2017.

H. Fu, Y. Chi, and Y. Liang, "Local Geometry of Cross Entropy Loss in Learning One-Hidden-Layer Neural Networks", *IEEE International Symposium on Information Theory*, Paris, France, 2019.

H. Fu, P. Wang, T. Koike-Akino, R. Ma, B.Wang, P.V. Orlik, W. Tsujitaz, K. Sadamotoz, Y. Sawaz, K. Katox and M. Nakajima, "Terahertz Imaging of Multi-Level Pseudo-Random Refectance", *International Conference on Infrared, Millimeter, and Terahertz Waves*, Nagoya, Japan, 2018.

P. Wang, H. Fu, T. Koike-Akino, and P.V. Orlik, "Multi-Layer Terahertz Imaging of Non-Overlapping Contents", *IEEE Sensor Array and Multi-Channel Signal Processing Workshop*, Sheffield, UK, 2018.

H. Fu, P. Wang, T. Koike-Akino, P.V. Orlik, and Y. Chi, "Terahertz imaging of binary reflectance with variational Bayesian inference", *IEEE International Conference on Acoustics, Speech and Signal Processing*, Alberta, Canada, 2018.

H. Fu and Y. Chi, "Compressive Spectrum Estimation using Quantized Measurements", *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, 2017.

H. Fu and Y. Chi, "Principal Subspace Estimation for Low-rank Toeplitz Covariance Matrices with Binary Sensing", *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, 2016.

Fields of Study

Major Field: Electrical and Computer Engineering

Table of Contents

Page

List of Figures

Chapter 1: Introduction

This thesis studies several problems on learning from coarse and nonlinear data. In this chapter, we first explain the motivation of these problems, and describe the major approach to solving them. We then summarize our main contributions with respect to the state-of-the-art in the literature.

1.1 Motivation

In order to extract latent information from the collected data samples in many machine learning, signal processing and statistical inference tasks, a parametric model is generally proposed to postulate a data generating mechanism. Fitting an effective and reasonable model is then equivalent to recovering a finite set of parameters. In this thesis, we follow this direction and focus on parametric models to fit the data. Assuming the observed data is generated from a true parametric model, our goal is to recover the parameters of the underlying model from the acquired data.

One example is the classical linear model, where we aim to estimate a vector $\boldsymbol{w}^\star \in \mathbb{R}^d$ from a set of independent and identically distributed (i.i.d.) data samples

$$\left\{(\boldsymbol{x}_i, y_i) : \boldsymbol{x}_i \in \mathbb{R}^d \text{ and } y_i \in \mathbb{R}\right\}_{i=1}^n,$$

by assuming the observation y_i depends linearly on the vector $\boldsymbol{w}^\star$, through the following model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{w}^\star + \nu_i, \quad i = 1, 2, \cdots, n, \tag{1.1}$$

where ν_i stands for some noise.

To find an estimator $\widehat{\boldsymbol{w}}$ that approximates $\boldsymbol{w}^\star$, a common strategy is to form a proper optimization problem with the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, i.e., solving the following optimization problem

$$\begin{aligned} \widehat{\boldsymbol{w}} = & \underset{\boldsymbol{w}}{\operatorname{argmin}} \quad L_n(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i), \\ & \text{subject to} \quad \boldsymbol{w} \in \mathcal{C}. \end{aligned} \tag{1.2}$$

Here, the loss function $L_n(\boldsymbol{w})$ measures the discrepancy between the candidate model and the true model, and the set $\mathcal{C}$ is the feasible region which has incorporated the constraint or prior information, if any. For example, when the noise ν_i follows i.i.d. Gaussian distribution and $\mathcal{C} = \mathbb{R}^d$, it is not a bad idea to take $\ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i) = \left(\boldsymbol{x}_i^\top \boldsymbol{w} - y_i\right)^2$, the least-squares, or the quadratic loss, which yields the maximum likelihood estimator (MLE).

Although the linear model (1.1) is equipped with many good properties, for example, it is easy to interpret, and computationally efficient, to name a few. However, the stringent assumption that the output depends on the input via a linear relationship has been a serious impediment for its accuracy and universality, since the relationship of the data can be nonlinear in general. One example of the non-linearity is when the observed data is quantized into a binary observation, i.e., $y_i \in \{-1, 1\}$. Such binary observations have emerged in many applications, such as 1-bit compressed sensing [1] in signal processing, and binary classification in machine learning. Another example

is that the observed data is continuous but does not necessarily depends linearly on the input data, such as the output of a neural network with nonlinear activations.

The binary or the general nonlinear nature of the output data has raised many new challenges which are not inherent in the linear model (1.1). In particular, the computational cost of the MLE with quantized data could be much higher. When the quantizer is not given, the effect of quantization has not been fully understood yet, a MLE cannot be formulated. Finding the MLE for a general nonlinear model often requires solving a non-convex optimization problem which is much more challenging. Hence, in this thesis we study the problem of learning from coarse and nonlinear data to shed light on understanding quantization effects as well as developing algorithms for solving a type of non-convex optimization problem.

1.2 Convex and Nonconvex Approaches

One of the most popular methods to recover the model parameterized with $\boldsymbol{w}^\star$ from the observed data is via solving the optimization problem (1.2). The loss function $L_n(\cdot)$ or the feasible region $\mathcal{C}$ is often nonconvex in many problems. The main challenge caused by the nonconvexity is the existence of local minima and saddle points that can attract the algorithm.

One way to circumvent this challenge is to reformulate the problem as a convex optimization problem via convex relaxation, which excuses the need to worry about the spurious local minima. For example, in the problem of spectrum estimation which will be studied later, we assume the signal $\boldsymbol{w}^\star$ is approximately sparse in the spectral domain. However, incorporating such sparse prior makes the feasible set $\mathcal{C}$ nonconvex, and searching in such a nonconvex space is not computationally tractable. To deal

with this problem, we take an atomic norm [2] constraint which convexifies the exact sparsity constraint on $\boldsymbol{w}$. Such a convex relaxation approach has achieved tremendous theoretical success in solving various problems [3, 4].

Differently from convex relaxation, an alternative approach is to directly solve the nonconvex optimization problem with careful local landscape analysis and initialization. Although solving general nonconvex optimization problems is computationally intractable due to the existence of spurious local minima or saddle points, solving certain nonconvex problems under proper statistical models may not be that difficult. For certain types of nonconvex problems, it was shown that in the local region around the global minima, the objective function usually possesses a benign landscape [5–7]. For example, in the problem of recovering neural-network-based models that will be studied later, we found that there often exists a reasonably large basin of attraction around the ground truth $\boldsymbol{w}^\star$. Particularly, we can either explore the second-order (Hessian) property of the loss function to establish the local strong convexity or check the first-order (gradient) property to verify a so-called regularity condition under some mild conditions. Once initialized in such local regions, simple algorithms like gradient descent can find the global optimum $\boldsymbol{w}^\star$.

This Thesis explores the convex approach to solve the problem of spectrum estimation from quantized measurements in the context of signal processing, and adopts the nonconvex approach to solve the neural-network-based model recovery problem in the context of machine learning.

1.3 Spectrum Estimation from Quantized Measurements

We first study a problem in the context of signal processing, i.e., estimating a spectrally-sparse signals from its heavy quantized linear measurements. In this problem, the non-linearity of the data comes from the quantization procedure, i.e., the output data depends on the input through

$$y_i = f\left(\boldsymbol{x}_i^\top \boldsymbol{w}^\star + \nu_i\right), \tag{1.3}$$

where $f(\cdot)$ is used to model the quantization non-linearity. More specifically, we study high-resolution spectrum estimation of a band-limited signal from quantization of its noisy random linear measurements. The signals of interest are spectrally sparse, which are modeled as a linear superposition of complex sinusoids with continuous-valued frequencies. In the extreme 1-bit case, the quantization is based on the quadrants of the complex-valued measurements. More generally, sophisticated quantization schemes such as Lloyds quantizer [8] can be used to allow a higher bit depth. The specific form of the quantizer can be either known or unknown. In addition, the quantized measurements may be additionally contaminated by a noise model, in order to model imperfections in the quantization.

The reasons we study this problem are two-fold. Compressed Sensing (CS) [9, 10] has emerged as an effective approach to allow sub-Nyquist sampling [11–13] when the wideband signal is approximately sparse in the spectral domain. The resulting paradigm is referred to as *Compressive Spectrum Sensing* [14, 15]. Significant focus has been put on reducing the sampling rates of the analog-to-digital converters (ADC), which only covers one aspect of the operations of ADCs. Quantization, which maps the analog samples into a finite number of bits for digital processing, is

another necessary step that requires careful treatments. Most existing works, with a few exceptions, assume that the samples are quantized at a high bit level so that the quantization error is relatively small and well-behaved. Another motivation comes from the application wideband spectrum sensing in bandwidth-constrained wireless networks [16, 17]. In order to reduce the communication overhead, each sensor transmits quantized messages, e.g. 1-bit messages; and it is necessary to estimate wideband spectrum from quantized measurements at the fusion center. Moreover, the quantization scheme might be unknown, due to lack of the knowledge of noise statistics or privacy constraints. Therefore, it is necessary to develop estimators that do not require exact knowledge of the quantizers.

Hence the study aims at understanding the fundamental limits of quantization, as well as developing computationally efficient algorithms, for compressive spectrum sensing and parameter estimation, in particular in the regime of *heavy* quantization where it is no longer appropriate to model quantization errors as bounded additive noise. Examining the figure-of-merit of ADCs, two key specifications are the sampling rate and the effective number of bits (ENOB), which is the number of bits per measurement, also known as the *bit depth*. Typically, a small bit depth allows a high sampling rate, and vice versa [18]. Therefore, it is critical to understand the fundamental trade-off between sampling rate and bit depth for high-resolution spectrum estimation. Though the importance of understanding such trade-off has been realized in the context of CS [19, 20], they haven't been studied for the task of *parameter estimation* using estimation-theoretic tools.

Contributions

In this study, we first derive the Cramér-Rao bound (CRB) for estimating multiple frequencies and their complex amplitudes assuming additive white Gaussian noise (AWGN) and the Lloyd's quantizer, using a fixed and deterministic CS measurement matrix. Our bounds suggest that the CRB experiences a phase transition depending on the signal-to-noise ratio (SNR) *before* quantization. In the low SNR regime it is *noise-limited*, and behaves similarly as if there was no quantization; in the high SNR regime, it is *quantization-limited*, and experiences severe performance degeneration due to quantization. Furthermore, we use the derived CRB to answer the following question: given the same budget of bits, should we use more measurements (high sample complexity) with low bit-depth, or fewer measurements (low sample complexity) with high bit-depth? We answer this question by comparing 1-bit versus 2-bit quantization schemes using the CRB, and demonstrate the answer depends on the SNR. At low SNR, 1-bit measurements are preferred, while at high SNR, 2-bit measurements are preferred.

It is well-known that maximum likelihood estimators approach the performance of CRB asymptotically at high SNR [21], however, their implementation requires exact knowledge of the likelihood function, which in our problem, includes the exact form of the quantizer and noise statistics. However, such knowledge may not be available in certain applications. Therefore, our goal is to develop estimators that do not require the knowledge of the quantization scheme. To mitigate basis mismatch [22], atomic norm [2, 23–30] has been proposed recently to promote spectral sparsity via convex optimization without discretizing the frequencies onto a finite grid, which has found

applications in signal denoising, interpolation of missing data, and frequency localization of spectrally-sparse signals. Existing atomic norm minimization algorithms assume unquantized measurements that are possibly contaminated by additive noise, and a direct application will lead to highly sub-optimal performance when a significant amount of the quantized measurements *saturate* [31].

We propose a novel atomic norm soft thresholding (AST) algorithm [29] to recover spectrally-sparse signals and estimate the frequencies from their 1-bit quantized measurements. Our algorithm is based on finding the proximal mapping of properly designed surrogate signals, that are formed by linear combinations of the sample-modulated measurement vectors, with respect to the atomic norm to promote spectral sparsity. In other words, we aim to find signals that balance between the proximity to the surrogate signals and the small atomic norm. Moreover, the frequencies can be localized without knowing the model order a priori, by examining the peak of a dual polynomial constructed from the dual solution. Alternatively, conventional subspace methods can be used to estimate the frequencies using the recovered spectral signal. The proposed algorithm can be generalized to handle quantizations of noisy random linear measurements of multiple spectrally-sparse signals [25], where each signal contains the same set of frequencies with different coefficients. The proposed algorithms do not require knowledge of the specific form of the quantizer, and therefore can be applied even when the quantizer is unknown.

Related Work

This study is closely related to 1-bit compressed sensing [1, 32–38], which aims to recover a sparse signal from signs of random linear measurements. In particular, Plan and Vershynin [34–37] generalize this idea to reconstructing signals that belong

to some low-dimensional set. Very recently, [39] studied a similar setup and proposed a new algorithm using projected gradient descent. The surrogate signals used in our algorithm can be traced back to [20, 37]. The difference lies in that instead of projecting the surrogate signals directly onto some low-dimensional set, we adopt the proximal mapping of the surrogate signals with respect to the atomic norm. Several algorithms have been proposed in the CS literature to deal with general quantization schemes [19, 40] and nonlinear measurement schemes [37, 39], however the focus has been on reconstruction of sparse signals in a finite dictionary, whereas our focus is on parameter estimation and reconstructing sparse signals in a parametric dictionary containing an infinite number of atoms.

There are also several conflicting evidence regarding the trade-offs between bit-depth and sample complexity [20, 41] for signal reconstruction, as they may vary for different problems when using specific algorithms. In contrast, we derive the Cramér-Rao bound for *parameter estimation* using quantized compressive random measurements, which provides an estimation-theoretic baseline for gauging the trade-off as well as benchmarking performances. Our CRB adds to existing literature of CRB calculations for 1-bit quantized single-tone frequency estimation [42] as well as for parameter estimation using compressive measurements [43].

1.4 Learning One-Hidden-Layer Neural Networks for Binary Classification

We then study the problem of learning neural-network-based models in the context of machine learning, i.e., we assume the training samples $(\boldsymbol{x}_i, y_i) \sim (\boldsymbol{x}, y)$, $i = 1, \ldots, n$, are generated independently and identically distributed (i.i.d.) from a distribution based on a neural network model with the ground truth parameter $\boldsymbol{W}^\star \in \mathbb{R}^{d\times K}$. The

output of an one-hidden-layer neural network can be written as

$$f\left(\boldsymbol{W}^{\star}, \boldsymbol{x}_i\right) = \frac{1}{K} \sum_{k=1}^{K} \phi\left(\boldsymbol{x}_i^{\top} \boldsymbol{w}_k^{\star}\right), \tag{1.4}$$

where $\phi(\cdot)$ is the activation function, and $\boldsymbol{w}_k^{\star}$ denotes the k-th column of $\boldsymbol{W}^{\star}$. The goal is to recover the underlying model parameter $\boldsymbol{W}^{\star}$ using the training samples. The non-linearity comes from two-fold. First the activation function considered is either Sigmoid $\phi(x) = \frac{1}{1+\exp(-x)}$ or ReLU $\phi(x) = \max(x, 0)$, both of which are nonlinear. Second with the binary classification setting we considered, y_i is quantized to discrete label $\{-1, 1\}$.

Neural networks have attracted a significant amount of research interest in recent years due to the success of deep neural networks [44] in practical domains such as computer vision and artificial intelligence [45–47]. Extensive studies have established the expressive power of neural networks. In particular, one-hidden-layer neural networks are sufficient to approximate any continuous function under certain conditions. However, the theoretical underpinnings behind this model remains mysterious to a large extent. This motivates us to study such a one-hidden-layer neural networks.

As with learning the linear model (1.1), an optimization problem (1.2) will be formulated for learning the neural-network-based model as well. Due to the non-linearity of the neural-network-based model, the optimization problem that needs to be solved is often non-convex. The intricate and unknown landscape of the non-convex objection function makes the problem much more difficult than the convex case.

Contributions

The main purpose of this study is to answer whether the true parameter can be recovered from its finite non-linear observations via solving the formulated non-convex optimization problem. As most machine learning tasks can be categorized as a classification problem or a regression problem, we considered both settings in this study, i.e.,

- *Regression*, where each sample $y \in \mathbb{R}$ is generated as

$$y = f(\boldsymbol{W}^\star, \boldsymbol{x}).$$

  This type of regression problem has been studied in various settings. In particular, [48] studied the single-neuron model under the Rectified Linear Unit (ReLU) activation, [7] studied the one-hidden-layer multi-neuron network model, and [49] studied a two-layer feedforward network with ReLU activations and identity mapping.

- *Classification*, where a label $y \in \{0, 1\}$ is drawn according to the conditional distribution

$$\mathbb{P}(y = 1|\boldsymbol{x}) = f(\boldsymbol{W}^\star, \boldsymbol{x}).$$

  Such a problem has been studied in [50] when the network contains only a single neuron.

For both cases, previous studies attempted to recover $\boldsymbol{W}^\star$, by minimizing an empirical loss function using the squared loss, i.e. $\min_{\boldsymbol{W}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{W}, \boldsymbol{x}_i))^2$, given the training data. Two types of statistical guarantees were provided for such model recovery problems using the squared loss. More specifically, [7] showed that in the

local neighborhood of the ground truth $\boldsymbol{W}^\star$, the *empirical* loss function is strongly convex for each *given* point under *independent* high probability event, which implies that *fresh samples* are required at *every* iteration for gradient descent to converge linearly with well-designed initializations. On the other hand, studies such as [50] established strong convexity in the entire local neighborhood of the ground truth in a uniform sense, so that resampling per iteration is not needed for gradient descent to have guaranteed linear convergence as long as it enters such a local neighborhood. Here, one weakness of the pointwise strong convexity in [7], compared to the uniform strong convexity in [50], is that independent fresh samples are required at each iteration to guarantee the linear convergence of gradient descent. Consequently, the sample complexity of [7] grows with respect to the recovery accuracy ϵ, typically with an extra factor of $\log(1/\epsilon)$ under linear convergence, which can be large when the desired accuracy is high. Therefore, the latter type of uniform strong convexity *without requiring per-iteration resampling* is much stronger and more desirable.

Considering the multi-neuron classification problem with either FCN or CNN. We show that the empirical risk function $L_n(\boldsymbol{W})$ is *uniformly* strongly convex in a local neighborhood of the ground truth $\boldsymbol{W}^\star$, hence if initialized in this neighborhood, gradient descent converges linearly to a critical point (which we show to exist). Due to the nature of quantized labels here, the recovery of the ground truth is only up to certain statistical accuracy. Finally, we show that a tensor method provably provides an initialization in the neighborhood of the ground truth both for FCN and CNN.

The cross-entropy loss is much more challenging to analyze than the squared loss, e.g., its gradient and Hessian take much more complicated forms compared with the squared loss; moreover, it is hard to control the values of gradient and Hessian due

to the saturation phenomenon, i.e., when $f(\boldsymbol{W}, \boldsymbol{x})$ approaches 0 or 1. In order to establish the uniform local strong convexity property for the cross-entropy loss, we first show the *population* loss is smooth regarding to $\boldsymbol{W}^\star$. Such a property was also established in [7] for the squared loss. However, considering the special form of Hessian under the cross-entropy loss, we need to apply Taylor's approximation together with certain probabilistic upper bounds to control the value of Hessian, and obtain the smooth property. Network-specific quantities to capture the local geometry of the population loss at $\boldsymbol{W}^\star$ for FCN and CNN are derived, which imply that the geometry of CNN is more benign than FCN, corroborated by the numerical experiments. Beyond these two steps, the additional uniform concentration property of the Hessian is of key importance for us to obtain the uniform local strong convexity of the *empirical* loss. To show the uniform concentration of the Hessian, we successfully apply a type of covering argument. Different from the arguments in [50], which deal with the squared loss and are facilitated by certain nice assumptions on the activation functions, the cross-entropy loss is more difficult to apply the covering argument, e.g., both the gradient and Hessian no longer have a deterministic upper bound. Hence, we exploit the property of the sigmoid activation to show that the gradient and the Hessian of the cross-entropy loss are upper bounded with high probability in order to establish the uniform concentration property.

Related Work

Due to the scope, we focus on the most relevant literature on theoretical and algorithmic aspects of learning shallow neural networks via nonconvex optimization. The parameter recovery viewpoint is relevant to the success of nonconvex learning in

signal processing problems such as matrix completion, phase retrieval, blind deconvolution, dictionary learning and tensor decomposition [51–58], to name a few; see also the overview article [59]. The statistical model for data generation effectively removes worst-case instances and allows us to focus on average-case performance, which often possess much benign geometric properties that enable global convergence of simple local search algorithms.

The studies of one-hidden-layer network model can be further categorized into two classes, landscape analysis and model recovery. In the landscape analysis, it is known that if the network size is large enough compared to the data input, then there are no spurious local minima in the optimization landscape, and all local minima are global [60–63]. For the case with multiple neurons ($2 \leq K \leq d$) in the under-parameterized setting, the work of Tian [64] studied the landscape of the population squared loss surface with ReLU activations. In particular, there exist spurious bad local minima in the optimization landscape [65,66] even at the population level. Zhong et. al. [7] provided several important geometric characterizations for the regression problem using a variety of activation functions and the squared loss.

In the model recovery problem, the number of neurons is smaller than the input dimension, and all the existing works discussed below assumed the squared loss and (sub-)Gaussian inputs. In the case with a single neuron ($K = 1$), [48] showed that gradient descent converges linearly when the activation function is ReLU, with a zero initialization, as long as the sample complexity is $O(d)$ for the regression problem. When the activation function is quadratic, [67] shows that randomly initialized gradient descent converges fast to the global optimum at a near-optimal sample complexity. On the other hand, [50] showed that when $\phi(\cdot)$ has bounded first, second and third

derivatives, there is no other critical points than the unique global minimum (within a constrained region of interest), and (projected) gradient descent converges linearly with an arbitrary initialization, as long as the sample complexity is $O(d \log^2 d)$ for the classification problem. Moreover, in the case with multiple neurons, [68] showed that projected gradient descent with a local initialization converges linearly for smooth activations with bounded second derivatives for the regression problem, [69] showed that gradient descent with tensor initialization converges linearly to a neighborhood of the ground truth using ReLU activations, and [70] showed the linear convergence of gradient descent with the spectral initialization using quadratic activations. For CNN with ReLU activations, [71] shows that gradient descent converges to the ground truth with random initialization for the population risk function based on the squared loss under Gaussian inputs. Moreover, [72] shows that gradient descent successfully learns a two-layer convolutional neural network despite the existence of bad local minima. From a technical perspective, our study differs from all the aforementioned work in that the cross entropy loss function we analyze has a very different form. Furthermore, we study the model recovery classification problem under the multi-neuron case, which has not been studied before.

1.5 Guaranteed Recovery of CNN with ReLU Activations

Moving beyond the classification setting, we further study a similar model recovery problem under the regression setting using the quadratic loss. In particular, we consider the training data that are generated by a one-hidden-layer convolutional neural network (CNN) with the ReLU activation function. Previous work [71, 73]

focused on analyzing the corresponding population risk function given by

$$L(\boldsymbol{w}) = \mathbb{E}_{\mathcal{D}}\left[\ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i)\right], \tag{1.5}$$

where $\mathcal{D}$ denotes the joint distribution of $(\boldsymbol{x}, y)$. They analyzed the performance of gradient descent that minimizes the population risk (1.5) with suitable assumptions on the joint distribution $\mathcal{D}$, and [71] further characterized the critical points of the population risk function. However, these studies considered only the first-order property about the population risk function. The second-order geometric property of the population risk function has not been explored, which is our interest here. We further wish to leverage the power of such second-order geometric properties for improving the previous results. Moreover, the performance of gradient descent on minimizing the empirical risk function has not been understood yet. Due to the non-smoothness of the ReLU activation, the landscape of the empirical risk function cannot be directly studied via its second-order property, i.e., the Hessian, and new methods need to be developed.

Contributions

We first provide a refined analysis of the landscape of the population risk function $L(\boldsymbol{w})$ under the Gaussian input assumption [71, 73], and show that it is strongly convex in a local neighborhood of the ground truth $\boldsymbol{w}^\star$. Then, we show that gradient descent with random initialization converges to the global minimum $\boldsymbol{w}^\star$ at a linear rate. We further study the empirical risk function and show that such a function satisfies a regularity condition within a local neighborhood of $\boldsymbol{w}^\star$. We then further establish that with a good initialization gradient descent converges linearly to the true weights $\boldsymbol{w}^\star$.

Related Work

Besides the related work already mentioned in section 1.4, we here focus only on the studies of the model recovery problem with ReLU activation. [64] considered a single neuron neural network with ReLU activation, and showed that gradient descent with random initialization learns the true weights. Later, [49] showed that for a one-hidden-layer feed forward neural network with ReLU activation, SGD converges to the global minimum in two phases with small initialization weights. For the one-hidden-layer non-overlap convolutional neural network with ReLU activation, [71] studied the convergence of gradient descent with a fixed output layer and proved that gradient descent with random initialization can recover the weights exactly. Subsequently [73] generalized their result by not fixing the output layer, and more specifically they showed that when the output layer is to be learned, there exists spurious local minimum, but gradient descent still converges to the true weights with random initialization. These works have all been obtained under the assumption that input training data follow an i.i.d. Gaussian distribution. Moreover, all the convergence analysis of gradient descent methods is with regard to the population risk function.

In the finite sample regime, training the neural network is based on the empirical risk function. [48] studied the single-neuron network with ReLU activation in the high dimensional regime, and showed that projected gradient descent with initialization at the origin can recover the true weights up to some numerical constant under the assumption that the weights belong to a closed set and the input follows a Gaussian distribution.

The most related work to our study is [71], which studied only the population risk function. Our work captures the second-order landscape property of the population risk function, based on which we significantly improve the convergence rate of gradient descent from polynomial in [71] to linear. Inspired by the success of nonconvex learning in signal processing problems [51, 58, 74], we exploit the geometric property of the nonconvex empirical risk function and show the convergence of gradient descent with well-designed initialization.

1.6 Notations

Throughout this thesis, we use boldface letters to denote vectors and matrices, e.g. $\boldsymbol{a}$ and $\boldsymbol{A}$. The Hermitian transpose of $\boldsymbol{a}$ is denoted by $\boldsymbol{a}^{\mathsf{H}}$, the transpose of $\boldsymbol{a}$ is denoted by $\boldsymbol{a}^{\mathsf{T}}$, and $\|\boldsymbol{A}\|$, $\|\boldsymbol{A}\|_{\mathrm{F}}$, $\mathrm{Tr}(\boldsymbol{A})$ denote the spectral norm, the Frobenius norm, and the trace of the matrix $\boldsymbol{A}$, respectively. If $\boldsymbol{A}$ is positive semidefinite (PSD), then $\boldsymbol{A} \succeq 0$. The identity matrix is denoted by $\boldsymbol{I}$. The gradient and the Hessian of a function $f(\boldsymbol{W})$ is denoted by $\nabla f(\boldsymbol{W})$ and $\nabla^2 f(\boldsymbol{W})$, respectively. We use $c, C, C_1, \ldots$ to denote constants whose values may vary from place to place. If necessary, we will introduce additional notations following the convention of notations in specific context in each chapter.

Chapter 2: Line Spectrum Estimation with Quantized Measurements

In this chapter, we consider the problem of recovering spectrally-sparse signals from heavy quantizations of their noisy complex-valued random linear measurements, e.g. only the quadrant information. We first derive the Cramér-Rao bound (CRB) for estimating multiple frequencies and their complex amplitudes assuming additive white Gaussian noise (AWGN) and the Lloyd's quantizer, using a fixed and deterministic compressed sensing (CS) measurement matrix. Furthermore, we use the derived CRB to answer the following question: given the same budget of bits, should we use more measurements (high sample complexity) with low bit-depth, or fewer measurements (low sample complexity) with high bit-depth? Moreover, we propose a novel atomic norm soft thresholding (AST) algorithm [29] to recover spectrally-sparse signals and estimate the frequencies from their 1-bit quantized measurements.

In this chapter we adopt the following notations. An indicator function for an event A is denoted as $\mathbb{I}_A$. Denote $\mathcal{T}(\boldsymbol{u}) \in \mathbb{C}^{n\times n}$ as the Hermitian Toeplitz matrix with $\boldsymbol{u}$ as the first column. Define the inner product between two vectors $\boldsymbol{a}, \boldsymbol{b}$ as $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a}^{\mathsf{H}}\boldsymbol{b}$. The cardinality of a set $\mathcal{D}$ is defined as $|\mathcal{D}|$. If $\boldsymbol{A}$ is positive semidefinite (PSD), then $\boldsymbol{A} \succeq 0$. $\Re(y)$ and $\Im(y)$ denote the real and imaginary part of a complex

number y, respectively. The expectation of a random variable a is written as $\mathbb{E}[a]$. Define $\odot$ as entry-wise product.

2.1 Problem Formulation

Let $\boldsymbol{x}^\star \in \mathbb{C}^n$ be a line spectrum signal, which is composed of a small number of spectral lines, defined as

$$\boldsymbol{x}^\star = \sum_{k=1}^{K} c_k \boldsymbol{v}(f_k), \tag{2.1}$$

where K is the number of frequencies or level of sparsity, $c_k = A_k e^{j2\pi\phi_k} \in \mathbb{C}$ is the kth coefficient, $A_k > 0$ is the kth amplitude, $\phi_k \in [0, 1)$ is the kth normalized phase, $f_k \in [0, 1)$ is the kth frequency, and

$$\boldsymbol{v}(f) = \begin{bmatrix} 1 & e^{j2\pi f} & \cdots & e^{j2\pi(n-1)f} \end{bmatrix}^{\mathsf{T}}.$$

In CS, we acquire a set of random linear measurements of $\boldsymbol{x}^\star$, contaminated by additive complex Gaussian noise, where each measurement is given as

$$z_i = \langle \boldsymbol{a}_i, \boldsymbol{x}^\star \rangle + \sigma\epsilon_i, \quad i = 1, \ldots, m, \tag{2.2}$$

where m is the number of measurements, $\boldsymbol{a}_i \in \mathbb{C}^n$'s are the measurement vectors composed of i.i.d. standard complex Gaussian entries $\mathcal{CN}(0, 1)$, σ is the noise level, and we further have i.i.d. $\epsilon_i \sim \mathcal{CN}(0, 1)$. In a vector notation, we write

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}^\star + \sigma\boldsymbol{\epsilon}, \tag{2.3}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m]^{\mathsf{H}} \in \mathbb{C}^{m\times n}$ is the measurement matrix, $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \ldots, \epsilon_m]^{\mathsf{T}}$, and $\boldsymbol{z} = [z_1, z_2, \ldots, z_m]^{\mathsf{T}}$. These measurements are then quantized into a finite number of bits for the ease of digital storage and processing. Denote $\mathcal{Q}(\cdot) : \mathbb{R} \mapsto \mathcal{D}$ as the quantizer that quantizes a real number into a finite alphabet $\mathcal{D}$, where the *bit depth*

is the smallest number of bits necessary to represent $\mathcal{D}$, i.e. $b^\star = \min\{b \in \mathbb{Z}^+ : |\mathcal{D}| \leq 2^b\}$. The quantized measurements $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]^\mathsf{T}$ of $\boldsymbol{z}$ are then denoted as

$$y_i = \mathcal{Q}(\Re(z_i)) + j\mathcal{Q}(\Im(z_i)), \quad i = 1, \ldots, m, \tag{2.4}$$

where we apply the same quantizer $\mathcal{Q}$ to both the real part and the imaginary part of the complex-valued measurement z_i. With slight abuse of notation, we denote the quantized measurements as

$$\boldsymbol{y} = \mathcal{Q}(\boldsymbol{z}). \tag{2.5}$$

Our goal is then to recover $\boldsymbol{x}^\star$, and the set of frequencies $\boldsymbol{f} = \{f_k\}_{k=1}^K$, from the quantized measurements $\boldsymbol{y}$, possibly without a priori knowing the sparsity level K, and the form of the quantizer $\mathcal{Q}$.

Several choices of the quantizer are of special interest. At the extreme, we consider only knowing the quadrature information of z_i, where

$$Q(a) = \operatorname{sign}(a), \quad a \in \mathbb{R}, \tag{2.6}$$

We refer to this quantizer as the *one-bit quantizer*, as only a single bit is used to quantize each real number.

More generally, we consider a quantizer $\mathcal{Q}(\cdot)$ that is fully characterized by the quantization intervals $\{[t_\ell, t_{\ell+1})\}_{\ell=1}^{|\mathcal{D}|-1}$, where $t_0 = -\infty$, $t_{|\mathcal{D}|} = \infty$, $\cup_{\ell=1}^{|\mathcal{D}|}[t_\ell, t_{\ell+1}) = \mathbb{R}$, as well as the representatives of each interval $\omega_\ell \in [t_\ell, t_{\ell+1})$, where

$$Q(a) = \omega_\ell, \quad \text{if} \quad a \in [t_\ell, t_{\ell+1}). \tag{2.7}$$

For example, the Lloyd's quantizer [8] belongs to this form. The choice of the quantization scheme plays an important role in determining the performance of parameter estimation.

2.2 Cramer-Rao Bounds and Trade-offs

In this section, we study the effects of quantization on parameter estimation by deriving the Cramér-Rao bound assuming the quantizer, the sparsity level, and the noise level are known. In particular, the bounds are calculated for 1-bit and general quantizations, respectively, which are then used to study the trade-off between sample complexity and bit depths for a fixed bit budget.

To begin with, we assume the set of parameters, including the frequencies, amplitudes, and phases, given as $\boldsymbol{\kappa} = \{f_k, A_k, \phi_k\}_{k=1}^K \in \mathbb{R}^{3K}$, is deterministic but unknown, the measurement matrix $\boldsymbol{A}$ is deterministic and known. Denote the probability mass function as $p(\boldsymbol{y}|\boldsymbol{\kappa})$, which is given as

$$p(\boldsymbol{y}|\boldsymbol{\kappa}) = \prod_{i=1}^{m} p(y_i|\boldsymbol{\kappa}) = \prod_{i=1}^{m} \left[p(\Re(y_i)|\boldsymbol{\kappa}) \cdot p(\Im(y_i)|\boldsymbol{\kappa})\right], \tag{2.8}$$

where the second equality follows from the fact that ϵ_i is proper. Moreover, let

$$p(\Re(y_i)|\boldsymbol{\kappa}) = \prod_{\omega \in \mathcal{D}} p_{\Re(y_i)}(\omega|\boldsymbol{\kappa})^{\mathbb{I}_{\{\Re(y_i)=\omega\}}} \tag{2.9}$$

$$p(\Im(y_i)|\boldsymbol{\kappa}) = \prod_{\omega \in \mathcal{D}} p_{\Im(y_i)}(\omega|\boldsymbol{\kappa})^{\mathbb{I}_{\{\Im(y_i)=\omega\}}} \tag{2.10}$$

be the probability mass function of $\Re(y)$ and $\Im(y)$, respectively. The Fisher Information Matrix (FIM), denoted by $\boldsymbol{I}(\boldsymbol{\kappa}) \in \mathbb{R}^{3K\times 3K}$, is given as

$$\boldsymbol{I}(\boldsymbol{\kappa}) = \mathbb{E}\left[\left(\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)\left(\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^{\top}\right]. \tag{2.11}$$

Note that for any $1 \le i, j \le m$,

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{\partial \log p(\Re(y_i)|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)\left(\frac{\partial \log p(\Im(y_j)|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^{\top}\right]\\ &= \mathbb{E}\left[\frac{\partial \log p(\Re(y_i)|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right] \cdot \mathbb{E}\left[\frac{\partial \log p(\Im(y_j)|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right]^{\top} = \boldsymbol{0},\end{aligned}$$

where the first equality follows from independence of $\Re(y_i)$ and $\Im(y_j)$, and the second equality follows from the fact

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\partial \log p(\Re(y_i)|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)\right] \\
&= \mathbb{E}\left[\sum_{\omega\in\mathcal{D}} \frac{\mathbb{I}_{\{\Re(y_i)=\omega\}}}{p_{\Re(y_i)}(\omega|\boldsymbol{\kappa})} \frac{\partial p_{\Re(y_i)}(\omega|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right] \\
&= \sum_{\omega\in\mathcal{D}} \frac{\partial p_{\Re(y_i)}(\omega|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} = \frac{\partial\left(\sum_{\omega\in\mathcal{D}} p_{\Re(y_i)}(\omega|\boldsymbol{\kappa})\right)}{\partial \boldsymbol{\kappa}} = \mathbf{0}.
\end{aligned}$$

Thus, plugging (2.8) into (2.11) all cross-terms will be zero and we have

$$\boldsymbol{I}(\boldsymbol{\kappa}) = \sum_{i=1}^{m} \left[\boldsymbol{I}_i^R(\boldsymbol{\kappa}) + \boldsymbol{I}_i^I(\boldsymbol{\kappa})\right], \tag{2.12}$$

where

$$\begin{aligned}
\boldsymbol{I}_i^R(\boldsymbol{\kappa}) &= \mathbb{E}\left[\left(\frac{\partial \log p(\Re(y_i)|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)\left(\frac{\partial \log p(\Re(y_i)|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^\top\right] \\
&= \sum_{\omega\in\mathcal{D}} \frac{1}{p_{\Re(y_i)}(\omega|\boldsymbol{\kappa})}\left(\frac{\partial p_{\Re(y_i)}(\omega|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)\left(\frac{\partial p_{\Re(y_i)}(\omega|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^\top,
\end{aligned}$$

and $\boldsymbol{I}_i^I(\boldsymbol{\kappa})$ can be given similarly by replacing $\Re(y_i)$ with $\Im(y_i)$.

The CRB for estimating $\boldsymbol{\kappa}$, is then given as $\mathrm{CRB}(\boldsymbol{\kappa}) = \boldsymbol{I}(\boldsymbol{\kappa})^{-1}$, and the CRB for estimating the ith parameter in $\boldsymbol{\kappa}$, is given as $[\boldsymbol{I}(\boldsymbol{\kappa})^{-1}]_{i,i}$.

2.2.1 CRB for 1-Bit Quantization

Our goal is then to calculate the FIM in (2.12). We will explain in details the calculations for the 1-bit case. First, since $\Re(z_i) \sim \mathcal{N}(\Re(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle), \frac{1}{2}\sigma^2)$, then

$$p_{\Re(y_i)}(\omega|\boldsymbol{\kappa}) = \mathbb{P}\left(\omega \cdot \Re(z_i) > 0|\boldsymbol{\kappa}\right) = \frac{1}{2} + \omega \cdot \Phi\left(\frac{\Re(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle)}{\sigma}\right),\ \omega = \pm 1, \tag{2.13}$$

where $\Phi(u) = \frac{1}{\sqrt{\pi}}\int_0^u e^{-t^2}dt$. Therefore, by the chain rule,

$$\begin{aligned}
\frac{\partial p_{\Re(y_i)}(\omega|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} &= \frac{\omega}{\sigma}\Phi'\left(\frac{\Re(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle)}{\sigma}\right)\frac{\partial \Re(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle)}{\partial \boldsymbol{\kappa}} \\
&= \frac{\omega}{\sqrt{\pi\sigma^2}}\exp\left(-\frac{\Re(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle)^2}{\sigma^2}\right)\frac{\partial \Re(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle)}{\partial \boldsymbol{\kappa}}.
\end{aligned} \tag{2.14}$$

As a short-hand notation, denote $s_i(\boldsymbol{\kappa}) = \Re(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star \rangle)$ and $r_i(\boldsymbol{\kappa}) = \Im(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star \rangle)$. Plug (2.14) into $\boldsymbol{I}_i^R(\boldsymbol{\kappa})$, we have

$$\boldsymbol{I}_i^R(\boldsymbol{\kappa}) = \frac{4\exp\left(-2s_i(\boldsymbol{\kappa})^2/\sigma^2\right)}{\pi\sigma^2\left[1-4\Phi^2\left(\frac{s_i(\boldsymbol{\kappa})}{\sigma}\right)\right]}\left(\frac{\partial s_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)\left(\frac{\partial s_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^\top,$$

and similarly,

$$\boldsymbol{I}_i^I(\boldsymbol{\kappa}) = \frac{4\exp\left(-2r_i(\boldsymbol{\kappa})^2/\sigma^2\right)}{\pi\sigma^2\left[1-4\Phi^2\left(\frac{r_i(\boldsymbol{\kappa})}{\sigma}\right)\right]}\left(\frac{\partial r_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)\left(\frac{\partial r_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^\top.$$

As a remark, when $\sigma = 0$, the amplitude of the signal cannot be recovered from the 1-bit measurements due to scaling ambiguity, and the FIM becomes singular in this case. Therefore, our expressions for CRB is valid when $\sigma \neq 0$.

2.2.2 CRB for General Quantization

We now explain the calculation for a general quantization scheme. For $\omega_\ell \in \mathcal{D}$, and a corresponding interval $[t_\ell, t_{\ell+1})$, we have

$$\begin{aligned} p_{\Re(y_i)}(\omega_\ell|\boldsymbol{\kappa}) &= \mathbb{P}\left(\Re(z_i) \in [t_\ell, t_{\ell+1})|\boldsymbol{\kappa}\right) \\ &= \int_{\frac{t_\ell - s_i(\boldsymbol{\kappa})}{\sigma}}^{\frac{t_{\ell+1}-s_i(\boldsymbol{\kappa})}{\sigma}} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \\ &= \Phi\left(\frac{t_{\ell+1}-s_i(\boldsymbol{\kappa})}{\sigma}\right) - \Phi\left(\frac{t_\ell - s_i(\boldsymbol{\kappa})}{\sigma}\right), \end{aligned} \tag{2.15}$$

then, following similar arguments, we have

$$\frac{\partial p_{\Re(y_i)}(\omega_\ell|\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} = \frac{1}{\sqrt{\pi\sigma^2}}\left[e^{-\frac{(t_{\ell+1}-s_i(\boldsymbol{\kappa}))^2}{\sigma^2}} - e^{-\frac{(t_\ell - s_i(\boldsymbol{\kappa}))^2}{\sigma^2}}\right]\frac{\partial \Re(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star\rangle)}{\partial \boldsymbol{\kappa}}.$$

Therefore, define

$$\Gamma_i^R(\boldsymbol{\kappa}) = \sum_{\ell=1}^{|\mathcal{D}|-1} \frac{\left[e^{-\frac{(t_{\ell+1}-s_i(\boldsymbol{\kappa}))^2}{\sigma^2}} - e^{-\frac{(t_\ell - s_i(\boldsymbol{\kappa}))^2}{\sigma^2}}\right]^2}{\Phi\left(\frac{t_{\ell+1}-s_i(\boldsymbol{\kappa})}{\sigma}\right) - \Phi\left(\frac{t_\ell - s_i(\boldsymbol{\kappa})}{\sigma}\right)}, \tag{2.16}$$

and

$$\Gamma_i^I(\boldsymbol{\kappa}) = \sum_{\ell=1}^{|\mathcal{D}|-1} \frac{\left[e^{-\frac{(t_{\ell+1}-r_i(\boldsymbol{\kappa}))^2}{\sigma^2}} - e^{-\frac{(t_\ell - r_i(\boldsymbol{\kappa}))^2}{\sigma^2}}\right]^2}{\Phi\left(\frac{t_{\ell+1}-r_i(\boldsymbol{\kappa})}{\sigma}\right) - \Phi\left(\frac{t_\ell - r_i(\boldsymbol{\kappa})}{\sigma}\right)}, \tag{2.17}$$

we have the following theorem for the expression of FIM in light of our derivations in the previous subsection.

Theorem 1. *The Fisher information matrix $\boldsymbol{I}(\boldsymbol{\kappa})$ for estimating the unknown parameter $\boldsymbol{\kappa}$ is given as*

$$\boldsymbol{I}(\boldsymbol{\kappa}) = \frac{1}{\pi\sigma^2} \sum_{i=1}^{m} \left(\Gamma_i^R(\boldsymbol{\kappa}) \frac{\partial s_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} \left(\frac{\partial s_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^\top + \Gamma_i^I(\boldsymbol{\kappa}) \frac{\partial r_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} \left(\frac{\partial r_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^\top \right). \tag{2.18}$$

It is worth mentioning the FIM depends only on the quantization intervals, *not* the value of representatives. In contrast, the FIM using the unquantized measurements $\boldsymbol{z}$ is given as

$$\boldsymbol{I}_{\text{unquantized}}(\boldsymbol{\kappa}) = \frac{2}{\sigma^2} \sum_{i=1}^{m} \frac{\partial s_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} \left(\frac{\partial s_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^\top + \frac{\partial r_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}} \left(\frac{\partial r_i(\boldsymbol{\kappa})}{\partial \boldsymbol{\kappa}}\right)^\top \Big).$$

It remains to evaluate $\frac{\partial s_i(\boldsymbol{\kappa})}{\partial \kappa}$ and $\frac{\partial r_i(\boldsymbol{\kappa})}{\partial \kappa}$. Following the Wirtinger calculus [75], we have $\frac{\partial s_i(\boldsymbol{\kappa})}{\partial \kappa} = \frac{\partial \Re(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star \rangle)}{\partial \kappa} = \frac{1}{2}\Re\left(\boldsymbol{a}_i^{\mathsf{H}} \frac{\partial \boldsymbol{x}^\star}{\partial \kappa}\right)$ and $\frac{\partial r_i(\boldsymbol{\kappa})}{\partial \kappa} = \frac{\partial \Im(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star \rangle)}{\partial \kappa} = \frac{1}{2}\Im\left(\boldsymbol{a}_i^{\mathsf{H}} \frac{\partial \boldsymbol{x}^\star}{\partial \kappa}\right)$. Define

$$\boldsymbol{g}(f) = \frac{\partial \boldsymbol{v}(f)}{\partial f} = \left[0, j2\pi e^{j2\pi f}, \cdots, j2\pi(n-1)e^{j2\pi(n-1)f}\right]^{\mathsf{T}}.$$

Then, for each of the parameters in $\boldsymbol{\kappa}$, we have, for $k = 1, \ldots, K$,

$$\frac{\partial \boldsymbol{x}^\star}{\partial f_k} = c_k \boldsymbol{g}(f_k), \tag{2.19a}$$

$$\frac{\partial \boldsymbol{x}^\star}{\partial A_k} = e^{j2\pi\phi_k} \boldsymbol{v}(f_k), \tag{2.19b}$$

$$\frac{\partial \boldsymbol{x}^\star}{\phi_k} = j2\pi c_k \boldsymbol{v}(f_k). \tag{2.19c}$$

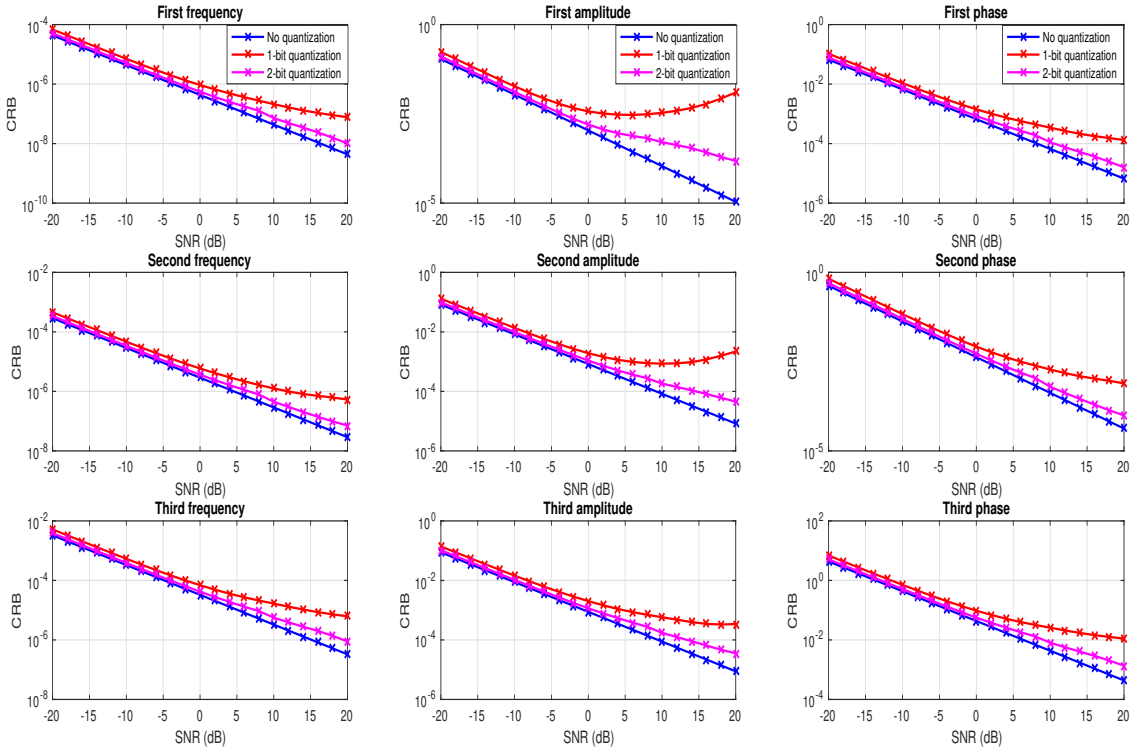


Figure 2.1: The CRB under different bit-depths with respect to SNR for a fixed number of measurements $m = 100$. Here, $n = 64$ and $K = 3$. Each row represents the CRB for estimating the frequency, amplitude and phase of one spectral atom.

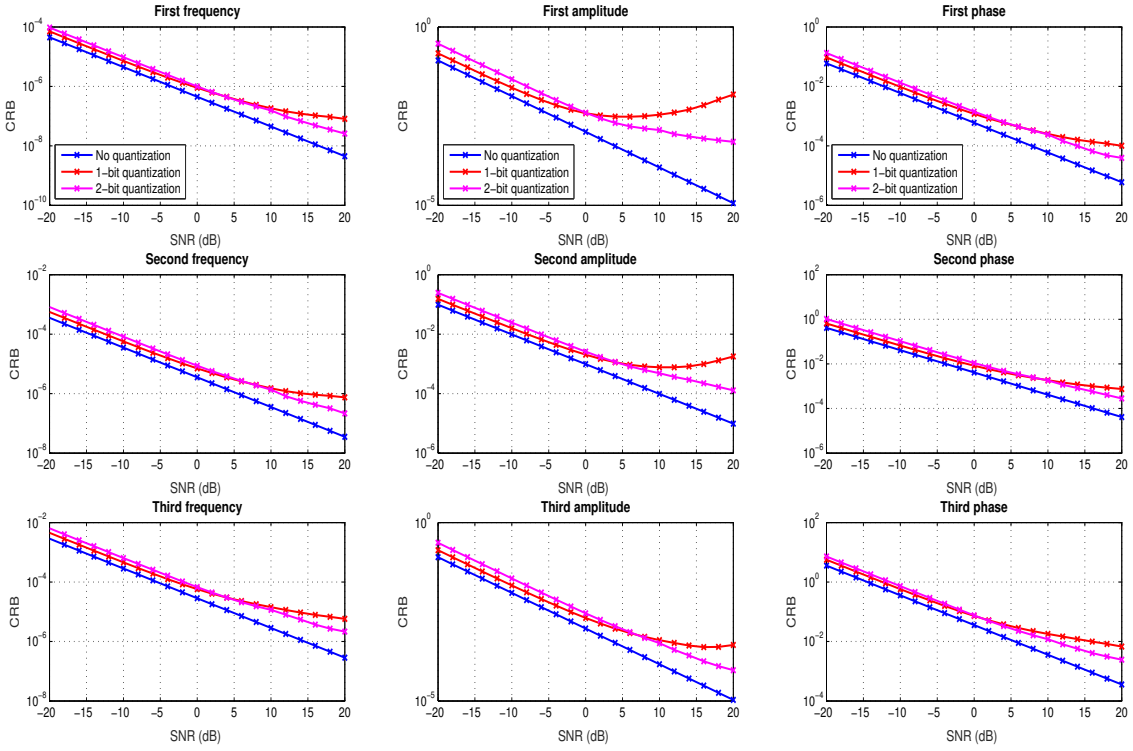


Figure 2.2: The CRB under different bit-depths with respect to SNR for a fixed number of bits $B = 100$. In this case, 2-bit quantization only has half the number of measurements of the 1-bit case. Here, $n = 64$ and $K = 3$. Each row represents the CRB for estimating the frequency, amplitude and phase of one spectral atom.

2.2.3 Numerical Evaluations of CRB

We now evaluate the CRB for 1-bit and 2-bit quantization schemes using the Lloyd's quantizer, and compare it against the CRB without quantization. We generate a spectrally-sparse signal $\boldsymbol{x}^\star$ of length $n = 64$ with frequencies $f_1 = 0.3$, $f_2 = 0.325$, and $f_3 = 0.8$, and complex coefficients $c_1 = 0.4e^{j2\pi\cdot 0.1}$, $c_2 = 0.15e^{j2\pi\cdot 0.55}$, and $c_3 = 0.05e^{j2\pi\cdot 0.75}$, which are selected arbitrarily.

We first fix the number of measurements as $m = 100$, and generate a measurement matrix with complex standard i.i.d. Gaussian entries. Fig. 2.1 shows the CRB for estimating all parameters with respect to the SNR, where it is defined as $\mathsf{SNR} = \|\boldsymbol{x}^\star\|_2^2/\sigma^2$. It is evident that increasing the bit depth improves the performance. In the low SNR regime performance is *noise-limited*, and behaves similarly as if there was no quantization; in the high SNR regime, performance is *quantization-limited*, and experiences severe performance degeneration due to quantization.

In many situations, we cannot simultaneously have high sample complexity and high bit depth, but rather, our budget is constrained by the number of total bits, which is $B = m \cdot b^\star = m \cdot \lceil \log_2 |\mathcal{D}| \rceil$. Therefore, it is useful to understand the trade-off between sample complexity and bit depth. Here, we use the CRB as a tool to compare the 1-bit and 2-bit quantization schemes. Fix the total number of bits as $B = 100$. In the 1-bit quantization scheme, we use a measurement matrix with $m = 100$ as generated earlier. In the 2-bit quantization scheme, we only use the first $m/2$ rows of the same measurement matrix. For comparison, we also plot the CRB assuming unquantized measurements using the same measurement matrix as the 1-bit case. Fig. 2.2 shows the CRB for estimating all parameters with respect to the SNR. It can be seen that in the low SNR regime, 1-bit quantization is preferred, as performance is

noise-limited, so higher sample complexity improves performance; in the high SNR regime, 2-bit quantization is preferred, as performance is quantization-limited, so higher bit depth improves performance. Our analysis is estimation-theoretic, and doesn't depend on the algorithm being adopted.

2.3 Atomic Norm Soft Thresholding for Quantized Spectral Compressed Sensing

It is well-known that maximum likelihood estimators approach the performance of CRB asymptotically at high SNR [21], however, their implementation requires exact knowledge of the likelihood function, which may not be available in certain applications. Therefore, in this section, we will develop estimators that do not require the knowledge of the quantization scheme using 1-bit measurements via atomic norm minimization [2]. We first provide the backgrounds on atomic norm for line spectrum estimation, and then describe the proposed algorithms for both the single vector case and the multiple vector case with performance guarantees.

2.3.1 Backgrounds on Atomic Norms

The atomic norm is originally proposed in [2] as a unified framework of convex regularizations for solving underdetermined linear inverse problems. Subsequently, [23–29] has tailored it to the estimation of spectrally-sparse signals. Most recently, [30] offered an extensive and comprehensive overview to atomic norm minimization as a canonical convex approach for super resolution.

For the single vector case, define the *atomic set* as

$$\mathcal{A}_s = \left\{ e^{j\phi} \boldsymbol{v}(f) : \ f \in [0,1), \phi \in [0, 2\pi) \right\},$$

then the atomic norm of a vector $\boldsymbol{x}$ is given as

$$\|\boldsymbol{x}\|_{\mathcal{A}} := \inf\{t > 0 : \ \boldsymbol{x} \in t \cdot \mathrm{conv}(\mathcal{A}_s)\} = \inf\left\{ \sum_i |\alpha_i| \ \Big| \ \boldsymbol{x} = \sum_i \alpha_i \boldsymbol{v}(f_i) \right\}, \quad (2.20)$$

where $\mathrm{conv}(\mathcal{A})$ denotes the convex hull of set $\mathcal{A}$. The atomic norm can be viewed as a continuous analog of the ℓ_1 norm over the continuous dictionary defined by the atomic set. Therefore, by promoting signals with small atomic norms, we encourage signals that can be expressed by a small number of spectral atoms. Appealingly, as shown in [23], it is possible to calculate $\|\boldsymbol{x}\|_{\mathcal{A}}$ using an equivalent semidefinite program, which can be computed efficiently using off-the-shelf solvers:

$$\|\boldsymbol{x}\|_{\mathcal{A}} = \min_{\boldsymbol{u}\in\mathbb{C}^n, w} \left\{ \frac{1}{2n}\mathrm{Tr}(\mathcal{T}(\boldsymbol{u})) + \frac{w}{2} \Big| \begin{bmatrix} \mathcal{T}(\boldsymbol{u}) & \boldsymbol{x} \\ \boldsymbol{x}^{\mathsf{H}} & w \end{bmatrix} \succeq 0 \right\},$$

where $\mathcal{T}(\boldsymbol{u})$ denotes the Hermitian Toeplitz matrix with $\boldsymbol{u}$ as the first column. The dual atomic norm $\|\cdot\|_{\mathcal{A}}^*$ for a vector $\boldsymbol{q} \in \mathbb{C}^n$, as will become useful later, is given as

$$\|\boldsymbol{q}\|_{\mathcal{A}}^* = \sup_{\|\boldsymbol{x}\|_{\mathcal{A}} \le 1} \langle \boldsymbol{q}, \boldsymbol{x} \rangle_{\mathbb{R}} = \sup_{f\in[0,1]} |\boldsymbol{q}^{\mathsf{H}} \boldsymbol{v}(f)|,$$

where the second equality follows from the fact the the extreme values are taken when $\boldsymbol{x}$ is aligned with $\boldsymbol{v}(f)$ due to convexity. From the above equation it is clear that $\|\boldsymbol{q}\|_{\mathcal{A}}^*$ can be interpreted as the largest absolute value of a polynomial of $e^{j2\pi f}$, denoted as $Q(f) = |\boldsymbol{q}^{\mathsf{H}} \boldsymbol{v}(f)|$.

2.3.2 Atomic Soft-Thresholding with Quantized Measurements

We first construct a surrogate signal from the quantized measurements as [37]

$$\boldsymbol{s} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i = \frac{1}{m}\boldsymbol{A}^{\mathsf{H}}\boldsymbol{y} \in \mathbb{C}^n, \quad (2.21)$$

and use the following atomic norm soft-thresholding (AST) algorithm to estimate the signal $\boldsymbol{x}$,

$$\hat{\boldsymbol{x}} = \mathrm{argmin}_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{s}\|_2^2 + \tau\|\boldsymbol{x}\|_{\mathcal{A}}, \quad (2.22)$$

which is the proximal mapping of the surrogate signal $\boldsymbol{s}$ with respect to the atomic norm, where $\tau > 0$ is a regularization parameter. One appealing feature of atomic norm minimization is that the set of frequencies can be recovered via the dual polynomial approach [29]. Namely, denote the dual variable as $\hat{\boldsymbol{q}} = (\boldsymbol{s} - \hat{\boldsymbol{x}})/\tau$, and $Q(f) = |\hat{\boldsymbol{q}}^{\mathsf{H}}\boldsymbol{v}(f)|$. Then the set of frequencies can be localized as $\hat{\mathcal{F}} = \{f : Q(f) = 1\}$. We refer interested readers to the details in [23]. Alternatively, the frequencies can be localized via performing conventional subspace methods using the estimated signal.

2.3.3 Performance Guarantees

In this section, we develop performance guarantees of the proposed AST algorithm under 1-bit quantization in the single vector case using the sign quantizer in (2.6). Note that in this case, it can be seen that $\boldsymbol{s}$ in (2.21) is an unbiased estimator of $\boldsymbol{x}^\star$ up to a scaling difference, i.e.

$$\mathbb{E}[\boldsymbol{s}] = \lambda \frac{\boldsymbol{x}^\star}{\|\boldsymbol{x}^\star\|_2},$$

where

$$\lambda = \frac{2\|\boldsymbol{x}^\star\|_2}{\sqrt{\pi(\sigma^2 + \|\boldsymbol{x}^\star\|_2^2)}} = \frac{2}{\sqrt{\pi(1/\mathsf{SNR} + 1)}} \tag{2.23}$$

depends on the SNR before quantization $\mathsf{SNR} = \|\boldsymbol{x}^\star\|_2^2/\sigma^2$. To illustrate, Fig. 2.3 depicts λ as a function of SNR, which is a monotonically increasing function with respect to SNR and approaches to the limit $2/\sqrt{\pi}$ as SNR goes to infinity.

Without loss of generality, we assume $\|\boldsymbol{x}^\star\|_2 = 1$. The performance of AST relies critically on the separation condition, which is defined as the minimum distance between distinct frequencies,

$$\Delta = \min_{k \neq j} |f_k - f_j| \geq \frac{4}{n}, \tag{2.24}$$

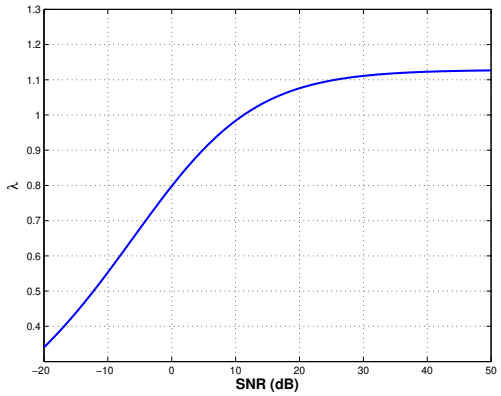


Figure 2.3: The value of λ with respect to SNR before quantization.

where $|f_k - f_j|$ is evaluated as the wrap-around difference on the unit modulus. Under the separation condition, we have the performance guarantee of the proposed algorithm in (2.22), stated below.

Theorem 2. *Set $\tau := \eta\sqrt{n \log n/m}$ for some constant $\eta \geq 1$. Under the separation condition, the solution $\hat{\boldsymbol{x}}$ satisfies*

$$\left\| \frac{\hat{\boldsymbol{x}}}{\lambda} - \boldsymbol{x}^\star \right\|_2 \lesssim \frac{1}{\lambda}\sqrt{\frac{K \log n}{m}}$$

with high probability.

The proof of Theorem 2 can be found in Appendix A.1. Theorem 2 suggests that the proposed algorithm accurately recovers the signal as soon as m is on the order of $K \log n$, which is order-wise near-optimal, since at least an order of $K \log(n/K)$ measurements are needed in order to recover a sparse signal in the DFT basis [35]. Moreover, the theorem also suggests that the normalized reconstruction error is inverse proportional to λ, which plays the role of SNR *after quantization* and is a

nonlinear function of the SNR before quantization. In the low SNR regime, λ scales as $1/\sqrt{\mathsf{SNR}}$, and the performance is comparable to that using unquantized measurements. However, in the high SNR regime, there is a saturation phenomenon, as evidenced by Fig. 2.3, and the performance does not improve as much with we increase SNR, which is also corroborated by numerical simulation in Section 2.5. These results are qualitatively in line with existing work on one-bit CS [35].

Remark: More generally, Theorem 2 can be extended to the generalized linear model following similar strategies in [34], as long as the 1-bit measurements y_i's are i.i.d. and satisfy $\mathbb{E}[y_i|\boldsymbol{a}_i] = g\left(\langle \boldsymbol{a}_i, \boldsymbol{x}^\star \rangle\right)$ for some link function $g(\cdot)$, and accordingly $\lambda = \mathbb{E}[g(\theta)\theta^{\mathsf{H}}]$ where the expectation is taken with respect to $\theta \sim \mathcal{CN}(0,1)$. This allows us to model other complex quantization schemes with non-Gaussian noise.

2.4 Extension to the Multiple Vector Case

In many applications, we encounter an ensemble of line spectrum signals, where each signal $\boldsymbol{x}_t \in \mathbb{C}^n$ contains a linear combination of spectral lines with the same set of frequencies $\mathcal{F}$, but with varying amplitudes, given as

$$\boldsymbol{x}_t^\star = \sum_{k=1}^{K} c_{k,t} \boldsymbol{v}\left(f_k\right), \quad 1 \le t \le T,$$

where $c_{k,t} \in \mathbb{C}$, and T is the number of snapshots. Denote $\boldsymbol{X}^\star = [\boldsymbol{x}_1^\star, \boldsymbol{x}_2^\star, \ldots, \boldsymbol{x}_T^\star] \in \mathbb{C}^{n\times T}$ as the signal ensemble. Similar to (2.3), the CS measurement of each snapshot is given as

$$\boldsymbol{z}_t = \boldsymbol{A}\boldsymbol{x}_t^\star + \sigma\boldsymbol{\epsilon}_t, \tag{2.25}$$

where $\boldsymbol{\epsilon}_t = [\epsilon_{1,t}, \epsilon_{2,t}, \ldots, \epsilon_{m,t}]^{\mathsf{T}}$ contains i.i.d. standard complex Gaussian $\mathcal{CN}(0,1)$ entries. Similar to (2.5), the quantized measurements of each $\boldsymbol{z}_t$ is then given as

$$\boldsymbol{y}_t = \mathcal{Q}(\boldsymbol{z}_t). \tag{2.26}$$

Denote $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_T]$ and $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ as the unquantized measurement ensemble and the quantized measurement ensemble, respectively. Our goal is then to recover $\boldsymbol{X}^\star$ and the set of frequencies from $\boldsymbol{Y}$, without assuming the knowledge of the sparsity level and the quantizer. The presence of multiple vectors can significantly improve the accuracy of frequency estimation.

It is possible to extend the atomic norm formulation to the multiple vector case [25]. Define the atomic set as

$$\mathcal{A}_m = \left\{ \boldsymbol{A}\left(f, \boldsymbol{b}\right) = \boldsymbol{v}\left(f\right)\boldsymbol{b} | f \in (0,1], \boldsymbol{b} \in \mathbb{C}^{1\times T}, \|\boldsymbol{b}\|_2 = 1 \right\},$$

then the atomic norm is defined as

$$\begin{aligned} \|\boldsymbol{X}\|_{\mathcal{A}} &= \inf\left\{t > 0 : \boldsymbol{X} \in t \cdot \operatorname{conv}\left(\mathcal{A}_m\right)\right\} \\ &= \inf\left\{ \sum_k |c_k| \,\middle|\, \boldsymbol{X} = \sum_k c_k \boldsymbol{A}\left(f_k, \boldsymbol{b}_k\right) \right\}, \end{aligned}$$

which can be computed similarly via solving the following semidefinite program [25]:

$$\|\boldsymbol{X}\|_{\mathcal{A}} = \min_{\boldsymbol{u}\in\mathbb{C}^n, \boldsymbol{W}\in\mathbb{C}^{T\times T}} \left\{ \frac{1}{2n}\operatorname{Tr}(\mathcal{T}(\boldsymbol{u})) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{W}) \middle| \begin{bmatrix} \mathcal{T}(\boldsymbol{u}) & \boldsymbol{X} \\ \boldsymbol{X}^{\mathsf{H}} & \boldsymbol{W} \end{bmatrix} \succeq 0 \right\}.$$

The dual norm for some $\boldsymbol{Q} \in \mathbb{C}^{n\times T}$ is given as

$$\|\boldsymbol{Q}\|_{\mathcal{A}}^* = \sup_{\|\boldsymbol{X}\|_{\mathcal{A}}\le 1} \langle \boldsymbol{Q}, \boldsymbol{X}\rangle_{\mathbb{R}} = \sup_{f\in[0,1]} \|\boldsymbol{Q}^{\mathsf{H}}\boldsymbol{v}(f)\|_2,$$

which is the largest absolute value of the polynomial $Q(f) = \|\boldsymbol{Q}^{\mathsf{H}}\boldsymbol{v}(f)\|_2$.

For reconstruction, we construct the surrogate signal ensemble from the quantized measurement ensemble $\boldsymbol{Y}$ as

$$\boldsymbol{S} = \frac{1}{m}\boldsymbol{A}^{\mathsf{H}}\boldsymbol{Y} \in \mathbb{C}^{n\times T}, \tag{2.27}$$

and use the following atomic norm soft-thresholding (AST) algorithm to estimate the signal ensemble $\boldsymbol{X}$,

$$\hat{\boldsymbol{X}} = \text{argmin}_{\boldsymbol{X}\in\mathbb{C}^{n\times T}} \|\boldsymbol{X} - \boldsymbol{S}\|_{\text{F}}^2 + \tau_T \|\boldsymbol{X}\|_{\mathcal{A}}, \tag{2.28}$$

where $\tau_T > 0$ is a regularization parameter. Moreover, define $\hat{\boldsymbol{Q}} = (\boldsymbol{S} - \hat{\boldsymbol{X}})/\tau_T$, and $Q(f) = \|\hat{\boldsymbol{Q}}^{\mathsf{H}}\boldsymbol{v}(f)\|_2$. Then the set of frequencies can be localized as $\hat{\mathcal{F}} = \{f : Q(f) = 1\}$. Alternatively, the frequencies can be localized via performing conventional subspace methods using the estimated snapshots.

2.5 Numerical Experiments

In this section, we conduct numerical experiments to evaluate the performance of the proposed AST algorithms for parameter estimation using quantized compressive measurements in both the single vector case and the multiple vector case. For implementation of the AST algorithms, we used the CVX toolbox [76]. There're several other fast solvers developed for atomic norm minimization that are more scalable to large problems, including ADMM [25, 29], ADCG [77], and CoGent [78], to name a few.

2.5.1 Single Vector Case

Let $n = 64$ and $K = 3$. The set of frequencies is located at $\boldsymbol{f} = \{0.3, 0.325, 0.8\}$, where the first two frequencies are separated barely more than $1/n$, the Rayleigh

limit. The number of bits is set as $m = 1000$, where the measurement vectors are generated with i.i.d. $\mathcal{CN}(0, 1)$ entries. The measurements are quantized according to (2.6). Fig. 2.4 shows the amplitude of the constructed dual polynomial by solving (2.22), where its peaks can be used to localize the frequencies. It can be seen that it matches accurately with the ground truth.

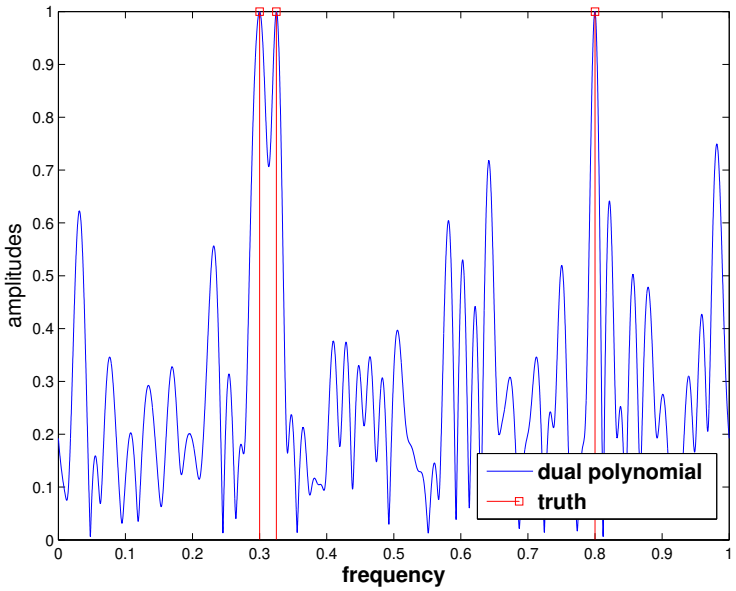


Figure 2.4: Frequency localization via peaks of the dual polynomial, superimposed on the ground truth.

Next, we compare the performance of signal reconstruction using atomic norm with unquantized measurements $\boldsymbol{z}$, by running the algorithm:

$$\hat{\boldsymbol{x}}_{\mathsf{UQ}} = \text{argmin}_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \tilde{\tau}\|\boldsymbol{x}\|_{\mathcal{A}},$$

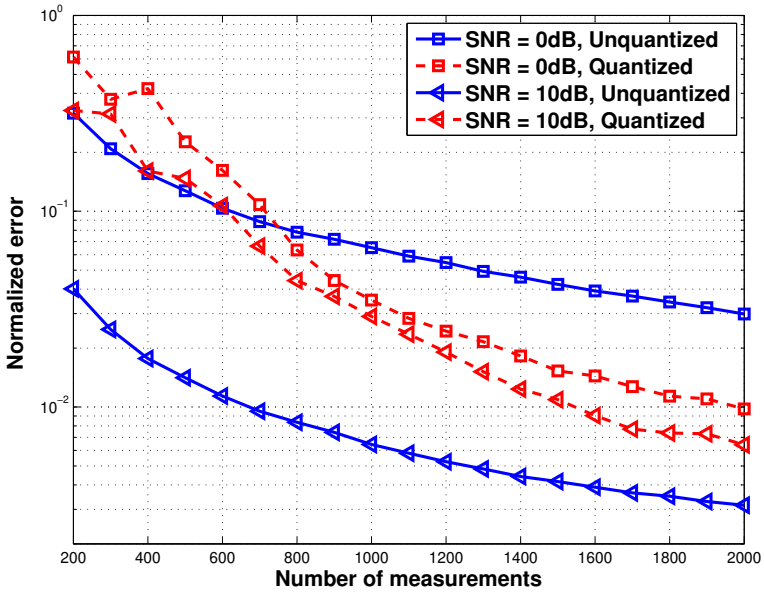


Figure 2.5: Normalized reconstruction error with respect to the number of measurements at different SNRs with or without quantization.

where $\tilde{\tau}$ is a properly tuned regularization parameter. The normalized reconstruction error is defined as $\sin^2(\angle\hat{\boldsymbol{x}}, \boldsymbol{x}) = 1 - |\langle\hat{\boldsymbol{x}}, \boldsymbol{x}^\star\rangle|^2/(\|\hat{\boldsymbol{x}}\|_2^2\|\boldsymbol{x}^\star\|_2^2)$, where $\hat{\boldsymbol{x}}$ is the reconstructed signal using either algorithm. Fig. 2.5 shows the normalized reconstruction error at different SNRs with comparisons to that using the quantized measurements and the AST algorithm (2.22), where SNR is defined again as $\mathsf{SNR} = \|\boldsymbol{x}^\star\|_2^2/\sigma^2$. It can be seen that the reconstruction accuracy improves as we increase the SNR as well as the number of measurements, validating the theoretical analysis. In particular, at low SNR, using quantized measurements can potentially achieve better reconstruction quality with much fewer measurement budgets in bits. It can also be seen that

improving the SNR before quantization does not have as strong impact as for the unquantized case.

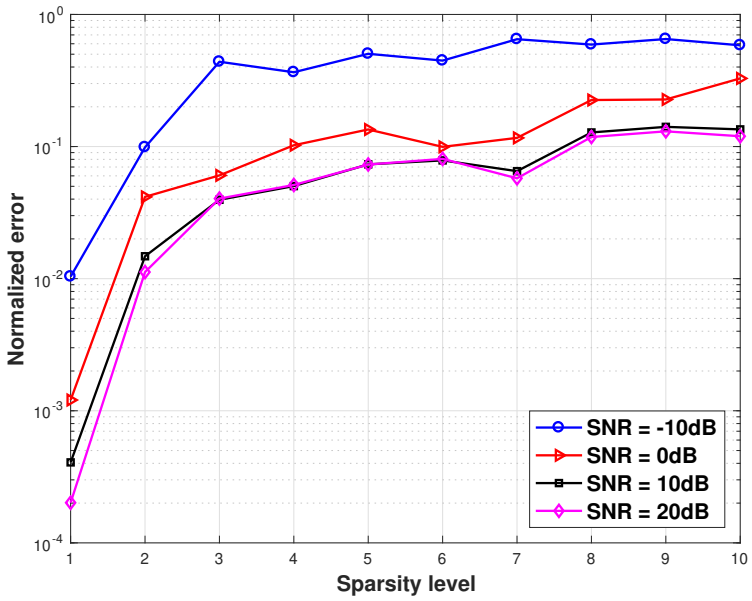


Figure 2.6: Normalized reconstruction error with respect to the spectral sparsity level at different SNRs before quantization.

Next, we examine the performance of the proposed algorithm as a function of the spectral sparsity level. Fix $n = 64$ and $m = 1000$. At each run, we randomly generate K different frequencies that satisfy the separate condition. Fig. 2.6 shows the normalized reconstruction error as a function of the sparsity level at various SNR, averaged over 200 Monte Carlo simulations. It can be seen that the reconstruction error is higher when the spectral sparsity level is higher, and the SNR is lower. Moreover, it can be seen that the reconstruction error stops to decrease when the

SNR is relatively high, indicating a saturation effect due to quantization, as predicted by our theory.

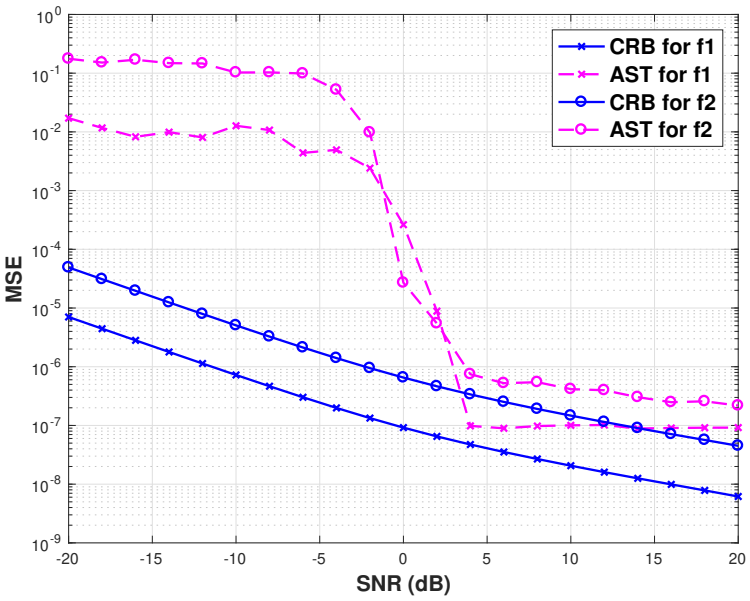
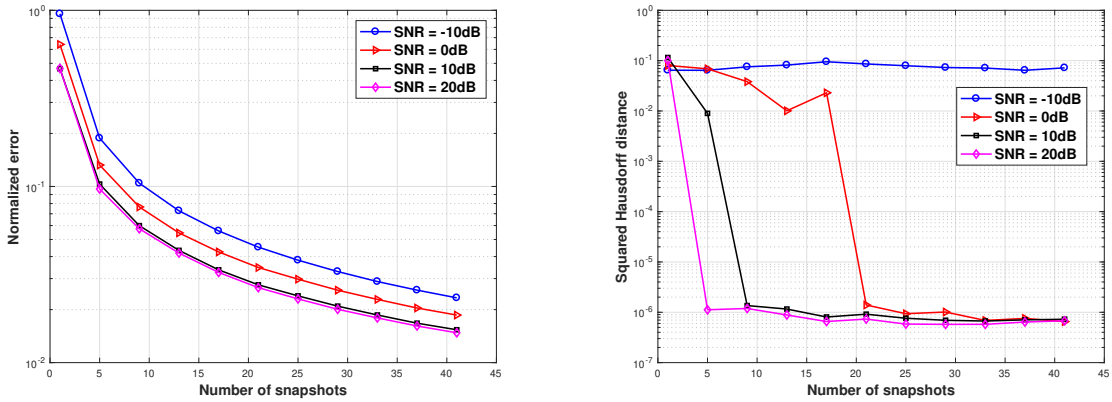


Figure 2.7: Mean square error of frequency localization with respect to SNR using 1-bit measurements, CRB is provided as a benchmark: (a) first frequency; (b) second frequency.

We further compare the performance of frequency localization using the proposed algorithm with the CRB. Fix $n = 64$ and $m = 1000$. We generate the ground signal with frequencies $f_1 = 0.3$, $f_2 = 0.325$ and amplitudes $c_1 = 0.4e^{j2\pi\cdot 0.1}$, $c_2 = 0.15e^{j2\pi\cdot 0.55}$. Fig. 2.7 shows the average mean squared error for each frequency over 200 Monte Carlo simulations, against the corresponding CRB calculated using the formulas in Section 2.2. The frequencies are estimated by using the MATLAB function `rootmusic` by assuming the correct model order, that is $K = 2$. The performance of the proposed

algorithm exhibits a threshold effect where it approaches that of CRB as soon as SNR is large enough. However, further increasing the SNR doesn't seem to improve the performance, which coincides with the saturation effect discussed earlier.



(a) (b)

Figure 2.8: Performance with respect to the number of snapshots at different SNRs using 1-bit measurements: (a) signal reconstruction error; (b) frequency estimation error measured in Hausdorff distance.

2.5.2 Multiple Vector Case

We evaluate the performance of the AST algorithm (2.28) in the multiple vector case. We follow the same setup as Fig. 2.4, where $n = 64$, the set of frequencies $\boldsymbol{f} = \{0.3, 0.325, 0.8\}$, and the number of measurements for each snapshot is $m = 50$. The coefficients of each snapshot in $\boldsymbol{X}$ is generated independently using the standard complex Gaussian distribution. The SNR *per snapshot* is defined as $\mathsf{SNR} = \|\boldsymbol{X}\|_{\mathrm{F}}^2/(T\sigma^2)$, where T is the number of snapshots. We set the regularization

parameter $\tau_T = \sqrt{n \log n/(10 \cdot mT)}$ in the experiment. The normalized reconstruction error is defined as $\sin^2(\angle \boldsymbol{X}; \hat{\boldsymbol{X}})$, where $\hat{\boldsymbol{X}}$ is the recovered signal containing multiple snapshots, and $\angle$ denotes the angle between the subspace spanned by $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$. Once $\hat{\boldsymbol{X}}$ is obtained, we estimate the frequencies by using the MATLAB function `rootmusic` by assuming the correct model order, that is $K = 3$. The accuracy of frequency estimation is evaluated by examining the Hausdorff distance between the recovered frequencies $\hat{\boldsymbol{f}}$ and the ground truth $\boldsymbol{f}$ as

$$d_H(\boldsymbol{f}, \hat{\boldsymbol{f}}) = \max \left\{ \sup_{f \in \boldsymbol{f}} \inf_{\hat{f} \in \hat{\boldsymbol{f}}} \|f - \hat{f}\|_2, \sup_{\hat{f} \in \hat{\boldsymbol{f}}} \inf_{f \in \boldsymbol{f}} \|f - \hat{f}\|_2 \right\}.$$

Fig. 2.8 shows the recovery performance with respect to the number of snapshots at different SNRs, averaged over 50 Monte Carlo simulations, where (a) depicts the normalized reconstruction error, and (b) depicts the squared Hausdorff distance. At a fixed SNR, it can be seen that both the normalized reconstruction error and frequency estimation error reduce, highlighting the benefit of having multiple snapshots. In particular, having multiple snapshots allows better frequency recovery once the number of snapshots is large enough. Moreover, performance improves as we increase the SNR.

Chapter 3: Learning One-Hidden-Layer Neural Network for Binary Classification

In this chapter, we study model recovery for data classification, where the training labels are generated from a one-hidden-layer neural network with sigmoid activations, and the goal is to recover the weights of the neural network. Considering the multi-neuron classification problem with either FCN or CNN, we first analyze the landscape of the cross-entropy loss and show that gradient descent converges linearly to a critical point, which is shown to exist. Due to the quantized nature of labels, the recovery of the ground truth is only up to certain statistical accuracy. At last, we show that a tensor method provably provides an initialization in the neighborhood of the ground truth both for FCN and CNN.

We adopt additional notations in this chapter as follows. Denote $\|\cdot\|_{\psi_1}$ as the sub-exponential norm of a random variable. For nonnegative functions $f(x)$ and $g(x)$, $f(x) = O\left(g(x)\right)$ means there exist positive constants c and a such that $f(x) \leq cg(x)$ for all $x \geq a$; $f(x) = \Omega\left(g(x)\right)$ means there exist positive constants c and a such that $f(x) \geq cg(x)$ for all $x \geq a$.

3.1 Problem Formulation

We consider two popular types of one-hidden-layer nonlinear neural networks illustrated in Fig. 3.1, i.e., a FCN [7] and a non-overlapping CNN [71]. For both cases, we let $\boldsymbol{x} \in \mathbb{R}^d$ be the input, $K \geq 1$ be the number of neurons, and the activation function be the sigmoid function

$$\phi(x) = \frac{1}{1+\exp(-x)}.$$

- *FCN:* the network parameter is $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K] \in \mathbb{R}^{d\times K}$, and

$$H_{\text{FCN}}(\boldsymbol{W}, \boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} \phi(\boldsymbol{w}_k^\top \boldsymbol{x}). \tag{3.1}$$

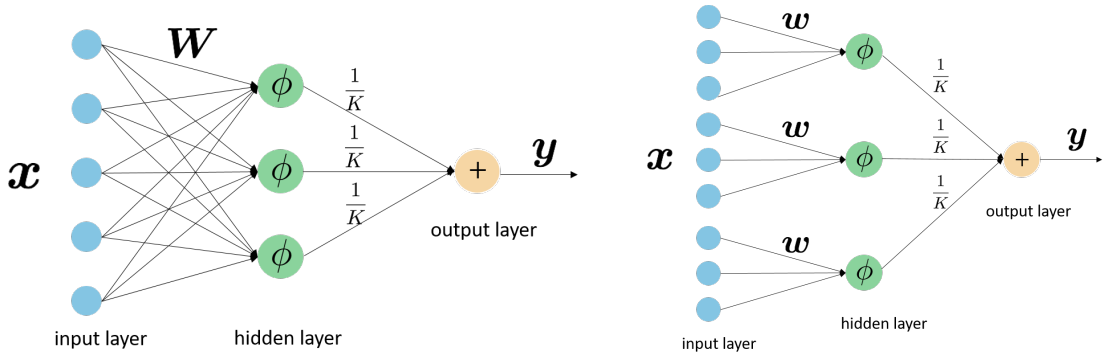
- *Non-overlapping CNN:* for simplicity we let $d = m \cdot K$ for some integers m. Let $\boldsymbol{w} \in \mathbb{R}^m$ be the network parameter, and the kth stride of $\boldsymbol{x}$ be given as $\boldsymbol{x}^{(k)} = \left[x_{m(k-1)+1}, \cdots x_{m\cdot k}\right]^\top \in \mathbb{R}^m$. Then,

$$H_{\text{CNN}}(\boldsymbol{w}, \boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} \phi(\boldsymbol{w}^\top \boldsymbol{x}^{(k)}). \tag{3.2}$$

The non-overlapping CNN model can be viewed as a highly structured instance of the FCN, where the weight matrix can be written as:

$$\boldsymbol{W}_{\text{CNN}} = \begin{bmatrix} \boldsymbol{w} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{w} & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{w} \end{bmatrix} \in \mathbb{R}^{d\times K}.$$

In a model recovery setting, we are given n training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim (\boldsymbol{x}, y)$ that are drawn i.i.d. from certain distribution regarding the ground truth network parameter $\boldsymbol{W}^\star$ (or resp. $\boldsymbol{w}^\star$ for CNN). Suppose the network input $\boldsymbol{x} \in \mathbb{R}^d$ is drawn from a standard Gaussian distribution $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. This assumption has been used



(a) FCN (b) CNN

Figure 3.1: Illustration of two types of one-hidden-layer neural networks considered in this Chapter: (a) a fully-connected network (FCN); (b) a non-overlapping convolutional neural network (CNN).

a lot in previous literature [48,68,71,72], to name a few. Then, conditioned on $\boldsymbol{x} \in \mathbb{R}^d$, the output y is mapped to $\{0, 1\}$ via the output of the neural network, i.e.,

$$\mathbb{P}\left(y = 1 | \boldsymbol{x}\right) = H\left(\boldsymbol{W}^\star, \boldsymbol{x}\right). \tag{3.3}$$

Our goal is to recover the network parameter, i.e., $\boldsymbol{W}^\star$, via minimizing the following empirical loss function:

$$f_n(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(\boldsymbol{W}; \boldsymbol{x}_i\right), \tag{3.4}$$

where $\ell\left(\boldsymbol{W}; \boldsymbol{x}\right) := \ell\left(\boldsymbol{W}; \boldsymbol{x}, y\right)$ is the cross-entropy loss function, i.e.,

$$\ell\left(\boldsymbol{W}; \boldsymbol{x}\right) = -y \cdot \log\left(H\left(\boldsymbol{W}, \boldsymbol{x}\right)\right) - (1 - y) \cdot \log\left(1 - H\left(\boldsymbol{W}, \boldsymbol{x}\right)\right), \tag{3.5}$$

where $H(\boldsymbol{W}, \boldsymbol{x})$ can subsume either H_{FCN} or H_{CNN}.

3.2 Gradient Descent and its Performance Guarantee

To estimate the network parameter $\boldsymbol{W}^\star$, since (3.4) is a highly nonconvex function, vanilla gradient descent with an arbitrary initialization may get stuck at local minima. Therefore, we implement gradient descent (GD) with a well-designed initialization scheme that is described in details in Section 3.3. In this section, we focus on the performance of the local update rule

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \nabla f_n\left(\boldsymbol{W}_t\right),$$

where η is the constant step size. The algorithm is summarized in Algorithm 1.

Algorithm 1 Gradient Descent (GD)

Input: Training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, step size η, iteration T
Initialization: $\boldsymbol{W}_0 \leftarrow \textsc{Initialization}\left(\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n\right)$
Gradient Descent: for $t = 0, 1, \cdots, T-1$

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \nabla f_n\left(\boldsymbol{W}_t\right).$$

Output: $\boldsymbol{W}_T$

Note that throughout the execution of GD, the same set of training samples is used which is the standard implementation of gradient descent. Consequently the analysis is challenging due to the statistical dependence of the iterates with the data.

3.2.1 Uniform local strong convexity

We first characterize the local strong convexity of $f_n(\cdot)$ in a neighborhood of the ground truth. We use the Euclidean ball to denote the local neighborhood of $\boldsymbol{W}^\star$ for

FCN or of $\boldsymbol{w}^\star$ for CNN.

$$\mathbb{B}\left(\boldsymbol{W}^\star, r\right) = \left\{\boldsymbol{W} \in \mathbb{R}^{d\times K} : \|\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}} \le r\right\}, \tag{3.6a}$$

$$\mathbb{B}\left(\boldsymbol{w}^\star, r\right) = \left\{\boldsymbol{w} \in \mathbb{R}^{m} : \|\boldsymbol{w} - \boldsymbol{w}^\star\|_{2} \le r\right\}, \tag{3.6b}$$

where r is the radius of the ball. With slight abuse of notations, we will drop the subscript FCN or CNN for simplicity, whenever it is clear from the context that the result is for FCN when the argument is $\boldsymbol{W} \in \mathbb{R}^{d\times K}$ and for CNN when the argument is $\boldsymbol{w} \in \mathbb{R}^m$. Further, $\sigma_i\left(\boldsymbol{W}\right)$ denotes the i-th largest singular value of $\boldsymbol{W}^\star$. Let the condition number be $\kappa = \sigma_1/\sigma_K$, and $\lambda = \prod_{i=1}^{K}\left(\sigma_i/\sigma_K\right)$. Moreover, we introduce an important quantity $\rho\left(\sigma\right)$ regarding $\phi(z)$, the sigmoid activation function, that captures the geometric properties of the loss function for neural networks (3.1) and (3.2).

Definition 1 (Key quantity for FCN)**.** *Let* $z \sim \mathcal{N}\left(0,1\right)$ *and define* $\alpha_q(\sigma) = \mathbb{E}[\phi'(\sigma \cdot z)z^q], \forall q \in \{0,1,2\}$*, and* $\beta_q(\sigma) = \mathbb{E}[\phi'(\sigma \cdot z)^2 z^q], \forall q \in \{0,2\}$*. Define* $\rho_{\mathrm{FCN}}(\sigma)$ *as*

$$\rho_{\mathrm{FCN}}(\sigma) = \min\left\{\beta_0(\sigma) - \alpha_0^2(\sigma), \beta_2(\sigma) - \alpha_2^2(\sigma)\right\} - \alpha_1^2(\sigma).$$

Definition 2 (Key quantity for CNN)**.** *Let* $z \sim \mathcal{N}\left(0,\sigma^2\right)$ *and define* $\rho_{\mathrm{CNN}}(\sigma)$ *as*

$$\rho_{\mathrm{CNN}}(\sigma) = \min\left\{\mathbb{E}[(\phi'(z)z)^2], \mathbb{E}[\phi'\left(z\right)^2]\right\}.$$

Note that Definition 1 for FCN is different from that in [7, Property 3.2] but consistent with [7, Lemma D.4] which removes the third term in [7, Property 3.2]. For the activation function considered in this Chapter, the first two terms suffice. Definition 2 for CNN is a newly distilled quantity in this Chapter tailored to the special structure of CNN.

The quantity $\rho(\sigma)$ plays an important role in the following theorem which guarantees the Hessian of the empirical risk function in the local neighborhood of the ground truth is positive definite with high probability for both FCN and CNN.

Theorem 3 (Local Strong Convexity)**.** *Consider the classification model with FCN* (3.1) *or CNN* (3.2) *and the sigmoid activation function.*

- *For FCN, assume* $\|\boldsymbol{w}_k^\star\|_2 \leq 1$ *for all* k*. There exist constants* c_1 *and* c_2 *such that as soon as sample size*

$$n_{\mathrm{FCN}} \geq c_1 \cdot dK^5 \log^2 d \cdot \left(\frac{\kappa^2\lambda}{\rho_{\mathrm{FCN}}(\sigma_K)}\right)^2,$$

with probability at least $1 - d^{-10}$*, we have for all* $\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})$*,*

$$\Omega\left(\frac{1}{K^2} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2\lambda}\right) \cdot \boldsymbol{I} \preceq \nabla^2 f_n(\boldsymbol{W}) \preceq \Omega(1) \cdot \boldsymbol{I},$$

where $r_{\mathrm{FCN}} := \frac{c_2}{\sqrt{K}} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2\lambda}$*.*

- *For CNN, assume* $\|\boldsymbol{w}^\star\|_2 \leq 1$*. There exist constants* c_3 *and* c_4 *such that as soon as sample size*

$$n_{\mathrm{CNN}} \geq c_3 \cdot dK^5 \log^2 d \cdot \left(\frac{1}{\rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)}\right)^2,$$

with probability at least $1 - d^{-10}$*, we have for all* $\boldsymbol{w} \in \mathbb{B}(\boldsymbol{w}^\star, r_{\mathrm{CNN}})$*,*

$$\Omega\left(\frac{1}{K} \cdot \rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)\right) \cdot \boldsymbol{I} \preceq \nabla^2 f_n(\boldsymbol{w}) \preceq \Omega(K) \cdot \boldsymbol{I},$$

where $r_{\mathrm{CNN}} := \frac{c_4}{K^2} \cdot \rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)$*.*

We note that for FCN (3.1), all column permutations of $\boldsymbol{W}^\star$ are equivalent global minimum of the loss function, and Theorem 3 applies to all such permutation matrices of $\boldsymbol{W}^\star$. The proof of Theorem 3 is outlined in Appendix B.2.

A pivot observation from the lower bound of the Hessian is that the sign of $\rho(\cdot)$ will determine whether the Hessian is positive definite or not, since K, κ, λ are all positive. We depict $\rho(\sigma)$ as a function of σ in a certain range for the sigmoid activation in Fig. 3.2. It can be seen from Fig. 3.2 that $\rho(\sigma)$ is monotonic increasing when σ increases, and we have $\rho(\sigma) > 0$ as long as $\sigma > 0$. When $\boldsymbol{W}^\star$ is orthogonal, κ and λ are both 1, $\rho(\sigma)$ is a constant, hence the lower bound of Hessian is on the order of $\frac{1}{K^2}$ for FCN. However, in the worst case where the columns of $\boldsymbol{W}^\star$ is linear dependent, then κ, λ, $\rho(\sigma)$ are infinite, and the local strong convexity doesn't hold for FCN case. Furthermore, the value of $\rho_{\mathrm{CNN}}(\sigma)$ is much larger than $\rho_{\mathrm{FCN}}(\sigma)$ for the same input.

Theorem 3 guarantees that for both FCN (3.1) and CNN (3.2) the Hessian of the empirical cross-entropy loss function $f_n(\boldsymbol{W})$ is positive definite in a neighborhood of the ground truth $\boldsymbol{W}^\star$, as long as the sample size n is sufficiently large and the columns of $\boldsymbol{W}^\star$ are linearly independent. The bounds in Theorem 3 depend on the dimension parameters of the network (n and K), as well as the ground truth ($\rho_{\mathrm{FCN}}(\sigma_K)$, λ, $\rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)$).

3.2.2 Performance Guarantees of GD

For the classification problem, due to the nature of quantized labels, $\boldsymbol{W}^\star$ is no longer a critical point of $f_n(\boldsymbol{W})$. By the strong convexity of the empirical risk function $f_n(\boldsymbol{W})$ in the local neighborhood of $\boldsymbol{W}^\star$, there can exist at most one critical point in $\mathbb{B}(\boldsymbol{W}^\star, r)$, which is the unique local minimizer in $\mathbb{B}(\boldsymbol{W}^\star, r)$ if it exists. The following theorem shows that there indeed exists such a critical point $\widehat{\boldsymbol{W}}_n$, which is provably close to the ground truth $\boldsymbol{W}^\star$, and gradient descent converges linearly to $\widehat{\boldsymbol{W}}_n$.

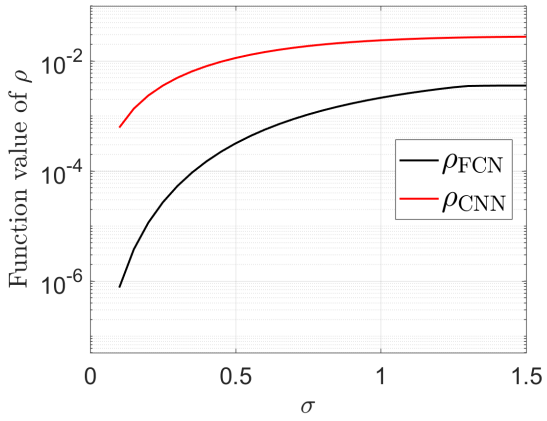


Figure 3.2: Illustration of $\rho(\sigma)$ for both FCN and CNN with the sigmoid activation.

Theorem 4 (Performance Guarantees of Gradient Descent)**.** *Assume the assumptions in Theorem 3 hold. Under the event that local strong convexity holds,*

- *for FCN, there exists a critical point in* $\mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})$ *such that*

$$\left\|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\right\|_{\mathrm{F}} \leq c_1 \frac{K^{9/4}\kappa^2\lambda}{\rho_{\mathrm{FNN}}(\sigma_K)}\sqrt{\frac{d\log n}{n}},$$

  and if the initial point $\boldsymbol{W}_0 \in \mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})$*, GD converges linearly to* $\widehat{\boldsymbol{W}}_n$*, i.e.*

$$\left\|\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n\right\|_{\mathrm{F}} \leq \left(1 - \frac{c_2\eta\rho_{\mathrm{FCN}}(\sigma_K)}{K^2\kappa^2\lambda}\right)^t \left\|\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_n\right\|_{\mathrm{F}},$$

  for $\eta \leq c_3$*, where* c_1, c_2, c_3 *are constants;*

- *for CNN, there exists a critical point in* $\mathbb{B}(\boldsymbol{w}^\star, r_{\mathrm{CNN}})$ *such that*

$$\|\widehat{\boldsymbol{w}}_n - \boldsymbol{w}^\star\|_2 \leq c_4 \frac{K}{\rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)} \cdot \sqrt{\frac{d\log n}{n}},$$

  and if the initial point $\boldsymbol{w}_0 \in \mathbb{B}(\boldsymbol{w}^\star, r_{\mathrm{CNN}})$*, GD converges linearly to* $\widehat{\boldsymbol{w}}_n$*, i.e.*

$$\|\boldsymbol{w}_t - \widehat{\boldsymbol{w}}_n\|_2 \leq \left(1 - \frac{c_5\eta\rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)}{K}\right)^t \|\boldsymbol{w}_0 - \widehat{\boldsymbol{w}}_n\|_2,$$

  for $\eta \leq c_6/K$*, where* c_4, c_5, c_6 *are constants.*

Similarly to Theorem 3, for FCN (3.1) Theorem 4 also holds for all column permutations of $\boldsymbol{W}^\star$. The proof can be found in Appendix B.3. Theorem 4 guarantees that the existence of critical points in the local neighborhood of the ground truth, which GD converges to, and also shows that the critical points converge to the ground truth $\boldsymbol{W}^\star$ at the rate of $O(K^{9/4}\sqrt{d\log n/n})$ for FCN (3.1) and $O\left(K\sqrt{d\log n/n}\right)$ for CNN(3.2) with respect to increasing the sample size n. Therefore, $\boldsymbol{W}^\star$ can be recovered consistently as n goes to infinity. Moreover, for both FCN (3.1) and CNN (3.2) gradient descent converges linearly to $\widehat{\boldsymbol{W}}_n$ (or resp. $\hat{\boldsymbol{w}}_n$) at a linear rate, as long as it is initialized in the basin of attraction. To achieve ϵ-accuracy, i.e. $\left\|\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n\right\|_{\mathrm{F}} \leq \epsilon$ (or resp. $\|\boldsymbol{w}_t - \widehat{\boldsymbol{w}}_n\|_2 \leq \epsilon$), it requires a computational complexity of $O\left(ndK^4\log\left(1/\epsilon\right)\right)$ (or resp. $O\left(ndK^2\log\left(1/\epsilon\right)\right)$), which is linear in n, d and $\log(1/\epsilon)$.

3.3 Initialization via Tensor Method

Our initialization adopts the tensor method proposed in [7]. The initialization method works for the FCN model directly, and works for the CNN model with slight modification as presented in [79]. To avoid unnecessary repetitions from the previous work, we focus on the FCN case to outline the algorithm and remark the difference. We recommend the readers refer to [7, 79] for more details.

3.3.1 Preliminary and Algorithm

We start with introducing the necessary definitions which can be found in [7]. We first define a product $\widetilde{\otimes}$ as follows. If $\boldsymbol{v} \in \mathbb{R}^d$ is a vector and $\boldsymbol{I}$ is the identity matrix, then $\boldsymbol{v}\widetilde{\otimes}\boldsymbol{I} = \sum_{j=1}^{d}[\boldsymbol{v}\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_j + \boldsymbol{e}_j\otimes\boldsymbol{v}\otimes\boldsymbol{e}_j + \boldsymbol{e}_j\otimes\boldsymbol{e}_j\otimes\boldsymbol{v}]$. If $\boldsymbol{M}$ is a symmetric

rank-r matrix factorized as $\boldsymbol{M} = \sum_{i=1}^{r} \boldsymbol{s}_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ and $\boldsymbol{I}$ is the identity matrix, then

$$\boldsymbol{M} \widetilde{\otimes} \boldsymbol{I} = \sum_{i=1}^{r} \boldsymbol{s}_i \sum_{j=1}^{d} \sum_{l=1}^{6} \boldsymbol{A}_{l,i,j}, \tag{3.7}$$

where $\boldsymbol{A}_{1,i,j} = \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_j$, $\boldsymbol{A}_{2,i,j} = \boldsymbol{v}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{v}_i \otimes \boldsymbol{e}_j$, $\boldsymbol{A}_{3,i,j} = \boldsymbol{e}_j \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i \otimes \boldsymbol{e}_j$, $\boldsymbol{A}_{4,i,j} = \boldsymbol{v}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{e}_j \otimes \boldsymbol{v}_i$, $\boldsymbol{A}_{5,i,j} = \boldsymbol{e}_j \otimes \boldsymbol{v}_i \otimes \boldsymbol{e}_j \otimes \boldsymbol{v}_i$ and $\boldsymbol{A}_{6,i,j} = \boldsymbol{e}_j \otimes \boldsymbol{e}_j \otimes \boldsymbol{v}_i \otimes \boldsymbol{v}_i$. This allows us to introduce the following quantities.

Definition 3. *Define $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, $\boldsymbol{M}_3$, $\boldsymbol{M}_4$ and $m_{1,i}$, $m_{2,i}$, $m_{3,i}$, $m_{4,i}$ as follows:*

$\boldsymbol{M}_1 = \mathbb{E}[y \cdot \boldsymbol{x}]$,

$\boldsymbol{M}_2 = \mathbb{E}[y \cdot (\boldsymbol{x} \otimes \boldsymbol{x} - \boldsymbol{I})]$,

$\boldsymbol{M}_3 = \mathbb{E}[y \cdot (\boldsymbol{x}^{\otimes 3} - \boldsymbol{x} \widetilde{\otimes} \boldsymbol{I})]$,

$\boldsymbol{M}_4 = \mathbb{E}[y \cdot (\boldsymbol{x}^{\otimes 4} - (\boldsymbol{x} \otimes \boldsymbol{x}) \widetilde{\otimes} \boldsymbol{I} + \boldsymbol{I} \tilde{\otimes} \boldsymbol{I})]$,

$m_{l,i} = g_{l,i}\left(\|\boldsymbol{w}_i^\star\|\right), \forall l = 0, 1, 2, 3, 4,$

where $g_{1,i}(\sigma) = \gamma_1(\sigma)$, $g_{2,i}(\sigma) = \gamma_2(\sigma) - \gamma_0(\sigma)$, $g_{3,i}(\sigma) = \gamma_3(\sigma) - 3\gamma_1(\sigma)$, $g_{4,i}(\sigma) = \gamma_4(\sigma) + 3\gamma_0(\sigma) - 6\gamma_2(\sigma)$, and $\gamma_j(\sigma) = \mathbb{E}_{z \sim \mathcal{N}(0,1)}[\phi(\sigma \cdot z) z^j]$, $\forall j = 0, 1, 2, 3, 4$.

We further define a tensor operation as follows. For a tensor $\boldsymbol{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and three matrices $\boldsymbol{A} \in \mathbb{R}^{n_1 \times d_1}, \boldsymbol{B} \in \mathbb{R}^{n_2 \times d_2}, \boldsymbol{C} \in \mathbb{R}^{n_3 \times d_3}$, the (i, j, k)-th entry of the tensor $\boldsymbol{T}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ is given by

$$\sum_{i'}^{n_1} \sum_{j'}^{n_2} \sum_{k'}^{n_3} \boldsymbol{T}_{i',j',k'} \boldsymbol{A}_{i',i} \boldsymbol{B}_{j',j} \boldsymbol{C}_{k',k}. \tag{3.8}$$

Armed this with definition, we define the following useful quantities.

Definition 4. *Let $\boldsymbol{\alpha} \in \mathbb{R}^d$ denote a randomly picked vector. We define $\boldsymbol{P}_2$ and $\boldsymbol{P}_3$ as follows: $\boldsymbol{P}_2 = \boldsymbol{M}_{j_2}(\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\alpha}, \cdots, \boldsymbol{\alpha})$, where $j_2 = \min\{j \geq 2 | \boldsymbol{M}_j \neq 0\}$, and $\boldsymbol{P}_3 = \boldsymbol{M}_{j_3}(\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\alpha}, \cdots, \boldsymbol{\alpha})$, where $j_3 = \min\{j \geq 3 | \boldsymbol{M}_j \neq 0\}$.*

We further denote $\overline{\boldsymbol{w}} = \boldsymbol{w}/\|\boldsymbol{w}\|$. An important implication of Definition 3 and 4 is that the non-zero matrix P_2 and non-zero tensor P_3 is in the form of $\sum_{i=1}^{K} m_{j_2,i} \left(\alpha^\top \overline{\boldsymbol{w}}_i^\star\right)^{j_2-2} \overline{\boldsymbol{w}}_i^{\star\otimes 2}$, $\sum_{i=1}^{K} m_{j_3,i} \left(\alpha^\top \overline{\boldsymbol{w}}_i^\star\right)^{j_3-3} \overline{\boldsymbol{w}}_i^{\star\otimes 3}$, see [7, Claim 5.5]. The basic strategy is to extract the direction, magnitude information from the empirical version of $\boldsymbol{P}_2$ and $\boldsymbol{P}_3$. Hence estimating $\boldsymbol{W}^\star$ can be decomposed as the following two steps.

- Estimate the direction of each column of $\boldsymbol{W}^\star$ by decomposing $\boldsymbol{P}_2$ to approximate the subspace spanned by $\{\overline{\boldsymbol{w}}_1^\star, \overline{\boldsymbol{w}}_2^\star, \cdots, \overline{\boldsymbol{w}}_K^\star\}$ (denoted by $\boldsymbol{V}$), then reduce the third-order tensor P_3 to a lower-dimension tensor $\boldsymbol{R}_3 = \boldsymbol{P}_3(\boldsymbol{V}, \boldsymbol{V}, \boldsymbol{V}) \in \mathbb{R}^{K\times K\times K}$, and apply non-orthogonal tensor decomposition on $\boldsymbol{R}_3$ to output the estimate $s_i \boldsymbol{V}^\top \overline{\boldsymbol{w}}_i^\star$, where $s_i \in \{1, -1\}$ is a random sign.
- Approximate the magnitude of $\boldsymbol{w}_i^\star$ and the sign s_i by solving a linear system of equations.

The initialization algorithm based on the tensor method is outlined in Algorithm 2. For more implementation details about Algorithm 2, e.g., power method, we refer to [7].

Algorithm 2 Initialization via Tensor Method

Require: Partition n pairs of data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ into three subsets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$.
Ensure:
1: Estimate $\widehat{\boldsymbol{P}}_2$ of $\boldsymbol{P}_2$ from data set $\mathcal{D}_1$.
2: $\boldsymbol{V} \leftarrow \textsc{PowerMethod}(\widehat{\boldsymbol{P}}_2, K)$.
3: Estimate $\widehat{\boldsymbol{R}}_3$ of $\boldsymbol{P}_3(\boldsymbol{V}, \boldsymbol{V}, \boldsymbol{V})$ from data set $\mathcal{D}_2$.
4: $\{\widehat{\boldsymbol{u}}_i\}_{i\in[K]} \leftarrow \mathrm{KCL}(\widehat{\boldsymbol{R}}_3)$.
5: $\{\boldsymbol{w}_i^{(0)}\}_{i\in[K]} \leftarrow \textsc{RecMag}(\boldsymbol{V}, \{\widehat{\boldsymbol{u}}_i\}_{i\in[K]}, \mathcal{D}_3)$.

3.3.2 Performance Guarantee of Initialization

For the classification problem, we make the following technical assumptions, similarly to [7, Assumption 5.3] for the regression problem.

Assumption 1. *The activation function $\phi(z)$ satisfies the following conditions:*

1. *If $M_j \neq 0$, then*

$$\sum_{i=1}^{K} m_{j,i} \left(\boldsymbol{w}_i^{\star\top} \boldsymbol{\alpha} \right)^{j-2} \overline{\boldsymbol{w}_i}^{\star} \overline{\boldsymbol{w}_i}^{\star\top} \neq \mathbf{0},$$
$$\sum_{i=1}^{K} m_{j,i} \left(\overline{\boldsymbol{w}}_i^{\star\top} \boldsymbol{\alpha} \right)^{j-3} (\boldsymbol{V}^\top \overline{\boldsymbol{w}}_i^{\star}) \mathrm{vec}((\boldsymbol{V}^\top \overline{\boldsymbol{w}}_i^{\star})(\boldsymbol{V}^\top \overline{\boldsymbol{w}}_i^{\star})^\top)^\top \neq 0,$$

   for $j \geq 3$.

2. *At least one of M_3 and M_4 is non-zero.*

Assumption 1 is to guarantee that the key terms still contain the magnitude information about $\boldsymbol{w}_j^{\star}$. It can be verified that for sigmoid activation $m_{3,i}$ is non-zero for $\sigma > 0$, hence it will satisfy Assumption 1. Furthermore, we do not require the homogeneous assumption (i.e., $\phi(az) = a^p z$ for an integer p) required in [7], which can be restrictive. Instead, we assume the following condition on the curvature of the activation function around the ground truth, which holds for a larger class of activation functions such as sigmoid and tanh.

Assumption 2. *Let l_1 be the index of the first nonzero M_i where $i = 1, \ldots, 4$. For the activation function $\phi(\cdot)$, there exists a positive constant δ such that $g_{l_1,i}(\cdot)$ is strictly monotone over the interval $(\|\boldsymbol{w}_i^{\star}\| - \delta, \|\boldsymbol{w}_i^{\star}\| + \delta)$, and the derivative of $g_{l_1,i}(\cdot)$ is lower bounded by some constant for all i.*

It can be numerically verified that sigmoid activation will also satisfy Assumption 2. We next present the performance guarantee for the initialization algorithm in the following theorem.

Theorem 5. *For the classification model* (3.1)*, under Assumptions 1 and 2, for any* $0 < \epsilon < 1$ *and* $\zeta > 1$*, if the sample size* $n \geq d \cdot \mathrm{poly}\left(K, \kappa, \zeta, \log d, 1/\epsilon\right)$*, then the output* $\boldsymbol{W}_0 \in \mathbb{R}^{d \times K}$ *of Algorithm 2 satisfies*

$$\|\boldsymbol{W}_0 - \boldsymbol{W}^\star\|_{\mathrm{F}} \leq \epsilon \mathrm{poly}\left(K, \kappa\right) \|\boldsymbol{W}^\star\|_{\mathrm{F}}, \tag{3.9}$$

with probability at least $1 - d^{-\Omega(\zeta)}$.

The proof of Theorem 5 consists of (a) showing the estimation of the direction of $\boldsymbol{W}^\star$ is sufficiently accurate and (b) showing the approximation of the norm of $\boldsymbol{W}^\star$ is accurate enough. The proof of part (a) is the same as that in [7], but our argument in part (b) is different, where we relax the homogeneous assumption on activation functions. More details can be found in the supplementary materials in Appendix B.5.

3.4 Numerical Experiments

For FCN, we first implement gradient descent to verify that the empirical risk function is strongly convex in the local region around $\boldsymbol{W}^\star$. If we initialize multiple times in such a local region, it is expected that gradient descent converges to the same critical point $\widehat{\boldsymbol{W}}_n$, with the same set of training samples. Given a set of training samples, we randomly initialize multiple times, and then calculate the variance of the output of gradient descent. Denote the output of the ℓth run as $\hat{\boldsymbol{w}}_n^{(\ell)} = \mathrm{vec}(\widehat{\boldsymbol{W}}_n^{(\ell)})$ and the mean of the runs as $\bar{\boldsymbol{w}}$. The error is calculated as $\mathrm{SD}_n = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L} \|\hat{\boldsymbol{w}}_n^{(\ell)} - \bar{\boldsymbol{w}}\|^2}$, where $L = 20$ is the total number of random initializations. Adopted from [50], it

quantifies the standard deviation of the estimator $\widehat{\boldsymbol{W}}_n$ under different initializations with the same set of training samples. We say an experiment is successful, if $\mathrm{SD}_n \leq 10^{-4}$. We generate the ground truth $\boldsymbol{W}^\star$ from Gaussian matrices, and the training samples are generated using the FCN (3.1). Figure 3.3 (a) shows the successful rate of gradient descent by averaging over 50 sets of training samples for each pair of n and d, where $K = 3$ and $d = 15, 20, 25$ respectively. The maximum iterations for gradient descent is set as $\mathrm{iter}_{\max} = 3500$. It can be seen that as long as the sample complexity is large enough, gradient descent converges to the same local minima with high probability.

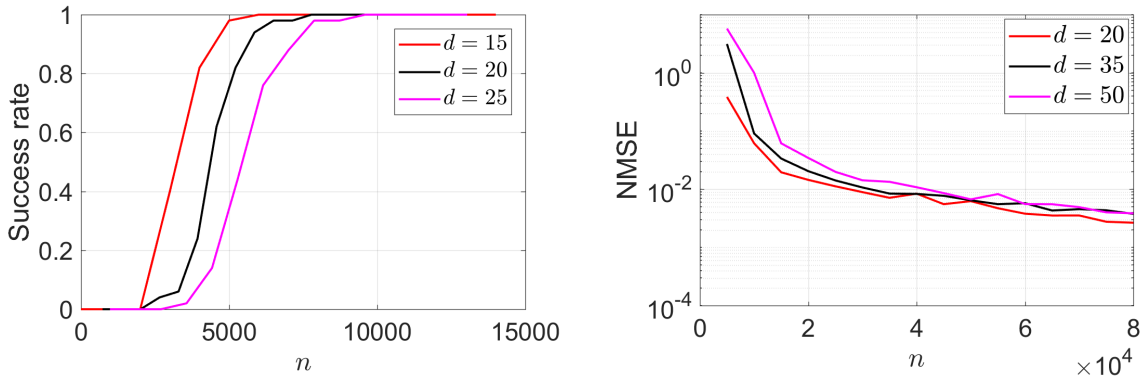
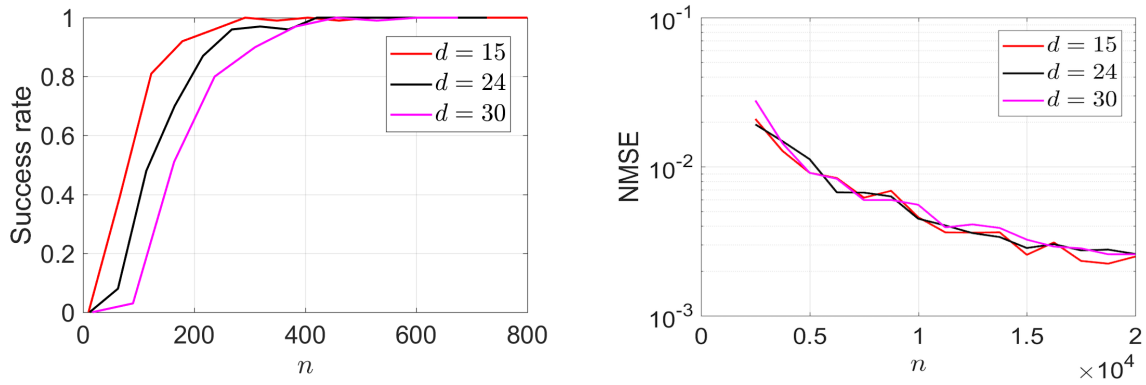


Figure 3.3: For FCN (3.1) fix $K = 3$. (a) Success rate of converging to the same local minima with respect to the sample complexity for various d with threshold 10^{-4}; (b) Average estimation error of gradient descent in a local neighborhood of the ground truth with respect to the sample complexity for various d. The x-axis is scaled to illuminate the correct scaling between n and d.

We next show that the statistical accuracy of the local minimizer for gradient descent if it is initialized close enough to the ground truth. Suppose we initialize around



(a) (b)

Figure 3.4: For CNN (3.2), fix $K = 3$. (a) Success rate of converging to the same local minima with respect to the sample complexity for various d with threshold 10^{-14}; (b) Average estimation error of gradient descent in a local neighborhood of the ground truth with respect to the sample complexity for various d. The x-axis is scaled to illuminate the correct scaling between n and d.

the ground truth such that $\|\boldsymbol{W}_0 - \boldsymbol{W}^\star\|_{\mathrm{F}} \leq 0.1 \cdot \|\boldsymbol{W}^\star\|_{\mathrm{F}}$. We calculate the average estimation error as $\sum_{\ell=1}^{L} \|\widehat{\boldsymbol{W}}_n^{(\ell)} - \boldsymbol{W}^\star\|_{\mathrm{F}}^2 / (L\|\boldsymbol{W}^\star\|_{\mathrm{F}}^2)$ over $L = 100$ Monte Carlo simulations with random initializations. Fig. 3.3 (b) shows the average estimation error with respect to the sample complexity when $K = 3$ and $d = 20, 35, 50$ respectively. It can be seen that the estimation error decreases gracefully as we increase the sample size and matches with the theoretical prediction of error rates reasonably well.

Similarly, for CNN, we first verify that the empirical risk function is locally strongly convex using the same method as before. We generate the entries of true weights $\boldsymbol{w}^\star$ from standard Gaussian distribution, and generate the training samples using the CNN model (3.2). In Fig. 3.4 (a), we say an experiment is successful if $\mathrm{SD}_n \leq 10^{-14}$, and the successful rate is calculated over 100 sets of training samples

with $K = 3$ and $d = 15, 24, 30$ respectively. Then we verify the performance of gradient descent in Fig. 3.4 (b). Suppose we initialized in the neighborhood of $\boldsymbol{w}^\star$, i.e., $\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|_2 \leq 0.9 \cdot \|\boldsymbol{w}^\star\|_2$, for fixed d, K, n, the average error is calculated over $L = 100$ Monte Carlo simulations. It can be seen that the error decreases as we increase the number of samples.

Chapter 4: Guaranteed Recovery of CNN with ReLU Activations

In this chapter, we study the model recovery problem when the data is generated by a one-hidden-layer non-overlap convolutional neural network with ReLU activations, and the goal is to recover the weights of the neural network. We first characterize the landscape of the population risk function $L(\boldsymbol{w})$ by exploiting the second-order property of $L(\boldsymbol{w})$, and then we show that gradient descent with random initialization converges to the global minimum $\boldsymbol{w}^\star$ at a much faster rate than previously established. Furthermore, we also study the empirical risk function which is practically used in training a neural network. We show that with a well designed initialization, gradient descent converges linearly to the true weights $\boldsymbol{w}^\star$ with high probability.

4.1 Problem Formulation

Given an input $\boldsymbol{x} \in \mathbb{R}^d$, the output of a one-hidden-layer non-overlapping convolutional neural network as illustrated in Fig. 4.1 is given by

$$f(\boldsymbol{x}; \boldsymbol{w}) = \frac{1}{K} \sum_{k=1}^{K} \phi\left(\boldsymbol{w}^\top \boldsymbol{x}^{(k)}\right), \tag{4.1}$$

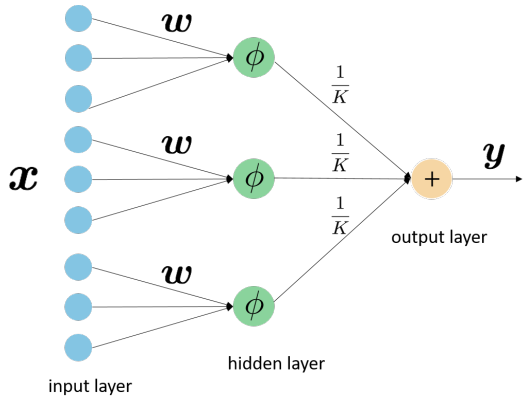


Figure 4.1: Illustration of a one-hidden-layer convolutional neural network without overlap

where K is the number of neurons, $\phi(\cdot)$ is the ReLU activation function, i.e., $\phi(z) = \max(0, z)$, $\boldsymbol{w} \in \mathbb{R}^m$ is the weight of the neural network, and the kth stride of $\boldsymbol{x}$ be given as $\boldsymbol{x}^{(k)} = \left[x_{m(k-1)+1}, \cdots x_{m \cdot k}\right]^\top \in \mathbb{R}^m$. Here, we assume $d = m \cdot K$ for simplicity.

Suppose we are given a set of training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where y_i is generated by the neural network with $\boldsymbol{w}^\star$, i.e., $y_i = f(\boldsymbol{x}_i; \boldsymbol{w}^\star)$. Our goal is to learn the true neural network $\boldsymbol{w}^\star$ via risk minimization. Under the regression setting, the squared loss is generally adopted as the risk function. Two types of risk functions have been considered in the literature, i.e., the population risk function

$$L(\boldsymbol{w}) = \mathbb{E}_{\mathcal{D}}\left[(y - f(\boldsymbol{x}; \boldsymbol{w}))^2\right], \tag{4.2}$$

where $\mathcal{D}$ denotes the joint distribution of $(\boldsymbol{x}, y)$, and the empirical risk function

$$L_n(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i; \boldsymbol{w}))^2. \tag{4.3}$$

In this work, we first consider minimizing the population risk function (4.2) via gradient descent, and then we further analyze the performance of gradient descent on minimizing the empirical risk function (4.3).

As shown in [71], even with infinitely many samples, the problem of minimizing the population risk without any assumption on the distribution $\mathcal{D}$ is NP-Complete. Hence in this work we adopt the standard assumption as in [71, 73] that the input $\{\boldsymbol{x}_i\}_{i=1}^n$ are composed of i.i.d. standard Gaussian vectors.

4.2 Minimizing the Population Risk

To minimize (4.2), we implement the gradient descent (GD) algorithm which is an iterative method and update the parameter as follows.

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t^{(1)} \nabla L\left(\boldsymbol{w}_t\right), \tag{4.4}$$

where $\eta_t^{(1)}$ is the step size or the learning rate, and the subscript t denotes the t-th update. This algorithm and its variant (such as stochastic gradient descent) have been widely used in practical machine learning and deep learning.

The population risk function has been studied in [71], in which only the first-order property was exploited. We take another path by leveraging the second-order property of the population risk function to provide a refined analysis. According to [71], under the Gaussian input assumption, we calculate the closed form of the population risk function (4.2) (up to the additive factors in $\boldsymbol{w}^\star$) as

$$\begin{aligned} L\left(\boldsymbol{w}\right) = \frac{1}{K^2}\Bigg[& \left(\frac{K^2-K}{2\pi} + \frac{K}{2}\right)\|\boldsymbol{w}\|_2^2 - \frac{K^2-K}{\pi}\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2 \\ & - \frac{K}{\pi}\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2\left(\sin\left(\theta\right) + \left(\pi-\theta\right)\cos\left(\theta\right)\right)\Bigg], \end{aligned} \tag{4.5}$$

where $\theta = \arccos\left(\frac{\boldsymbol{w}^\top \boldsymbol{w}^\star}{\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2}\right)$ denotes the angle between the vectors $\boldsymbol{w}$ and $\boldsymbol{w}^\star$.

The population risk function (4.5) in general is nonconvex, and with the closed form expression of the population risk function, the critical points can be characterized. According to [71, Lemma 5.1], the population risk function (4.5) has three critical points, i.e.,

(a) A local maximum at $\boldsymbol{w} = \mathbf{0}$;

(b) A unique global minimum at $\boldsymbol{w} = \boldsymbol{w}^\star$;

(c) A degenerate saddle point at $\boldsymbol{w} = -\frac{K^2-K}{K^2+(\pi-1)K}\boldsymbol{w}^\star$.

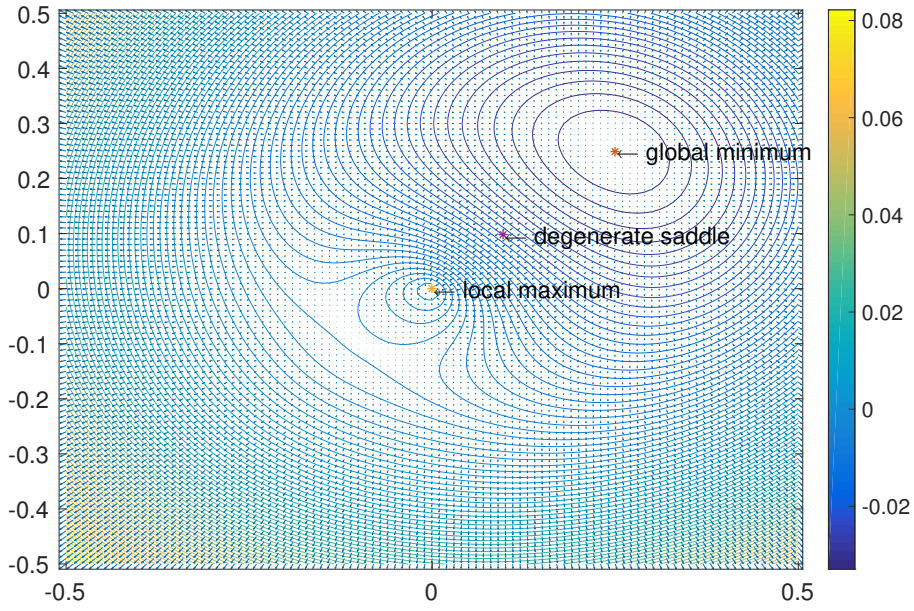


Figure 4.2: Contour plot of the population risk function (4.5) when $m = 2$ and $K = 3$.

We illustrate the landscape of the population risk function (4.5) in Fig 4.2 and mark the three critical points. With the aid of such critical point characterization, [71]

further showed that gradient descent with random initialization converges to a point with ϵ-accuracy to $\boldsymbol{w}^\star$ after $O\left(\frac{1}{\epsilon^4}\right)$ steps for which we suspect is not optimal. Moreover, the analysis of the critical points is not sufficient to fully understand the optimization landscape. In particular, we are interested in understanding whether the population risk is strongly convex in a region around the ground truth $\boldsymbol{w}^\star$; as demonstrated in Chapter 3 that the nonconvex objective functions for learning one-hidden-layer neural network usually have benign geometric properties locally. Hence, beyond the analysis of the critical points which only requires the first-order information of the population risk, we further analyze the Hessian of the population risk. We show that the Hessian of (4.2) can be lower and upper bounded in the local neighborhood of $\boldsymbol{w}^\star$ with a radius $\Omega\left(\frac{1}{K}\right)$. With the aid of this property, we are able to improve the performance analysis of gradient descent and show that with random initialization, gradient descent converges much faster to $\boldsymbol{w}^\star$.

Before presenting the result, we denote $\mathbb{B}\left(\boldsymbol{w}^\star, r\right)$ as a Euclidean ball centered at $\boldsymbol{w}^\star \in \mathbb{R}^m$ with a radius r, i.e.,

$$\mathbb{B}\left(\boldsymbol{w}^\star, r\right) = \left\{\boldsymbol{w} \in \mathbb{R}^m : \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \leq r\right\}. \tag{4.6}$$

The second-order property of the population risk function (4.2) is summarized as follows.

Theorem 6. *Consider the regression model with the ReLU activation function* (4.1)*, and the population risk* (4.2)*. If the input follows a Gaussian distribution, i.e.,* $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right)$*, then the following inequality*

$$\frac{1}{4K} \cdot \boldsymbol{I} \preceq \nabla^2 L\left(\boldsymbol{w}\right) \preceq 3 \cdot \boldsymbol{I} \tag{4.7}$$

holds for all $\boldsymbol{w} \in \mathbb{B}(\boldsymbol{w}^\star, r)$*, as long as* $\boldsymbol{w} \neq -\lambda \cdot \boldsymbol{w}^\star$*, where* λ *is any nonnegative constant and* $r := \frac{1}{3K+3}$*.*

The proof of Theorem 6 is provided in Appendix C.2. Theorem 6 guarantees that the population loss is strongly convex in the neighborhood of $\boldsymbol{w}^\star$, and the size of the neighborhood is roughly $O\left(\frac{1}{K}\right)$ which shrinks when the number of neurons increases. The implication of such a local strong convexity is that once initialized in the local region, if the updates of gradient descent $\boldsymbol{w}_t$ never lies on the line of $-\boldsymbol{w}^\star$, gradient descent converges to the unique critical point $\boldsymbol{w}^\star$ at a linear rate. In fact, it has been shown in [71] that the angle between the updates $\boldsymbol{w}_t$ and $\boldsymbol{w}^\star$ is decreasing, and thus $\boldsymbol{w}_t$ never lies on the line of $-\boldsymbol{w}^\star$.

It has been shown in [71] that the estimation error $\|\boldsymbol{w}_t - \boldsymbol{w}^\star\|_2$ keeps decreasing during the execution of gradient descent. The analysis can be used to guarantee that $\boldsymbol{w}_t$ enters such a local region. Hence, combining these two phases yields the following guarantee for (4.4).

Theorem 7. *Assume* $\|\boldsymbol{w}^\star\|_2 = 1$*, and suppose the initial point* $\boldsymbol{w}_0$ *is uniformly drawn from the unit sphere. Further set the learning rate* $\eta_t^{(1)} = \eta < \frac{1}{3}$*. Then the gradient descent updates* $\boldsymbol{w}_t$ *satisfy*

$$\|\boldsymbol{w}_t - \boldsymbol{w}^\star\|_2 \leq \epsilon \tag{4.8}$$

after $O\left(K^4 + K \cdot \log\left(\frac{1}{\epsilon}\right)\right)$ *iterations.*

Theorem 7 shows that even though the population loss (4.5) is nonconvex, gradient descent still converges fast to the global minimum with random initialization. The theoretical guarantee suggests that this convergence exhibits two phases: first the estimation error contracts relatively slowly; once it enters the local neighborhood

of $\boldsymbol{w}^\star$ after $O\left(K^4\right)$ steps, the convergence rate speeds up. Comparing to the result of [71] which requires $O\left(\frac{1}{\epsilon^4}\right)$ iterations to guarantee the ϵ-accuracy, our result shows that gradient descent converges much faster. This improvement highly depends on the property of the Hessian that we characterize. We sketch the proof of Theorem 7 as follows.

Proof. Let $\boldsymbol{w}_t$ be the estimator at the t-th iteration. Due to the proof of [71, Theorem 5.2], we have that with constant step size $0 < \eta < 1$,

$$\|\boldsymbol{w}_t - \boldsymbol{w}^\star\|_2 \leq O\left(\frac{1}{K}\right) \tag{4.9}$$

holds after $O\left(K^4\right)$ iterations, and $\frac{\boldsymbol{w}^\top \boldsymbol{w}^\star}{\|\boldsymbol{w}\|_2 \|\boldsymbol{w}^\star\|_2} \neq -1$. Hence, after $O\left(K^4\right)$ iterations, w_t enters the local region where

$$\frac{1}{4K} \cdot \boldsymbol{I} \preceq \nabla^2 L\left(\boldsymbol{w}\right) \preceq 3 \cdot \boldsymbol{I} \tag{4.10}$$

holds according to Theorem 6. Then following the same proof as in [59, Lemma 1], we obtain the following error contraction result,

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^\star\|_2 \leq \left(1 - \frac{1}{K}\right) \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|_2, \tag{4.11}$$

as long as $\eta < \frac{1}{3}$. Hence after $O\left(K^4\right)$ iterations, gradient descent converges to $\boldsymbol{w}^\star$ linearly. □

4.3 Minimizing the Empirical Risk

In practice, it is common to minimize the empirical risk function formed with finite samples. Hence, in this section we study the empirical risk function. Particularly, we implement the same gradient descent algorithm to minimize (4.3), which updates

$\boldsymbol{w}_{t+1}$ via

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t^{(2)} \nabla L_n \left(\boldsymbol{w}_t \right), \tag{4.12}$$

where $\eta_t^{(2)}$ is the step size or the learning rate. In this section, we analyze the performance of gradient descent (4.12). Note that throughout the execution of GD, the same set of training samples is used which is the standard implementation of gradient descent. Consequently the analysis is challenging due to the statistical dependence of the iterates with the data. Another challenge of analyzing the performance of GD on minimizing the empirical risk function is due to the non-smoothness of the empirical risk function because ReLU is not differentiable at 0. In our algorithm we define the gradient of ReLU activation $\phi(\cdot)$ as $\phi'(\boldsymbol{x}) = 1_{\{x>0\}}$. Then we calculate the gradient of the empirical risk function (4.3) as

$$\nabla L_n \left(\boldsymbol{w} \right) = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{1}{K^2} \sum_{j=1}^{K} \sum_{l=1}^{K} \left(\phi \left(\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)} \right) - \phi \left(\boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)} \right) \right) \phi' \left(\boldsymbol{w}^\top \boldsymbol{x}_i^{(l)} \right) \boldsymbol{x}_i^{(l)}. \tag{4.13}$$

Since the empirical risk function (4.3) is highly non-convex and non-smoothness, it is not possible to exploit the geometric property of (4.3) by analyzing its Hessian directly. As the analysis developed in Chapter 3, the empirical risk function usually concentrates around the population risk function when the sample size is large enough. However, due to the unboundedness of ReLU activation, it is very difficult to implement this approach and guarantee the concentration.

Instead, we resort to a so-called regularity condition, stated as follows, which is widely used to establish the geometric property of the loss function.

Definition 5 (Regularity condition [59]). *The function $f(\cdot)$ is said to obey the regularity condition RC(μ, λ, ζ) for some μ,λ,$\zeta > 0$ if*

$$2\left\langle \nabla f\left(\boldsymbol{w}\right), \boldsymbol{w}-\boldsymbol{w}^{\star}\right\rangle \geq \mu\|\nabla f\left(\boldsymbol{w}\right)\|_{2}^{2}+\lambda\|\boldsymbol{w}-\boldsymbol{w}^{\star}\|_{2}^{2} \tag{4.14}$$

for all $\boldsymbol{w} \in \mathbb{B}\left(\boldsymbol{w}^{\star}, \zeta\right)$.

Note that this condition does not require $\nabla f(\cdot)$ to be differentiable and it does not require the loss function $f(\cdot)$ to be convex, which is applicable to our case. Moreover, such a condition implies that at any point $\boldsymbol{w} \in \mathbb{B}\left(\boldsymbol{w}^{\star}, \zeta\right)$ the associated gradient $\nabla f(\cdot)$ is positively correlated with the estimation error $\boldsymbol{w}-\boldsymbol{w}^{\star}$, and hence the update rule (4.12) reduces the error $\boldsymbol{w}-\boldsymbol{w}^{\star}$. Next we show that the regularity condition indeed holds for (4.13) in Lemma 1. By leveraging such a geometric property, we show that gradient descent (4.12) converges linearly.

Lemma 1. *For all $\boldsymbol{w} \in \mathbb{B}\left(\boldsymbol{w}^{\star}, r\right)$ with $r:=\frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^{\star}\|_{2}$, the following regularity condition*

$$\left\langle \nabla L_{n}\left(\boldsymbol{w}\right), \boldsymbol{w}-\boldsymbol{w}^{\star}\right\rangle \geq \frac{1}{72K^{2}}\|\nabla L_{n}\left(\boldsymbol{w}\right)\|_{2}^{2}+\frac{1}{4K^{2}}\|\boldsymbol{w}-\boldsymbol{w}^{\star}\|_{2}^{2} \tag{4.15}$$

holds with probability at least $1-\frac{1}{d^{10}}$, as long as the sample complexity satisfies $n \geq c \cdot mK^{2} \cdot \log\left(n\right)$ for some sufficiently large constant c.

The proof of Lemma 1 can be found in Appendix C.3.1, in which we do not optimize the constant. Lemma 1 captures the particular geometric property of the empirical risk function within a local neighborhood of $\boldsymbol{w}^{\star}$. Utilizing such a geometric property, we next show the performance guarantee of gradient descent (4.12) as below.

Theorem 8. *Consider the problem of minimizing* (4.3) *with gradient descent* (4.12). *Assume the input $\boldsymbol{x}_{i} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right)$ and the initial point satisfies $\boldsymbol{w}_{0} \in \mathbb{B}\left(\boldsymbol{w}^{\star}, r\right)$ with*

$r := \frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2$. *Then, with the step size* $\eta_t^{(2)} = \frac{1}{4K^2}$, *the following inequality*

$$\|\boldsymbol{w}_t - \boldsymbol{w}^\star\|_2^2 \leq \left(1 - \frac{1}{288K^4}\right)^t \|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|_2^2 \tag{4.16}$$

holds with probability at least $1 - \frac{1}{d^{10}}$ *as long as the sample complexity satisfies* $n \geq c \cdot mK^2 \cdot \log(n)$ *for some sufficiently large constant* c.

Theorem 8 implies that in the finite sample regime, gradient descent with a well designed initial point still converges to the ground truth although the empirical risk function is not convex. Note that the performance guarantee requires a local initialization with radius $\frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2$, the sample complexity is roughly on the order of $O(dK)$ where $d = mK$, and the sample complexity is linear in terms of both the input dimension d and the number of neurons K.

Proof. Following the same proof of [59, Lemma 2], setting $\eta_t^{(2)} = \frac{1}{4K^2}$, and then applying Lemma 1, we obtain

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^\star\|_2^2 \leq \left(1 - \frac{1}{288K^4}\right) \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|_2^2. \tag{4.17}$$

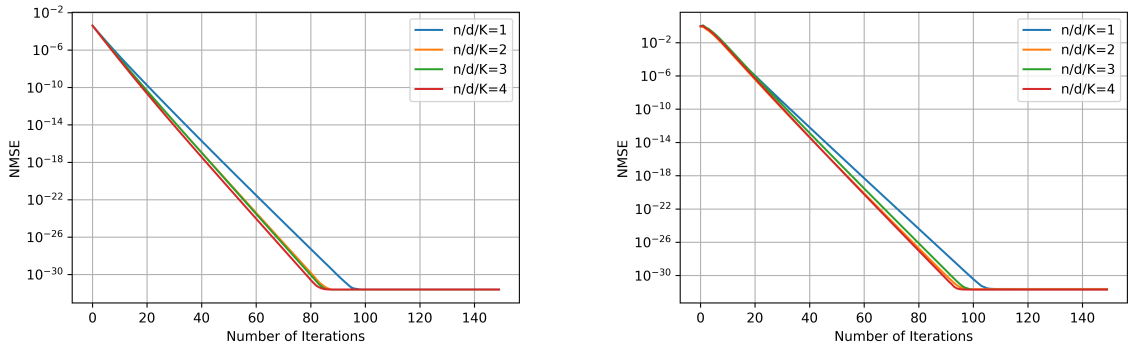
□

4.4 Numerical Experiments

In this section, we implement gradient descent on synthetic data to demonstrate the statistical accuracy of the optimizer if $\boldsymbol{w}$ is initialized close enough to the ground truth $\boldsymbol{w}^\star$, i.e., $\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|_2 \leq \frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2$. We first generate $\left\{\boldsymbol{x}_i \in \mathbb{R}^d\right\}_{i=1}^n$ as Gaussian random vector, i.e., each entry of $\boldsymbol{x}_i$ follows a standard Gaussian distribution, and then we generate the output y_i via (4.1). Given the set of training samples, we randomly initialize $\boldsymbol{w}_0$ and calculate the error of the outputs of gradient descent at

each step. Denoting the output at the t-th step as $\boldsymbol{w}_t$, we define the normalized mean squared error (NMSE) as $\frac{\|\boldsymbol{w}_t-\boldsymbol{w}^\star\|_2^2}{\|\boldsymbol{w}^\star\|_2^2}$. We first initialize $\boldsymbol{w}_0$ as $\boldsymbol{w}_0 = \boldsymbol{w}^\star + \epsilon$ where $\epsilon \sim \mathcal{N}\left(0, \frac{1}{K^3}\|\boldsymbol{w}^\star\|_2^2\right)$, such that $\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|_2 \leq \frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2$, and Fig 4.3 (a) shows that with local initialization, gradient descent converges linearly in log scale, and $\boldsymbol{w}^\star$ can be recovered exactly after the gradient descent converges.

Although the theoretical guarantee of gradient descent needs a local initialization, we observe an interesting phenomenon that gradient descent converges to the ground truth $\boldsymbol{w}^\star$ even without local initialization. We use the same set of training samples and run gradient descent starting from $\boldsymbol{w}_0$ which follows a Gaussian distribution. As can be seen from Fig 4.3 (b), gradient descent still converges to $\boldsymbol{w}^\star$ at a linear rate.



(a) (b)

Figure 4.3: For CNN, fix $K = 10$, $m = 35$ and $d = 350$, the NMSE with respect to the number of steps of gradient descent for various n. (a) local initialization; (b) Gaussian initialization.

Chapter 5: Conclusion and Future Work

5.1 Concluding Remarks

In the study of line spectrum estimation from quantized measurement, we examined the effect of (heavy) quantization in spectral compressed sensing that is useful for understanding wideband spectral signal acquisition and processing. Our contributions are two-fold. We first derived the Cramér-Rao bound for parameter estimation with multiple complex sinusoids using quantized compressed linear measurements. This bound is instrumental in describing the trade-offs between bit depth and sample complexity at different SNR regimes. Such an estimation-theoretical perspective is independent of the algorithm and hasn't been exploited in the previous literature. Secondly, we developed algorithms for spectral-sparse signal recovery using quantized measurements via atomic norm minimization, which do not require knowledge of the quantizer in recovery. Under a mild separation condition, we establish that we can accurately recover a spectrally-sparse signal from the signs of $O(K \log n)$ random linear measurements. The proposed algorithm also can be extended to handle multiple signal snapshots. This generalizes the literature on one-bit compressed sensing to the

important class of spectrally sparse signals using atomic norms, and we carefully examined the performance of the proposed algorithms via numerical experiments. The content of this Chapter is published in [80, 81].

In the study of model recovery of neural-network-based models we recover an one-hidden-layer neural network using the cross-entropy loss in a multi-neuron classification problem. In particular, we have characterized the sample complexity to guarantee local strong convexity in a neighborhood (whose size we have characterized as well) of the ground truth when the training data are generated from a classification model for two types of neural network models: fully-connected network and non-overlapping convolutional network. This guarantees that with high probability, gradient descent converges linearly to the ground truth if initialized properly. The content of this chapter is summarized in a research paper and can be found in [82].

In the study of recovering a one-hidden-layer non-overlap convolutional neural network under regression setting, where the activation is the commonly used ReLU activation in practice. We first analyze the landscape of the population risk function more carefully and find that its Hessian is lower and upper bounded by some positive quantities. Leveraging such a good property, we improve the convergence rate of gradient descent in the existing literature significantly. Secondly, we analyze the non-smooth empirical risk function. We show that a so-called regularity condition holds uniformly in a neighborhood of $\boldsymbol{w}^\star$ with high probability as soon as the sample size is $O(mK^2)$. This further implies that gradient descent finds the global optimal $\boldsymbol{w}^\star$ at a linear convergence rate provided a good initialization.

5.2 Future Work

Along the study of estimating model paratemeters from coarse and nonlinear data, we obtained some insightful results and also found some insteresting open problems. Hence we discuss several potential directions of future work in this section.

Alternative Algorithms for Spectrally-Sparse Signal Estimation

In the work of estimating spectrally sparse signal from its quantized linear measurements as described in Chapter 2, an alternative convex relaxation for spectrally-sparse signal recovery is based on Hankel matrix enhancement and nuclear norm minimization [83, 84]. In the single vector case, instead of imposing the atomic norm regularizer as in (2.22), one may consider

$$\hat{\boldsymbol{x}} = \text{argmin}_{\boldsymbol{x}\in\mathbb{C}^n} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{s}\|_2^2 + \tau_H \|\mathcal{H}(\boldsymbol{x})\|_*. \tag{5.1}$$

Here, $\mathcal{H}(\boldsymbol{x})$ denotes a Hankel matrix given as

$$\mathcal{H}(\boldsymbol{x}) = \begin{bmatrix} x_1 & x_2 & & \\ x_2 & & \iddots & \\ \vdots & \iddots & & \\ x_{n_1} & x_{n_1+1} & \cdots & x_n \end{bmatrix},$$

where n_1 is set as $\lfloor n/2 \rfloor$ to make the matrix $\mathcal{H}(\boldsymbol{x})$ as square as possible, $\|\cdot\|_*$ is the nuclear norm, and τ_H is a regularization parameter. The preliminary numerical simulations suggest this method is also effective for promoting spectral sparsity, hence it's interesting to analyze (5.1) in the future work. Further more, since the Cramér-Rao bounds assume perfect knowledge of the quantizers, they may not be indicative to benchmark the performance of the atomic norm minimization algorithms as proposed in this paper, since these algorithms do not make use of such knowledge. In the future,

it might be interesting to develop estimation-theoretical bounds that only assume partial or little knowledge about the quantizer.

Learning One-Hidden-Layer Neural Networks with ReLU Activations

In the work of learning an one-hidden-layer neural network as discussed in Chapter 3 and 4, we considered different activation functions and network structures. However, ReLU activation is only considered with CNN network, an interesting future direction would be considering learn a FCN network with ReLU activation. Next we briefly introduce the problem. Suppose we are given a set of training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is drawn from an i.i.d. distribution, i.e., $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. And y_i is generated according to

$$y_i = \sum_{j=1}^{K} \phi\left(\boldsymbol{w}_j^{\star\top} \boldsymbol{x}_i\right), \tag{5.2}$$

where $\phi(x) = \max(0, x)$ is the ReLu activation, K is the number of neurons and we are interested in recovering $\boldsymbol{W}^\star = [\boldsymbol{w}_1^\star, \cdots, \boldsymbol{w}_K^\star] \in \mathbb{R}^{d\times K}$ by minimizing the following empirical risk function, i.e.

$$f_n(\boldsymbol{W}) = \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{j=1}^{K} \phi\left(\boldsymbol{w}_j^\top \boldsymbol{x}_i\right) - y_i \right)^2. \tag{5.3}$$

One challenge of solving this problem is that the loss function is non-convex and non-smoothness. It's hard to make use of the second order information about the empirical risk function. Previously, [69] studies the same problem, i.e. for the particular FCN network, they study minimizing the empirical risk function (5.3) via gradient descent. In [69] the authors showed that $\boldsymbol{W}^\star$ can be recovered up to certain estimation error. Such an estimation error is upper bounded by two quantities. The first quantity comes from the optimization aspect, it decreases extremely fast as the algorithm iterates

and it's not avoidable. However, the second term doesn't decrease as the algorithm iterates it diminishes only when the sample size increases to infinity. Hence in the finite sample regime gradient descent is not guaranteed to converge to $\boldsymbol{W}^\star$ exactly. Thus it's desired to further understand the performance of gradient descent, develop new analysis to avoid the second type of error and see if we can get an exact recover of $\boldsymbol{W}^\star$ in finite sample regime.

Another open question about learning the one-hidden-layer CNN with ReLU activations is how gradient descent performs with random initialization. We have already numerically demonstrate in Chapter 4 that gradient descent with random initialization still converges linearly to the ground truth. Moreover, it was shown in [71] that gradient descent with random initialization finds the global optimal of the population risk function. Hence, it's natural to postulate that apply gradient descent to learn CNN doesn't require well designed initialization.

Appendix A: Proofs for Chapter 2

A.1 Proof of Theorem 2

An alternative way to represent the atomic decomposition is to write it as an integration of certain point measure [28]. Define the representing measure of $\boldsymbol{x}^\star$ as

$$\mu(f) = \sum_{k=1}^{K} c_k \delta(f - f_k),$$

where $\delta(\cdot)$ is the delta function. Then we can rewrite $\boldsymbol{x}^\star$ as

$$\boldsymbol{x}^\star = \int_0^1 \boldsymbol{v}(f) d\mu(f) = \sum_{k=1}^{K} c_k \boldsymbol{v}(f_k). \tag{A.1}$$

Correspondingly, denote $\hat{\mu}(f)$ as the representing measure for the solution $\hat{\boldsymbol{x}}$ of (2.22), which means $\hat{\boldsymbol{x}} = \int_0^1 \boldsymbol{v}(f) d\hat{\mu}(f)$.

Denote the reconstruction error as $\boldsymbol{e} = \lambda \boldsymbol{x}^\star - \hat{\boldsymbol{x}}$, and its representing measure is $\gamma = \lambda\mu - \hat{\mu}$. With these definitions, applying [28, Lemma 1], we can bound the error as [28]

$$\|\boldsymbol{e}\|_2^2 \leq \|\boldsymbol{e}\|_{\mathcal{A}}^* \left(\int_F |\gamma| (df) + I_0 + I_1 + I_2 \right), \tag{A.2}$$

where $I_\ell = \sum_{k=1}^{K} I_\ell^k$, for $\ell = 0, 1, 2$, with $I_0^k = \left| \int_{N_k} \gamma (df) \right|$, $I_1^k = n \left| \int_{N_k} (f - f_k) \gamma (df) \right|$, $I_2^k = \frac{n^2}{2} \int_{N_k} (f - f_k)^2 |\gamma| (df)$, where $N_k = \{ f \in \mathbb{T} : d(f, f_k) \leq 0.16/n \}$ as the neighborhoods around each frequency, and $F = \mathbb{T} \setminus \cap_{k=1}^{K} N_k$.

To bound the first term in (A.2), let us denote the deviation

$$\boldsymbol{w} = \boldsymbol{s} - \mathbb{E}[\boldsymbol{s}] = \boldsymbol{s} - \lambda \boldsymbol{x}^\star, \tag{A.3}$$

where $\mathbb{E}[\boldsymbol{w}] = 0$. We have

$$\begin{aligned} \|\boldsymbol{e}\|_{\mathcal{A}}^* &\leq \|\boldsymbol{w}\|_{\mathcal{A}}^* + \|\boldsymbol{s} - \hat{\boldsymbol{x}}\|_{\mathcal{A}}^* \\ &\leq \|\boldsymbol{w}\|_{\mathcal{A}}^* + \tau, \end{aligned} \tag{A.4}$$

where the first line follows from the triangle inequality, and the second line follows from the optimality condition of the AST algorithm in (2.22) in the following lemma.

Lemma 2 (Optimality conditions [29])**.** *$\hat{\boldsymbol{x}}$ is the solution of* (2.22) *if and only if $\|\boldsymbol{s} - \hat{\boldsymbol{x}}\|_{\mathcal{A}}^* \leq \tau$, and $\langle \boldsymbol{s} - \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}} \rangle = \tau \|\hat{\boldsymbol{x}}\|_{\mathcal{A}}$.*

Therefore, if we set $\tau \geq \eta \|\boldsymbol{w}\|_{\mathcal{A}}^*$, where $\eta \geq 1$ is some constant, then plugging this into (A.4) we can show that

$$\|\boldsymbol{e}\|_{\mathcal{A}}^* \leq (\eta^{-1} + 1)\tau \leq 2\tau. \tag{A.5}$$

The second term in (A.2) can be bounded in exactly the same manner as in [28], as long as (A.5) holds. In effect, [28] proved the following bound, under the separation condition, with high probability we have

$$\left(\int_F |\gamma| \, (df) + I_0 + I_1 + I_2 \right) \leq C \frac{K\tau}{n}. \tag{A.6}$$

The following lemma bounds $\|\boldsymbol{w}\|_{\mathcal{A}}^*$, whose proof is provided in Appendix A.2.

Lemma 3. *With probability at least $1 - 1/(\pi n \log n)$, we have*

$$\|\boldsymbol{w}\|_{\mathcal{A}}^* \leq C \cdot \sqrt{\frac{n \log n}{m}},$$

where C is some universal constant.

Therefore, set $\tau = C\eta\sqrt{n\log n/m}$, and plug (A.5) and (A.6) into (A.2), we have

$$\|\boldsymbol{e}\|_2^2 \leq C' \cdot \frac{K\tau^2}{n} \leq C' \frac{K\log n}{m}. \tag{A.7}$$

which is equivalent to

$$\left\| \frac{\hat{\boldsymbol{x}}}{\lambda} - \boldsymbol{x}^\star \right\|_2 \lesssim \frac{1}{\lambda}\sqrt{\frac{K\log n}{m}}.$$

The proof is complete.

A.2 Proof of Lemma 3

By definition, we can write $\|\boldsymbol{w}\|_{\mathcal{A}}^*$ as

$$\begin{aligned} \|\boldsymbol{w}\|_{\mathcal{A}}^* &= \sup_{f\in[0,1)} |\langle \boldsymbol{s} - \lambda\boldsymbol{x}^\star, \boldsymbol{v}(f)\rangle| \\ &= \sup_{f\in[0,1)} |\langle \boldsymbol{s}, \boldsymbol{v}(f)\rangle - \mathbb{E}[\langle \boldsymbol{s}, \boldsymbol{v}(f)\rangle]| \\ &= \sup_{f\in[0,1)} |g_{\boldsymbol{x}^\star}(f) - \mathbb{E}[g_{\boldsymbol{x}^\star}(f)]| \end{aligned} \tag{A.8}$$

where $g_{\boldsymbol{x}^\star}(f) := \langle \boldsymbol{s}, \boldsymbol{v}(f)\rangle = \frac{1}{m}\sum_{i=1}^m y_i \langle \boldsymbol{a}_i, \boldsymbol{v}(f)\rangle$.

To proceed, we use the following symmetrization bound, which is the complex-valued version of [35, Lemma 5.1].

Lemma 4. *Let $\{\epsilon_i\}_{i=1}^m$ be a sequence of independent complex-valued random variables, where $\epsilon_i \sim \epsilon = e^{j2\pi\theta}$, where θ uniformly distributed between $[0,1)$. Then*

$$\begin{aligned} \mu :&= \mathbb{E}\left[\sup_{f\in[0,1)} |g_{\boldsymbol{x}^\star}(f) - \mathbb{E}[g_{\boldsymbol{x}^\star}(f)]|\right] \\ &\leq 2\mathbb{E}\left[\sup_{f\in[0,1)} \frac{1}{m}\left|\sum_{i=1}^m \epsilon_i y_i \langle \boldsymbol{a}_i, \boldsymbol{v}(f)\rangle\right|\right]. \end{aligned} \tag{A.9}$$

Furthermore, we have the deviation inequality

$$\begin{aligned} &\mathbb{P}\left\{\sup_{f\in[0,1)} |g_{\boldsymbol{x}^\star}(f) - \mathbb{E}[g_{\boldsymbol{x}^\star}(f)]| \geq 2\mu + t\right\} \\ &\qquad \leq 4\mathbb{P}\left\{\sup_{f\in[0,1)} \frac{1}{m}\left|\sum_{i=1}^m \epsilon_i y_i \langle \boldsymbol{a}_i, \boldsymbol{v}(f)\rangle\right| > \frac{t}{2}\right\}. \end{aligned} \tag{A.10}$$

Before applying Lemma 4, note that by symmetrization and rotational invariance, $\epsilon_i y_i \boldsymbol{a}_i$ have the same i.i.d. distribution of $\sqrt{2}\boldsymbol{a}_i$. Therefore, the following quantities are equivalent in distribution:

$$\begin{aligned}\sup_{f\in[0,1)} \frac{1}{m}\left|\sum_{i=1}^{m} \epsilon_i y_i \left\langle \boldsymbol{a}_i, \boldsymbol{v}\left(f\right)\right\rangle\right| &\sim \frac{\sqrt{2}}{m} \sup_{f\in[0,1)} \left|\sum_{i=1}^{m} \left\langle \boldsymbol{a}_i, \boldsymbol{v}\left(f\right)\right\rangle\right| \\ &\sim \sqrt{\frac{2}{m}} \sup_{f\in[0,1)} |\left\langle \boldsymbol{g}, \boldsymbol{v}\left(f\right)\right\rangle|,\end{aligned}$$

where $\boldsymbol{g}$ is a vector composed of i.i.d. $\mathcal{CN}(0,1)$.

Applying (A.10) in Lemma 4 to (A.8), we have

$$\mathbb{P}\left(\|\boldsymbol{w}\|_{\mathcal{A}}^{*} \geq 2\mu + t\right) \leq 4\mathbb{P}\left(\sqrt{\frac{2}{m}} \sup_{f\in[0,1)} |\left\langle \boldsymbol{g}, \boldsymbol{v}\left(f\right)\right\rangle| \geq \frac{t}{2}\right). \tag{A.11}$$

From (A.9) in Lemma 4, we have

$$\begin{aligned}\mu = \mathbb{E}\left[\|\boldsymbol{w}\|_{\mathcal{A}}^{*}\right] &\leq 2\sqrt{\frac{2}{m}}\mathbb{E}\left[\sup_{f\in[0,1)} |\left\langle \boldsymbol{g}, \boldsymbol{v}\left(f\right)\right\rangle|\right] \\ &\leq C\sqrt{\frac{n\log n}{m}},\end{aligned} \tag{A.12}$$

where the second line follows from [29, Appendix C,D] as

$$\mathbb{E}\left[\sup_{f\in[0,1]} |\left\langle \boldsymbol{g}, \boldsymbol{v}\left(f\right)\right\rangle|\right] \leq C_1\sqrt{n\log\left(n\right)}.$$

Moreover, from [29, Appendix C], we have

$$\sup_{f\in[0,1)} |\langle \boldsymbol{g}, \boldsymbol{v}\left(f\right)\rangle| \leq C_2 \cdot \sqrt{n\log n}$$

hold with probability at least $1 - 1/(\pi n \log n)$. Set $t = 2C_2\sqrt{n\log n}$ and plug in the above two inequalities in (A.11), we have that

$$\|\boldsymbol{w}\|_{\mathcal{A}}^{*} \leq C \cdot \sqrt{\frac{n\log n}{m}}$$

holds with probability at least $1 - 1/(\pi n \log n)$.

Appendix B: Proofs for Chapter 3

B.1 Gradient and Hessian of the Population Loss

For the convenience of analysis, we first provide the gradient and the Hessian formula for the cross-entropy loss using FCN and CNN here.

B.1.1 The FCN case

Consider the population loss function $f(\boldsymbol{W}) = \mathbb{E}\left[f_n(\boldsymbol{W})\right] = \mathbb{E}\left[\ell\left(\boldsymbol{W};\boldsymbol{x}\right)\right]$, where $\ell\left(\boldsymbol{W};\boldsymbol{x}\right)$ is associated with network $H_{\mathrm{FCN}}\left(\boldsymbol{W},\boldsymbol{x}\right) = \frac{1}{K}\sum_{k=1}^{K}\phi(\boldsymbol{w}_k^\top\boldsymbol{x})$. Hiding the dependence on $\boldsymbol{x}$ for notational simplicity, we can calculate the gradient and the Hessian as

$$\mathbb{E}\left[\frac{\partial\ell\left(\boldsymbol{W}\right)}{\partial\boldsymbol{w}_j}\right] = \mathbb{E}\left[-\frac{1}{K}\frac{(y-H(\boldsymbol{W}))}{H(\boldsymbol{W})\left(1-H(\boldsymbol{W})\right)}\phi'\left(\boldsymbol{w}_j^\top\boldsymbol{x}\right)\boldsymbol{x}\right], \tag{B.1}$$

$$\mathbb{E}\left[\frac{\nabla^2\ell\left(\boldsymbol{W}\right)}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_l}\right] = \mathbb{E}\left[\xi_{j,l}\left(\boldsymbol{W}\right)\cdot\boldsymbol{x}\boldsymbol{x}^\top\right], \tag{B.2}$$

for $1 \le j, l \le K$. Here, when $j \neq l$,

$$\xi_{j,l}\left(\boldsymbol{W}\right) = \frac{1}{K^2}\phi'\left(\boldsymbol{w}_j^\top\boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_l^\top\boldsymbol{x}\right)\cdot\frac{H\left(\boldsymbol{W}\right)^2 + y - 2y\cdot H\left(\boldsymbol{W}\right)}{H^2(\boldsymbol{W})\left(1-H(\boldsymbol{W})\right)^2},$$

and when $j = l$,

$$\xi_{j,j}(\boldsymbol{W}) = \frac{1}{K^2}\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)^2 \cdot \frac{H(\boldsymbol{W})^2 + y - 2y \cdot H(\boldsymbol{W})}{H^2(\boldsymbol{W})\left(1 - H(\boldsymbol{W})\right)^2} - \frac{1}{K}\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right) \cdot \frac{y - H(\boldsymbol{W})}{H(\boldsymbol{W})\left(1 - H(\boldsymbol{W})\right)}.$$

B.1.2 The CNN case

For the CNN case, i.e., $H(\boldsymbol{w}) := H_{\text{CNN}}(\boldsymbol{w}, \boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K}\phi(\boldsymbol{w}^\top \boldsymbol{x}^{(k)})$, the corresponding gradient and Hessian of the population loss function $\ell(\boldsymbol{w})$ is given by

$$\mathbb{E}\left[\frac{\partial \ell(\boldsymbol{w})}{\partial \boldsymbol{w}}\right] = \mathbb{E}\left[-\phi'(\boldsymbol{w}^\top \boldsymbol{x}^{(1)}) \cdot \frac{y - H(\boldsymbol{w})}{H(\boldsymbol{w})\left(1 - H(\boldsymbol{w})\right)} \cdot \boldsymbol{x}^{(1)}\right], \tag{B.3}$$

$$\mathbb{E}\left[\frac{\nabla^2 \ell(\boldsymbol{w})}{\partial \boldsymbol{w}^2}\right] = \mathbb{E}\left[\sum_{j=1}^{K}\sum_{l=1}^{K} g_{j,l}(\boldsymbol{w})\, \boldsymbol{x}^{(j)} \boldsymbol{x}^{(l)\top}\right], \tag{B.4}$$

where when $j \neq l$,

$$g_{j,l}(\boldsymbol{w}) = \frac{1}{K^2} \cdot \frac{H(\boldsymbol{w})^2 + y - 2y \cdot H(\boldsymbol{w})}{\left(H(\boldsymbol{w})\left(1 - H(\boldsymbol{w})\right)\right)^2}\phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)}\right)\phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(l)}\right),$$

and when $j = l$,

$$g_{j,j}(\boldsymbol{w}) = \frac{1}{K^2} \cdot \frac{H(\boldsymbol{w})^2 + y - 2y \cdot H(\boldsymbol{w})}{\left(H(\boldsymbol{w})\left(1 - H(\boldsymbol{w})\right)\right)^2} \cdot \phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)}\right)^2 - \frac{1}{K} \cdot \frac{y - H(\boldsymbol{w})}{H(\boldsymbol{w})\left(1 - H(\boldsymbol{w})\right)} \cdot \phi''\left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)}\right).$$

B.2 Proof of Theorem 3

In order to show that the empirical loss possesses a local strong convexity, we follow the following steps:

1. We first show that the Hessian $\nabla^2 f(\boldsymbol{W})$ of the population loss function is smooth with respect to $\nabla^2 f(\boldsymbol{W}^\star)$ (Lemma 5);

2. We then show that $\nabla^2 f(\boldsymbol{W})$ satisfies local strong convexity and smoothness in a neighborhood of $\boldsymbol{W}^\star$ with appropriately chosen radius, $\mathbb{B}(\boldsymbol{W}^\star, r)$, by leveraging similar properties of $\nabla^2 f(\boldsymbol{W}^\star)$ (Lemma 6);

3. Next, we show that the Hessian of the empirical loss function $\nabla^2 f_n(\boldsymbol{W})$ is close to its population counterpart $\nabla^2 f(\boldsymbol{W})$ uniformly in $\mathbb{B}(\boldsymbol{W}^\star, r)$ with high probability (Lemma 7).

4. Finally, putting all the arguments together, we establish $\nabla^2 f_n(\boldsymbol{W})$ satisfies local strong convexity and smoothness in $\mathbb{B}(\boldsymbol{W}^\star, r)$.

To begin, we first show that the Hessian of the population risk is smooth enough around $\boldsymbol{W}^\star$ in the following lemmas.

Lemma 5 (Hessian Smoothness of Population Loss)**.** *Suppose the loss $\ell(\cdot)$ associates with FCN* (3.1)*, and assume $\|\boldsymbol{w}_k^\star\|_2 \le 1$ for all k and $\|\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}} \le 0.7$. Then we have*

$$\|\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)\| \le \frac{C_1}{K^{\frac{3}{2}}} \cdot \|\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}}, \tag{B.5}$$

holds. Similarly, suppose the loss $\ell(\cdot)$ associates with CNN (3.2)*, and assume $\|\boldsymbol{w}^\star\|_2 \le 1$ and $\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \le 0.7$. We have*

$$\|\nabla^2 f(\boldsymbol{w}) - \nabla^2 f(\boldsymbol{w}^\star)\| \le C_2 \cdot K \cdot \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2, \tag{B.6}$$

holds. Here C_1 and C_2 denote some large constants.

The proof is provided in Appendix B.4.1. Together with the fact that $\nabla^2 f(\boldsymbol{W}^\star)$ be lower and upper bounded, Lemma 5 allows us to bound $\nabla^2 f(\boldsymbol{W})$ in a neighborhood around ground truth, given below.

Lemma 6 (Local Strong Convexity and Smoothness of Population Loss)**.** *If the loss* $\ell(\cdot)$ *associates with FCN* (3.1)*, there exists some constant* C_1*, such that*

$$\frac{4}{K^2} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2 \lambda} \cdot \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{W}) \preceq C_1 \cdot \boldsymbol{I},$$

holds for all $\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})$ *with* $r_{\mathrm{FCN}} := \frac{C_2}{K^{\frac{1}{2}}} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2 \lambda}$*. Moreover, if loss* $\ell(\cdot)$ *associates with CNN* (3.2)*, then we have*

$$C_3 \cdot \frac{\rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)}{K} \cdot \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{w}) \preceq C_4 \cdot K \cdot \boldsymbol{I}, \tag{B.7}$$

holds for all $\boldsymbol{w} \in \mathbb{B}(\boldsymbol{w}^\star, r_{\mathrm{CNN}})$ *with* $r_{\mathrm{CNN}} := C_5 \cdot \frac{\rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)}{K^2}$*.*

The proof is provided in Appendix B.4.2. The next step is to show the Hessian of the empirical loss function is close to the Hessian of the population loss function in a uniform sense, which can be summarized as follows.

Lemma 7. *For the loss* $\ell(\cdot)$ *associated with FCN* (3.1)*, there exists a constant* C *such that as long as* $n \geq C \cdot dK \log dK$*, with probability at least* $1 - d^{-10}$*, the following holds*

$$\sup_{\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})} \|\nabla^2 f_n(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W})\| \leq C \sqrt{\frac{dK \log n}{n}}, \tag{B.8}$$

where $r_{\mathrm{FCN}} := \frac{C}{K^{\frac{1}{2}}} \cdot \frac{\rho(\sigma_K)}{\kappa^2 \lambda}$*. For the loss* $\ell(\cdot)$ *associated with CNN* (3.2)*, we have*

$$\sup_{\boldsymbol{w} \in \mathbb{B}(\boldsymbol{w}^\star, r_{\mathrm{CNN}})} \|\nabla^2 f_n(\boldsymbol{w}) - \nabla^2 f(\boldsymbol{w})\| \leq C K^2 \sqrt{\frac{\frac{d}{K} \cdot \log(n)}{n}}, \tag{B.9}$$

holds with probability at least $1 - d^{-10}$*, as long as* $n \geq \frac{d}{K} \log\left(\frac{d}{K}\right)$*, and* $r_{\mathrm{CNN}} := C \cdot \frac{\rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)}{K^2}$*.*

The proof is provided in Appendix B.4.3. Combining the above results will give us the result. Next we assume that the loss $\ell(\cdot)$ associates with FCN, and take it as

an example in the proof. Then if the loss $\ell(\cdot)$ associates with CNN, the proof follows in the same manner.

Proof of Theorem 3. With probability at least $1 - d^{-10}$,

$$\begin{aligned}\nabla^2 f_n(\boldsymbol{W}) &\succeq \nabla^2 f(\boldsymbol{W}) - \left\|\nabla^2 f_n(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W})\right\| \cdot \boldsymbol{I} \\ &\succeq \Omega\left(\frac{1}{K^2} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2 \lambda}\right) \cdot \boldsymbol{I} - \Omega\left(C \cdot \sqrt{\frac{dK \log n}{n}}\right) \cdot \boldsymbol{I}.\end{aligned}$$

As long as the sample size n is set to satisfy

$$C \cdot \sqrt{\frac{dK \log n}{n}} \leq \frac{1}{K^2} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2 \lambda},$$

i.e. $n \geq C \cdot dK^5 \log^2 d \cdot \left(\frac{\kappa^2 \lambda}{\rho_{\mathrm{FCN}}(\sigma_K)}\right)^2$, we have

$$\nabla^2 f_n(\boldsymbol{W}) \succeq \Omega\left(\frac{1}{K^2} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2 \lambda}\right) \cdot \boldsymbol{I}.$$

holds for all $\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})$. Similarly, we have

$$\nabla^2 f_n(\boldsymbol{W}) \preceq C \cdot \boldsymbol{I}$$

holds for all $\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})$. □

B.3 Proof of Theorem 4

We have established that $f_n(\boldsymbol{W})$ is strongly convex in $\mathbb{B}(\boldsymbol{W}^\star, r)$ in Theorem 3. Thus there exists at most one critical point in $\mathbb{B}(\boldsymbol{W}^\star, r)$. The proof of Theorem 4 follows the steps below:

1. We first show that the gradient $\nabla f_n(\boldsymbol{W})$ concentrates around $\nabla f(\boldsymbol{W})$ in $\mathbb{B}(\boldsymbol{W}^\star, r)$ (Lemma 8), and then invoke [50, Theorem 2] to guarantee that there indeed exists a critical point $\widehat{\boldsymbol{W}}_n$ in $\mathbb{B}(\boldsymbol{W}^\star, r)$;

2. We next show that $\widehat{\boldsymbol{W}}_n$ is close to $\boldsymbol{W}^\star$ and gradient descent converges linearly to $\hat{\boldsymbol{W}}_n$ with a properly chosen step size.

To begin, the following lemma establishes that $\nabla f_n(\boldsymbol{W})$ uniformly concentrates around $\nabla f(\boldsymbol{W})$.

Lemma 8. *If the loss $\ell(\cdot)$ associates with FCN (3.1) with $r_{\mathrm{FCN}} := \frac{C}{K^{\frac{1}{2}}} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2 \lambda}$, and $\|\boldsymbol{w}_k^\star\|_2 \leq 1$ for all k, then*

$$\sup_{\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})} \|\nabla f_n(\boldsymbol{W}) - \nabla f(\boldsymbol{W})\| \leq C\sqrt{\frac{d\sqrt{K}\log n}{n}}$$

holds with probability at least $1 - d^{-10}$, as long as $n \geq CdK\log(dK)$. If the loss $\ell(\cdot)$ associates with CNN (3.2), with $r_{\mathrm{CNN}} := C \cdot \frac{\rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)}{K^2}$ and $\|\boldsymbol{w}^\star\|_2 \leq 1$, then

$$\sup_{\boldsymbol{w} \in \mathbb{B}(\boldsymbol{w}^\star, r_{\mathrm{CNN}})} \|\nabla f_n(\boldsymbol{w}) - \nabla f(\boldsymbol{w})\| \leq C \cdot \sqrt{\frac{d\log n}{n}} \tag{B.10}$$

holds with probability at least $1 - d^{-10}$ as long as $n \geq C\frac{d}{K}\log\left(\frac{d}{K}\right)$.

The proof is provided in Appendix B.4.4. Notice that for the population risk function $f(\boldsymbol{W})$, $\boldsymbol{W}^\star$ is the unique critical point in $\mathbb{B}(\boldsymbol{W}^\star, r)$ due to local strong convexity. With Lemma 7 and Lemma 8, we can invoke [50, Theorem 2], which guarantees the following.

Corollary 1. *If the loss $\ell(\cdot)$ associates with FCN or CNN, there exists one and only one critical point $\widehat{\boldsymbol{W}}_n \in \mathbb{B}(\boldsymbol{W}^*, r)$ that satisfies $\nabla f_n\left(\widehat{\boldsymbol{W}}_n\right) = \boldsymbol{0}$ correspondingly.*

Again, since the proof for the case with the loss $\ell(\cdot)$ associating with FCN is the same as that for CNN, we next take FCN as an example.

We first show that $\widehat{\boldsymbol{W}}_n$ is close to $\boldsymbol{W}^\star$. By the extended mean value theorem, there exists $\boldsymbol{W}'$ on the straight line connecting $\boldsymbol{W}^\star$ and $\widehat{\boldsymbol{W}}_n$ such that

$$\begin{aligned} f_n\left(\widehat{\boldsymbol{W}}_n\right) &= f_n\left(\boldsymbol{W}^\star\right) + \left\langle \nabla f_n\left(\boldsymbol{W}^\star\right), \mathrm{vec}\left(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\right)\right\rangle \\ &\quad + \frac{1}{2}\mathrm{vec}\left(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\right)^\top \nabla^2 f_n\left(\boldsymbol{W}'\right) \mathrm{vec}\left(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\right) \\ &\leq f_n\left(\boldsymbol{W}^\star\right), \end{aligned} \tag{B.11}$$

where the last inequality follows from the optimality of $\widehat{\boldsymbol{W}}_n$. By Theorem 3, we have

$$\begin{aligned} &\frac{1}{2}\mathrm{vec}\left(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\right)^\top \nabla^2 f_n\left(\boldsymbol{W}'\right) \mathrm{vec}\left(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\right) \\ &\geq \Omega\left(\frac{1}{K^2}\cdot\frac{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}{\kappa^2\lambda}\right)\left\|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\right\|_{\mathrm{F}}^2. \end{aligned} \tag{B.12}$$

On the other hand, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \left|\left\langle \nabla f_n\left(\boldsymbol{W}^\star\right), \mathrm{vec}\left(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\right)\right\rangle\right| &\leq \|\nabla f_n\left(\boldsymbol{W}^\star\right)\|_2 \|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\|_{\mathrm{F}} \\ &\leq \Omega\left(\sqrt{\frac{dK^{1/2}\log n}{n}}\right)\|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\|_{\mathrm{F}}, \end{aligned} \tag{B.13}$$

where the last line follows from Lemma 8. Plugging (B.12) and (B.13) into (B.11), we have

$$\|\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^\star\|_{\mathrm{F}} \leq \Omega\left(\frac{K^{\frac{9}{4}}\kappa^2\lambda}{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}\sqrt{\frac{d\log n}{n}}\right). \tag{B.14}$$

Now we have established that there indeed exists a critical point in $\mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})$. We can then establish the local linear convergence of gradient descent as below. Let $\boldsymbol{W}_t$ be the estimate at the t-th iteration. Due to the update rule, we have

$$\begin{aligned} \boldsymbol{W}_{t+1} - \widehat{\boldsymbol{W}}_n &= \boldsymbol{W}_t - \eta\nabla f_n\left(\boldsymbol{W}_t\right) - \left(\widehat{\boldsymbol{W}}_n - \eta\nabla f_n\left(\widehat{\boldsymbol{W}}_n\right)\right) \\ &= \left(\boldsymbol{I} - \eta\int_0^1 \nabla^2 f_n\left(\boldsymbol{W}\left(\gamma\right)\right)\right)\left(\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n\right), \end{aligned}$$

where $\boldsymbol{W}(\gamma) = \widehat{\boldsymbol{W}}_n + \gamma\left(\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n\right)$ for $\gamma \in [0,1]$. If $\boldsymbol{W}_t \in \mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})$, it is obvious that $\boldsymbol{W}(\gamma) \in \mathbb{B}(\boldsymbol{W}^\star, r_{\mathrm{FCN}})$, and by Theorem 3, we have

$$H_{\min} \cdot \boldsymbol{I} \preceq \nabla^2 f_n\left(\boldsymbol{W}(\gamma)\right) \preceq H_{\max} \cdot \boldsymbol{I},$$

where $H_{\min} = \Omega\left(\frac{1}{K^2} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2 \lambda}\right)$ and $H_{\max} = C$. Therefore, we have

$$\begin{aligned}\|\boldsymbol{W}_{t+1} - \widehat{\boldsymbol{W}}_n\|_{\mathrm{F}} &\leq \|\boldsymbol{I} - \eta \int_0^1 \nabla^2 f_n\left(\boldsymbol{W}\left(\gamma\right)\right) \| \|\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n\|_{\mathrm{F}} \\ &\leq \left(1 - \eta H_{\min}\right) \|\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n\|_{\mathrm{F}}. \end{aligned} \tag{B.15}$$

Hence, by setting $\eta = \frac{1}{H_{\max}} := \Omega\left(C\right)$, we obtain

$$\|\boldsymbol{W}_{t+1} - \widehat{\boldsymbol{W}}_n\|_{\mathrm{F}} \leq \left(1 - \frac{H_{\min}}{H_{\max}}\right) \|\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n\|_{\mathrm{F}}, \tag{B.16}$$

which implies that gradient descent converges linearly to the local minimizer $\widehat{\boldsymbol{W}}_n$.

B.4 Proof of Auxiliary Lemmas

B.4.1 Proof of Lemma 5.

We prove the two claims for FCN and CNN separately as below.

- **The FCN case:** Let $\boldsymbol{\Delta} = \nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)$. For each $(j,l) \in [K] \times [K]$, let $\boldsymbol{\Delta}_{j,l} \in \mathbb{R}^{d \times d}$ denote the (j,l)-th block of $\boldsymbol{\Delta}$. Let $\boldsymbol{a} = [\boldsymbol{a}_1^\top, \cdots, \boldsymbol{a}_K^\top]^\top \in \mathbb{R}^{dK}$. By definition,

$$\begin{aligned}\|\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)\| &= \max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^\top (\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)) \boldsymbol{a} \\ &= \max_{\|\boldsymbol{a}\|=1} \sum_{j=1}^K \sum_{l=1}^K \boldsymbol{a}_j^\top \boldsymbol{\Delta}_{j,l} \boldsymbol{a}_l. \end{aligned} \tag{B.17}$$

From (B.2) we know that

$$\boldsymbol{\Delta}_{j,l} = \frac{\partial^2 f\left(\boldsymbol{W}\right)}{\partial \boldsymbol{w}_j \partial \boldsymbol{w}_l} - \frac{\partial^2 f\left(\boldsymbol{W}^\star\right)}{\partial \boldsymbol{w}_j^\star \partial \boldsymbol{w}_l^\star} = \mathbb{E}\left[\left(\xi_{j,l}\left(\boldsymbol{W}\right) - \xi_{j,l}\left(\boldsymbol{W}^\star\right)\right) \cdot \boldsymbol{x}\boldsymbol{x}^\top\right], \tag{B.18}$$

and then by the mean value theorem, we can further expand $\xi_{j,l}(\boldsymbol{W})$ as

$$\xi_{j,l}(\boldsymbol{W}) = \xi_{j,l}(\boldsymbol{W}^\star) + \sum_{k=1}^{K} \left\langle \frac{\partial \xi_{j,l}\left(\tilde{\boldsymbol{W}}\right)}{\partial \tilde{\boldsymbol{w}}_k}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle, \tag{B.19}$$

where $\tilde{\boldsymbol{W}} = \eta \cdot \boldsymbol{W} + (1-\eta)\boldsymbol{W}^\star$ for some $\eta \in (0,1)$. Thus we can write $\boldsymbol{\Delta}_{j,l}$ as

$$\boldsymbol{\Delta}_{j,l} = \mathbb{E}\left[\left(\sum_{k=1}^{K} \left\langle \frac{\partial \xi_{j,l}\left(\tilde{\boldsymbol{W}}\right)}{\partial \tilde{\boldsymbol{w}}_k}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle\right) \cdot \boldsymbol{x}\boldsymbol{x}^\top\right], \tag{B.20}$$

which can be further simplified as

$$\boldsymbol{\Delta}_{j,l} = \mathbb{E}\left[\left(\sum_{k=1}^{K} T_{j,l,k} \left\langle \boldsymbol{x}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle\right) \cdot \boldsymbol{x}\boldsymbol{x}^\top\right], \tag{B.21}$$

by the fact that $\frac{\partial \xi_{j,l}(\tilde{\boldsymbol{W}})}{\partial \tilde{\boldsymbol{w}}_k}$ can be written as $T_{j,l,k} \cdot \boldsymbol{x}$, where $T_{j,l,k} \in \mathbb{R}$ is a scalar depending on $\boldsymbol{x}$. When $j = l$, we calculate $\frac{\partial \xi_{j,l}(\tilde{\boldsymbol{W}})}{\partial \tilde{\boldsymbol{w}}_k}$ for illustration,

$$\begin{aligned}
\frac{\partial \xi_{j,j}(\boldsymbol{W})}{\partial \boldsymbol{w}_k} &= \left(-\frac{2}{K^2}\frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)^2}{H(\boldsymbol{W})^3} + \frac{1}{K}\frac{\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})^2}\right)\frac{1}{K}\phi\left(\boldsymbol{w}_k^\top \boldsymbol{x}\right)\boldsymbol{x}, \quad k \neq j \\
\frac{\partial \xi_{j,j}(\boldsymbol{W})}{\partial \boldsymbol{w}_k} &= \left(\frac{2}{K^2}\left(\frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})^2} - \frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)^2}{H(\boldsymbol{W})^3}\right)\right. \\
&\quad \left. + \frac{1}{K}\left(\frac{\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})^2} - \frac{\phi'''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})}\right)\right)\frac{1}{K}\phi\left(\boldsymbol{w}_k^\top \boldsymbol{x}\right)\boldsymbol{x}, \qquad k = j
\end{aligned} \tag{B.22}$$

where we have simplified the presentation by setting $y = 1$, since y is a binary random variable, and we will show that in either case $|T_{j,j,k}|$ is upper bounded, i.e., in this case

$$|T_{j,j,k}| \leq \begin{cases} \max\left\{\frac{2}{K^3}\frac{1}{H(\tilde{\boldsymbol{W}})^3}, \frac{1}{K^2}\frac{1}{H(\tilde{\boldsymbol{W}})^2}\right\} & y = 1 \\ \max\left\{\frac{2}{K^3}\frac{1}{\left(1-H(\tilde{\boldsymbol{W}})\right)^3}, \frac{1}{K^2}\frac{1}{\left(1-H(\tilde{\boldsymbol{W}})\right)^2}\right\} & y = 0 \end{cases},$$

since $\phi(\cdot), \phi'(\cdot), \phi''(\cdot), \phi'''(\cdot)$ are bounded. More generally, by calculating the other case we can claim that

$$|T_{j,l,k}| \leq \max\left\{\frac{2}{K^3}\frac{1}{H(\tilde{\boldsymbol{W}})^3}, \frac{1}{K^2}\frac{1}{H(\tilde{\boldsymbol{W}})^2}, \frac{2}{K^3}\frac{1}{(1-H(\tilde{\boldsymbol{W}}))^3}, \frac{1}{K^2}\frac{1}{(1-H(\tilde{\boldsymbol{W}}))^2}\right\}, \tag{B.23}$$

holds for all j, l, k. Then, we can upper bound $\boldsymbol{a}_j^\top \boldsymbol{\Delta}_{j,l} \boldsymbol{a}_l$ using Cauchy-Schwarz inequality,

$$\begin{aligned}
\boldsymbol{a}_j^\top \boldsymbol{\Delta}_{j,l} \boldsymbol{a}_l &= \mathbb{E}\left[\left(\sum_{k=1}^K T_{j,l,k} \left\langle \boldsymbol{x}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle\right) \cdot \left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)\left(\boldsymbol{a}_l^\top \boldsymbol{x}\right)\right] \\
&\leq \sqrt{\mathbb{E}\left[\sum_{k=1}^K T_{j,l,k}^2\right] \cdot \mathbb{E}\left[\sum_{k=1}^K \left(\left\langle \boldsymbol{x}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle \left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)\left(\boldsymbol{a}_l^\top \boldsymbol{x}\right)\right)^2\right]} \\
&\leq \sqrt{\sum_{k=1}^K \mathbb{E}\left[T_{j,l,k}^2\right]} \cdot \sqrt{\sum_{k=1}^K \|\boldsymbol{w}_k - \boldsymbol{w}_k^\star\|_2^2 \cdot \|\boldsymbol{a}_j\|_2^2 \cdot \|\boldsymbol{a}_l\|_2^2}. \qquad \text{(B.24)}
\end{aligned}$$

Plug it back to (B.17) we can obtain the following inequality,

$$\begin{aligned}
&\|\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)\| \\
&\leq \max_{\|\boldsymbol{a}\|=1} \sum_{j=1}^K \sum_{l=1}^K \sqrt{\sum_{k=1}^K \mathbb{E}\left[T_{j,l,k}^2\right]} \cdot \sqrt{\sum_{k=1}^K \|\boldsymbol{w}_k - \boldsymbol{w}_k^\star\|_2^2 \cdot \|\boldsymbol{a}_j\|_2^2 \cdot \|\boldsymbol{a}_l\|_2^2}. \qquad \text{(B.25)}
\end{aligned}$$

Then the problem boils down to upper bound $\mathbb{E}\left[T_{i,j,k}^2\right]$, which we can apply the following lemma, whose proof can be found in Section B.4.5.

Lemma 9. *Let $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $t = \max\{\|\boldsymbol{w}_1\|_2, \cdots \|\boldsymbol{w}_K\|_2\}$ and $z \in \mathbb{Z}$ such that $z \geq 1$, for the sigmoid activation function $\phi(x) = \frac{1}{1+e^{-x}}$, the following*

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{\frac{1}{K}\sum_{j=1}^K \phi\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}\right)^z\right] \leq C_1 \cdot e^{t^2}, \\
&\mathbb{E}\left[\left(\frac{1}{\left(1 - \frac{1}{K}\sum_{j=1}^K \phi\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\right)}\right)^z\right] \leq C_2 \cdot e^{t^2} \qquad \text{(B.26)}
\end{aligned}$$

holds for some large enough constants C_1, C_2 that depend on the constant z.

Setting $z = 4$ and $z = 6$ in Lemma 9, together with (B.23) we obtain that

$$\mathbb{E}\left[T_{j,l,k}^2\right] \leq \frac{C}{K^4} \cdot e^{\max_{1\leq i\leq k} \|\tilde{\boldsymbol{w}}_i\|_2^2}, \qquad \text{(B.27)}$$

holds for some constant C. Plugging (B.27) into (B.25), we obtain

$$\begin{aligned}\|\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)\| &\leq \frac{C}{K^{\frac{3}{2}}} e^{\|\tilde{\boldsymbol{W}}\|_{\mathrm{F}}^2} \cdot \|\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}} \cdot \max_{\|\boldsymbol{a}\|=1} \sum_{j=1}^{K}\sum_{l=1}^{K} \|\boldsymbol{a}_j\|_2 \|\boldsymbol{a}_l\|_2 \\ &\leq \frac{C}{K^{\frac{3}{2}}} e^{\|\tilde{\boldsymbol{W}}\|_{\mathrm{F}}^2} \cdot \|\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}}. \end{aligned} \tag{B.28}$$

Further since $e^{\max_{1\leq i\leq k} \|\tilde{\boldsymbol{w}}_i\|_2^2} \leq C$ gives that $\|\boldsymbol{w}_i - \boldsymbol{w}_i^\star\|_2 \leq 0.7$, where we have used the assumption that $\max_{1\leq i\leq k} \|\boldsymbol{w}_i^\star\|_2^2 \leq 1$, we conclude that

$$\|\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)\| \leq \frac{C}{K^{\frac{3}{2}}} \|\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}} \tag{B.29}$$

holds for some constant C.

- **The CNN case:** according to (B.4), we can calculate the upper bound of $\|\nabla^2 f(\boldsymbol{w}) - \nabla^2 f(\boldsymbol{w}^\star)\|$ by definition as

$$\max_{\|\boldsymbol{u}\|_2=1} \sum_{j=1}^{K}\sum_{l=1}^{K} \mathbb{E}\left[\left(g_{j,l}(\boldsymbol{w}) - g_{j,l}(\boldsymbol{w}^\star)\right) \cdot \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \cdot \boldsymbol{x}^{(l)\top} \boldsymbol{u}\right]. \tag{B.30}$$

We then again apply the mean value theorem to $g_{j,l}(\boldsymbol{w})$, such that there exists $\tilde{\boldsymbol{w}} = \eta \boldsymbol{w} + (1-\eta)\boldsymbol{w}$ for some $\eta \in (0,1)$,

$$g_{j,l}(\boldsymbol{w}) - g_{j,l}(\boldsymbol{w}^\star) = \left\langle \nabla g_{j,l}(\tilde{\boldsymbol{w}}), \boldsymbol{w} - \boldsymbol{w}^\star \right\rangle.$$

Similarly to the FCN case, we can write $\nabla g_{j,l}(\tilde{\boldsymbol{w}})$ in the form of

$$\nabla g_{j,l}(\tilde{\boldsymbol{w}}) = \sum_{k=1}^{K} S_{j,l,k} \cdot \boldsymbol{x}^{(k)},$$

where $S_{j,l,k}$ is a scalar that depends on $\tilde{\boldsymbol{w}}$ and $\boldsymbol{x}^{(k)}, k = 1, \cdots, K$. Again we take $j \neq l$ as an example to calculate $S_{j,l,k}$, by definition, and obtain

$$\begin{aligned} K^2 \cdot \frac{\partial g_{j,l}(\boldsymbol{w})}{\partial \boldsymbol{w}} = &\frac{(1 - H(\boldsymbol{w}))\, \phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)}\right) \phi''\left(\boldsymbol{w}^\top \boldsymbol{x}^{(l)}\right)}{(1 - H(\boldsymbol{w}))^3} \cdot \boldsymbol{x}^{(l)} \\ &+ \frac{(1 - H(\boldsymbol{w}))\, \phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(l)}\right) \phi''\left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)}\right)}{(1 - H(\boldsymbol{w}))^3} \cdot \boldsymbol{x}^{(j)} \\ &- \frac{\phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(l)}\right) \phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)}\right)}{(1 - H(\boldsymbol{w}))^3} \cdot \left(\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}^{(k)}\right), \end{aligned} \tag{B.31}$$

where we set $y = 0$ for simplification. Then we obtain

$$S_{j,l,l} = \frac{1}{K^2} \frac{\left(1 - H\left(\boldsymbol{w}\right)\right) \phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)}\right) \phi''\left(\boldsymbol{w}^\top \boldsymbol{x}^{(l)}\right)}{\left(1 - H\left(\boldsymbol{w}\right)\right)^3} - \frac{1}{K^3} \frac{\phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(l)}\right) \phi'\left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)}\right)}{\left(1 - H\left(\boldsymbol{w}\right)\right)^3}. \tag{B.32}$$

and

$$|S_{j,l,l}| \leq \frac{1}{K^2} \frac{1}{\left(1 - H\left(\tilde{\boldsymbol{w}}\right)\right)^3}, \tag{B.33}$$

hold, where we used the fact that $0 \leq H\left(\boldsymbol{w}\right) \leq 1$ and $\phi'\left(\cdot\right), \phi''\left(\cdot\right)$ are bounded. Hence in the same way, we can obtain

$$|S_{j,l,k}| \leq \begin{cases} \max\left\{\frac{1}{K^2} \frac{1}{(1-H(\tilde{\boldsymbol{w}}))^3}, \frac{1}{K^2} \frac{1}{(H(\tilde{\boldsymbol{w}}))^3}\right\} & j \neq l \\ \max\left\{\frac{1}{K} \frac{1}{(1-H(\tilde{\boldsymbol{w}}))^2}, \frac{1}{K} \frac{1}{(H(\tilde{\boldsymbol{w}}))^2}\right\} & j = l \end{cases}. \tag{B.34}$$

Plug these back to (B.30) we obtain

$$\begin{aligned}
&\|\nabla^2 f\left(\boldsymbol{w}\right) - \nabla^2 f\left(\boldsymbol{w}^\star\right)\| \\
&\leq \max_{\|\boldsymbol{u}\|_2=1} \sum_{j=1}^{K} \sum_{l=1}^{K} \mathbb{E}\left[\sum_{k=1}^{K} \left\langle S_{j,l,k} \cdot \boldsymbol{x}^{(k)}, \boldsymbol{w} - \boldsymbol{w}^\star \right\rangle \cdot \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \cdot \boldsymbol{x}^{(l)\top} \boldsymbol{u}\right] \\
&= \max_{\|\boldsymbol{u}\|_2=1} \sum_{j=1}^{K} \sum_{l=1}^{K} \mathbb{E}\left[\sum_{k=1}^{K} S_{j,l,k} \cdot \left(\boldsymbol{w} - \boldsymbol{w}^\star\right)^\top \boldsymbol{x}^{(k)} \cdot \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \cdot \boldsymbol{x}^{(l)\top} \boldsymbol{u}\right] \\
&\leq \max_{\|\boldsymbol{u}\|_2=1} \sum_{j=1}^{K} \sum_{l=1}^{K} \sqrt{\mathbb{E}\left[\sum_{k=1}^{K} S_{j,l,k}^2\right] \cdot \mathbb{E}\left[\sum_{k=1}^{K} \left(\left(\boldsymbol{w} - \boldsymbol{w}^\star\right)^\top \boldsymbol{x}^{(k)}\right)^2 \left(\boldsymbol{u}^\top \boldsymbol{x}^{(j)}\right)^2 \left(\boldsymbol{x}^{(l)\top} \boldsymbol{u}\right)^2\right]} \\
&\leq \max_{\|\boldsymbol{u}\|_2=1} \sum_{j=1}^{K} \sum_{l=1}^{K} \sqrt{\mathbb{E}\left[\sum_{k=1}^{K} S_{j,l,k}^2\right] \cdot \sum_{k=1}^{K} \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 \cdot \|\boldsymbol{u}\|_2^2 \cdot \|\boldsymbol{u}\|_2^2} \\
&\leq C \cdot K \cdot e^{\|\tilde{\boldsymbol{w}}\|_2^2} \cdot \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2,
\end{aligned} \tag{B.35}$$

where the second inequality follows from Cauchy-Schwarz inequality, and the last inequality follows from (B.34) and Lemma 9. Further since $e^{\|\tilde{\boldsymbol{w}}\|_2^2} \leq C \cdot (1 + \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2)$ given that $\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \leq 0.7$, we conclude that

$$\|\nabla^2 f\left(\boldsymbol{w}\right) - \nabla^2 f\left(\boldsymbol{w}^\star\right)\| \leq C \cdot K \cdot \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \tag{B.36}$$

holds for some constant C and $\|\boldsymbol{w} - \boldsymbol{w}^\star\| \leq 0.7$.

B.4.2 Proof of Lemma 6

We first present upper and lower bounds on the Hessian $\nabla^2 f(\boldsymbol{W}^\star)$ of the population risk at ground truth, and then apply Lemma 5 to obtain a uniform bound in the neighborhood of $\boldsymbol{W}^\star$.

- **The FCN case:** Recall

$$\frac{\partial^2 f\left(\boldsymbol{W}^\star\right)}{\partial \boldsymbol{w}_j^2} = \mathbb{E}\left[\frac{1}{K^2}\cdot\left(\frac{\phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)^2}{H\left(\boldsymbol{W}^\star\right)\left(1-H\left(\boldsymbol{W}^\star\right)\right)}\right)\boldsymbol{x}\boldsymbol{x}^\top\right],$$
$$\frac{\partial^2 f\left(\boldsymbol{W}^\star\right)}{\partial \boldsymbol{w}_j \partial \boldsymbol{w}_l} = \mathbb{E}\left[\frac{1}{K^2}\cdot\left(\frac{\phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_l^{\star\top}\boldsymbol{x}\right)}{H\left(\boldsymbol{W}^\star\right)\left(1-H\left(\boldsymbol{W}^\star\right)\right)}\right)\boldsymbol{x}\boldsymbol{x}^\top\right],$$

where we have applied the fact that $\mathbb{E}\left[y|\boldsymbol{x}\right] = H\left(\boldsymbol{W}^\star\right)$. Let $\boldsymbol{a} = [\boldsymbol{a}_1^\top, \cdots, \boldsymbol{a}_K^\top]^\top \in \mathbb{R}^{dK}$. Then we can write

$$\begin{aligned}\nabla^2 f\left(\boldsymbol{W}^\star\right) &\succeq \left(\min_{\|\boldsymbol{a}\|_2=1} \boldsymbol{a}^\top \nabla^2 f\left(\boldsymbol{W}^\star\right)\boldsymbol{a}\right)\cdot \boldsymbol{I} \\ &= \min_{\|\boldsymbol{a}\|_2=1} \frac{1}{K^2}\mathbb{E}\left[\frac{\left(\sum_{j=1}^K \phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(\boldsymbol{a}_j^\top\boldsymbol{x}\right)\right)^2}{H\left(\boldsymbol{W}^\star\right)\left(1-H\left(\boldsymbol{W}^\star\right)\right)}\right]\cdot \boldsymbol{I}. \end{aligned} \tag{B.37}$$

Since $0 \le H\left(\boldsymbol{W}^\star\right) \le 1$, we have that $H\left(\boldsymbol{W}^\star\right)\left(1-H\left(\boldsymbol{W}^\star\right)\right) \le \frac{1}{4}$. Hence,

$$\nabla^2 f\left(\boldsymbol{W}^\star\right) \succeq \min_{\|\boldsymbol{a}\|_2=1} \frac{4}{K^2}\mathbb{E}\left[\left(\sum_{j=1}^K \phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(\boldsymbol{a}_j^\top\boldsymbol{x}\right)\right)^2\right]\cdot \boldsymbol{I} \succeq \frac{4}{K^2}\cdot\frac{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}{\kappa^2\lambda}\cdot\boldsymbol{I}, \tag{B.38}$$

where the last inequality follows from [7, Lemmas D.4 and D.6]. To derive an upper bound of $\nabla^2 f\left(\boldsymbol{W}^\star\right)$, we have

$$\begin{aligned}\nabla^2 f\left(\boldsymbol{W}^\star\right) &\preceq \left(\max_{\|\boldsymbol{a}\|_2=1} \boldsymbol{a}^\top \nabla^2 f\left(\boldsymbol{W}^\star\right)\boldsymbol{a}\right)\cdot \boldsymbol{I} \\ &= \max_{\|\boldsymbol{a}\|_2=1} \frac{1}{K^2}\mathbb{E}\left[\frac{\left(\sum_{j=1}^K \phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(\boldsymbol{a}_j^\top\boldsymbol{x}\right)\right)^2}{\frac{1}{K^2}\sum_{j,l}\phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1-\phi\left(\boldsymbol{w}_l^{\star\top}\boldsymbol{x}\right)\right)}\right]. \end{aligned} \tag{B.39}$$

Then by Cauchy-Schwarz inequality, we have

$$\frac{\left(\sum_{j=1}^{K}\phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(\boldsymbol{a}_j^{\top}\boldsymbol{x}\right)\right)^2}{\frac{1}{K^2}\sum_{j,l}\phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1-\phi\left(\boldsymbol{w}_l^{\star\top}\boldsymbol{x}\right)\right)} \leq \frac{\left(\sum_{j=1}^{K}\phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)^2\right)\cdot\left(\sum_{j=1}^{K}\left(\boldsymbol{a}_j^{\top}\boldsymbol{x}\right)^2\right)}{\frac{1}{K^2}\sum_{j,l}\phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1-\phi\left(\boldsymbol{w}_l^{\star\top}\boldsymbol{x}\right)\right)}.$$

Further since $\phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\leq\frac{1}{4}$, and

$$\begin{aligned}\sum_{j,l}\phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1-\phi\left(\boldsymbol{w}_l^{\star\top}\boldsymbol{x}\right)\right) &\geq \sum_{j=1}^{K}\phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1-\phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\right)\\ &= \sum_{j=1}^{K}\phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\\ &\geq 4\sum_{j=1}^{K}\phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)^2, \end{aligned} \tag{B.40}$$

we obtain

$$\boldsymbol{a}^{\top}\nabla^2 f\left(\boldsymbol{W}^{\star}\right)\boldsymbol{a} \preceq \frac{1}{K^2}\mathbb{E}\left[\frac{CK^2}{4}\sum_{j=1}^{K}\left(\boldsymbol{a}_j^{\top}\boldsymbol{x}\right)^2\right]. \tag{B.41}$$

Plugging (B.41) back to (B.39), we obtain

$$\nabla^2 f\left(\boldsymbol{W}^{\star}\right)\preceq C\cdot\boldsymbol{I}. \tag{B.42}$$

Thus together with the lower bound (B.38), we conclude that

$$\frac{4}{K^2}\cdot\frac{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}{\kappa^2\lambda}\cdot\boldsymbol{I}\preceq\nabla^2 f\left(\boldsymbol{W}^{\star}\right)\preceq C\cdot\boldsymbol{I}. \tag{B.43}$$

From Lemma 5, we have

$$\|\nabla^2 f(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W}^{\star})\|\lesssim\frac{C}{K^{\frac{3}{2}}}\|\boldsymbol{W}-\boldsymbol{W}^{\star}\|_F. \tag{B.44}$$

Therefore, if $\|\boldsymbol{W}^{\star}-\boldsymbol{W}\|_F\leq 0.7$ and

$$\frac{C}{K^{\frac{3}{2}}}\cdot\|\boldsymbol{W}-\boldsymbol{W}^{\star}\|_F\leq\frac{4}{K^2}\cdot\frac{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}{\kappa^2\lambda},$$

i.e., if $\|\boldsymbol{W}-\boldsymbol{W}^\star\|_F \le \min\left\{\frac{C}{K^{\frac{1}{2}}}\cdot\frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2\lambda}, 0.7\right\}$ for some constant C, we have

$$\begin{aligned}\sigma_{\min}\left(\nabla^2 f\left(\boldsymbol{W}\right)\right) &\ge \sigma_{\min}\left(\nabla^2 f\left(\boldsymbol{W}^\star\right)\right) - \|\nabla^2 f\left(\boldsymbol{W}\right)-\nabla^2 f\left(\boldsymbol{W}^\star\right)\| \\ &\gtrsim \frac{4}{K^2}\cdot\frac{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}{\kappa^2\lambda} - \frac{C}{K^{\frac{3}{2}}}\|\boldsymbol{W}-\boldsymbol{W}^\star\|_F \\ &\gtrsim \frac{4}{K^2}\cdot\frac{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}{\kappa^2\lambda}.\end{aligned}$$

Moreover, within the same neighborhood, by the triangle inequality we have

$$\|\nabla^2 f\left(\boldsymbol{W}\right)\| \le \|\nabla^2 f\left(\boldsymbol{W}\right)-\nabla^2 f\left(\boldsymbol{W}^\star\right)\| + \|\nabla^2 f\left(\boldsymbol{W}^\star\right)\| \lesssim C.$$

- **The CNN case:** Following from (B.4), we have

$$\nabla^2 f\left(\boldsymbol{w}^\star\right) = \mathbb{E}\left[\frac{\frac{1}{K^2}\sum_{j,l}\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(l)}\right)\boldsymbol{x}^{(j)}\boldsymbol{x}^{(l)\top}}{H\left(\boldsymbol{w}^\star\right)\left(1-H\left(\boldsymbol{w}^\star\right)\right)}\right]. \tag{B.45}$$

By definition, we lower bound $\nabla^2 f\left(\boldsymbol{w}^\star\right)$ by

$$\begin{aligned}&\min_{\|\boldsymbol{u}\|=1}\ \mathbb{E}\left[\frac{\frac{1}{K^2}\sum_{j,l}\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(l)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(l)}}{H\left(\boldsymbol{w}^\star\right)\left(1-H\left(\boldsymbol{w}^\star\right)\right)}\right]\cdot\boldsymbol{I} \\ &\succeq \min_{\|\boldsymbol{u}\|=1}\ \mathbb{E}\left[\frac{4}{K^2}\sum_{j,l}\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(l)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(l)}\right]\cdot\boldsymbol{I} \\ &= \frac{4}{K^2}\cdot\left(\min_{\|\boldsymbol{u}\|=1}\ \sum_{j\neq l}\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right]\cdot\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(l)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(l)}\right]\right. \\ &\qquad \left.+\sum_{j=1}^{K}\mathbb{E}\left[\left(\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right)^2\right]\right)\cdot\boldsymbol{I},\end{aligned}$$

where the last equality follows from the fact that $\boldsymbol{x}^{(j)}$ is independent from $\boldsymbol{x}^{(l)}$ given that $j\neq l$. Next we decompose $\boldsymbol{u}$ as $\boldsymbol{u}=\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^\star+\left(\boldsymbol{u}-\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^\star\right)$,

and calculate the expectation as

$$
\begin{aligned}
&\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right] \\
&= \mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\left(\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^\star+\left(\boldsymbol{u}-\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^\star\right)\right)^\top\boldsymbol{x}^{(j)}\right] \\
&= \mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right] \\
&\quad+\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\right]\cdot\mathbb{E}\left[\left(\boldsymbol{u}-\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^\star\right)^\top\boldsymbol{x}^{(j)}\right] \\
&= \frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right],
\end{aligned}
$$

where the second equality follows from the independence of $\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}$ and $\left(\boldsymbol{u}-\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^\star\right)^\top\boldsymbol{x}^{(j)}$. Hence,

$$
\begin{aligned}
&\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right]\cdot\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(l)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(l)}\right] \\
&= \left(\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\right)^2\left(\mathbb{E}\left[\phi'\left(z\right)z\right]\right)^2 \\
&= 0,
\end{aligned}
$$

where $z=\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\sim\mathcal{N}\left(0,\|\boldsymbol{w}^\star\|_2^2\right)$, and the last equality follows because $\phi'\left(z\right)z=-\left(\phi'\left(-z\right)\cdot\left(-z\right)\right)$. Similarly,

$$
\begin{aligned}
&\mathbb{E}\left[\left(\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right)^2\right] \\
&= \mathbb{E}\Bigg[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2\cdot\left(\left(\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2+\left(\left(\boldsymbol{u}-\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^\star\right)^\top\boldsymbol{x}^{(j)}\right)^2\right)\Bigg] \\
&= \left(\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\right)^2\cdot\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2\right] \\
&\quad+\left(\|\boldsymbol{u}\|_2^2-\frac{\left(\boldsymbol{u}^\top\boldsymbol{w}^\star\right)^2}{\|\boldsymbol{w}^\star\|_2^2}\right)\cdot\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2\right].
\end{aligned}
\tag{B.46}
$$

Together with Definition 2, we have

$$
\mathbb{E}\left[\left(\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right)^2\right]\geq\rho_{\mathrm{CNN}}\left(\|\boldsymbol{w}^\star\|_2\right).
\tag{B.47}
$$

Hence,

$$\nabla^2 f\left(\boldsymbol{w}^\star\right) \succeq \frac{4}{K} \cdot \rho_{\mathrm{CNN}}\left(\left\|\boldsymbol{w}^\star\right\|_2\right) \cdot \boldsymbol{I}. \tag{B.48}$$

Moreover, we apply Cauchy-Schwarz inequality and upper bound the Hessian as

$$\begin{aligned}\nabla^2 f\left(\boldsymbol{w}^\star\right) &\leq \left(\max_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top \nabla^2 f\left(\boldsymbol{w}^\star\right) \boldsymbol{u}\right) \cdot \boldsymbol{I} \\ &\leq \max_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\left[\frac{\sum_{j=1}^K \left(\frac{1}{K}\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\right)^2 \cdot \sum_{j=1}^K \left(\boldsymbol{u}^\top \boldsymbol{x}^{(j)}\right)^2}{H\left(\boldsymbol{w}^\star\right)\left(1-H\left(\boldsymbol{w}^\star\right)\right)}\right] \cdot \boldsymbol{I}. \end{aligned} \tag{B.49}$$

Using (B.40), i.e.,

$$\frac{\frac{1}{K^2}\sum_{j=1}^K \phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2}{H\left(\boldsymbol{w}^\star\right)\left(1-H\left(\boldsymbol{w}^\star\right)\right)} \leq \frac{1}{4}, \tag{B.50}$$

we upper bound the right-hand side of (B.49) as

$$\nabla^2 f\left(\boldsymbol{w}^\star\right) \preceq \max_{\|\boldsymbol{u}\|_2=1} \mathbb{E}\left[\frac{1}{4}\sum_{j=1}^K \left(\boldsymbol{u}^\top \boldsymbol{x}^{(j)}\right)^2\right] \cdot \boldsymbol{I} = \frac{K}{4} \cdot \boldsymbol{I}. \tag{B.51}$$

Together with the lower bound, we now conclude that

$$\frac{4}{K} \cdot \rho_{\mathrm{CNN}}\left(\left\|\boldsymbol{w}^\star\right\|_2\right) \cdot \boldsymbol{I} \preceq \nabla^2 f\left(\boldsymbol{w}^\star\right) \preceq \frac{K}{4} \cdot \boldsymbol{I}. \tag{B.52}$$

And following from (B.6) in Lemma 5, we have

$$\left\|\nabla^2 f\left(\boldsymbol{w}\right) - \nabla^2 f\left(\boldsymbol{w}^\star\right)\right\| \leq C \cdot K \cdot \left\|\boldsymbol{w} - \boldsymbol{w}^\star\right\|_2. \tag{B.53}$$

Thus if $\left\|\boldsymbol{w} - \boldsymbol{w}^\star\right\| \leq \min\left\{0.7, C \cdot \frac{\rho_{\mathrm{CNN}}\left(\left\|\boldsymbol{w}^\star\right\|_2\right)}{K^2}\right\}$, we have

$$C \cdot \frac{\rho_{\mathrm{CNN}}\left(\left\|\boldsymbol{w}^\star\right\|_2\right)}{K} \cdot \boldsymbol{I} \preceq \nabla^2 f\left(\boldsymbol{w}\right) \preceq C \cdot K \cdot \boldsymbol{I}. \tag{B.54}$$

B.4.3 Proof of Lemma 7

We apply a covering type of argument to show that the Hessian of the empirical risk function concentrates around the Hessian of the population risk function uniformly, and the argument applies to both the loss associated with FCN and CNN. We first take the FCN case as an example and then we provide the necessary modifications for the proof of the CNN case.

- **The FCN case:** We adapt the analysis in [50] to our setting. Let N_ϵ be the ϵ-covering number of the Euclidean ball $\mathbb{B}\left(\boldsymbol{W}^\star, r\right)$. Here, we omit the subscript FCN of r for simplicity. It is known that $\log N_\epsilon \leq dK \log\left(3r/\epsilon\right)$ [85]. Let $\mathcal{W}_\epsilon = \{\boldsymbol{W}_1, \cdots, \boldsymbol{W}_{N_\epsilon}\}$ be the ϵ-cover set with N_ϵ elements. For any $\boldsymbol{W} \in \mathbb{B}\left(\boldsymbol{W}^\star, r\right)$, let $j\left(\boldsymbol{W}\right) = \operatorname{argmin}_{j\in[N_\epsilon]} \left\|\boldsymbol{W} - \boldsymbol{W}_{j(\boldsymbol{W})}\right\|_{\mathrm{F}} \leq \epsilon$ for all $\boldsymbol{W} \in \mathbb{B}\left(\boldsymbol{W}^\star, r\right)$.

  For any $\boldsymbol{W} \in \mathbb{B}\left(\boldsymbol{W}^\star, r\right)$, we have

$$\begin{aligned}\left\|\nabla^2 f_n\left(\boldsymbol{W}\right) - \nabla^2 f(\boldsymbol{W})\right\| \leq \frac{1}{n}&\left\|\sum_{i=1}^n \left[\nabla^2 \ell\left(\boldsymbol{W}; \boldsymbol{x}_i\right) - \nabla^2 \ell\left(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i\right)\right]\right\| \\ &+ \left\|\frac{1}{n}\sum_{i=1}^n \nabla^2 \ell\left(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i\right) - \mathbb{E}\left[\nabla^2 \ell\left(\boldsymbol{W}_{j(\boldsymbol{w})}; \boldsymbol{x}\right)\right]\right\| \\ &+ \left\|\mathbb{E}\left[\nabla^2 \ell\left(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}\right)\right] - \mathbb{E}\left[\nabla^2 \ell\left(\boldsymbol{W}; \boldsymbol{x}\right)\right]\right\|.\end{aligned}$$

  Hence, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star, r)} \left\|\nabla^2 f_n\left(\boldsymbol{W}\right) - \nabla^2 f(\boldsymbol{W})\right\| \geq t\right) \leq \mathbb{P}\left(A_t\right) + \mathbb{P}\left(B_t\right) + \mathbb{P}\left(C_t\right),$$

where the events A_t, B_t and C_t are defined as

$$A_t = \left\{ \sup_{\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^\star, r)} \frac{1}{n} \left\| \sum_{i=1}^{n} \left[\nabla^2 \ell \left(\boldsymbol{W}; \boldsymbol{x}_i \right) - \nabla^2 \ell \left(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i \right) \right] \right\| \geq \frac{t}{3} \right\},$$

$$B_t = \left\{ \sup_{\boldsymbol{W} \in \mathcal{W}_\epsilon} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell \left(\boldsymbol{W}; \boldsymbol{x}_i \right) - \mathbb{E} \left[\nabla^2 \ell \left(\boldsymbol{W}; \boldsymbol{x} \right) \right] \right\| \geq \frac{t}{3} \right\},$$

$$C_t = \left\{ \sup_{\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^\star, r)} \left\| \mathbb{E} \left[\nabla^2 \ell \left(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x} \right) \right] - \mathbb{E} \left[\nabla^2 \ell \left(\boldsymbol{W}; \boldsymbol{x} \right) \right] \right\| \geq \frac{t}{3} \right\}.$$

In the sequel, we bound the terms $\mathbb{P}(A_t)$, $\mathbb{P}(B_t)$, and $\mathbb{P}(C_t)$, separately.

1. **Upper bound on $\mathbb{P}(B_t)$.** Before continuing, we state a useful technical lemma, whose proof can be found in [50].

   Lemma 10. *Let $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ be a symmetric $d \times d$ matrix and V_ϵ be an ϵ-cover of unit-Euclidean-norm ball $\mathbb{B}(\boldsymbol{0}, 1)$, then*

   $$\|\boldsymbol{M}\| \leq \frac{1}{1 - 2\epsilon} \sup_{\boldsymbol{v} \in V_\epsilon} |\langle \boldsymbol{v}, \boldsymbol{M}\boldsymbol{v} \rangle|. \tag{B.55}$$

   Let $V_{\frac{1}{4}}$ be a $\left(\frac{1}{4}\right)$-cover of the ball $\mathbb{B}(\boldsymbol{0}, 1) = \{\boldsymbol{W} \in \mathbb{R}^{d \times K} : \|\boldsymbol{W}\|_{\mathrm{F}} = 1\}$, where $\log |V_{\frac{1}{4}}| \leq dK \log 12$. Following from Lemma 10, we have

   $$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell \left(\boldsymbol{W}; \boldsymbol{x}_i \right) - \mathbb{E} \left[\nabla^2 \ell \left(\boldsymbol{W}; \boldsymbol{x} \right) \right] \right\| \\ &\leq 2 \sup_{\boldsymbol{v} \in V_{\frac{1}{4}}} \left| \left\langle \boldsymbol{v}, \left(\frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell \left(\boldsymbol{W}; \boldsymbol{x}_i \right) - \mathbb{E} \left[\nabla^2 \ell \left(\boldsymbol{W}; \boldsymbol{x} \right) \right] \right) \boldsymbol{v} \right\rangle \right|. \end{aligned}$$

   Taking the union bound over $\mathcal{W}_\epsilon$ and $V_{\frac{1}{4}}$ yields

   $$\begin{aligned} \mathbb{P}(B_t) &\leq \mathbb{P} \left(\sup_{\boldsymbol{W} \in \mathcal{W}_\epsilon, \boldsymbol{v} \in V_{\frac{1}{4}}} \left| \frac{1}{n} \sum_{i=1}^{n} G_i \right| \geq \frac{t}{6} \right) \\ &\leq e^{dK \left(\log \frac{3r}{\epsilon} + \log 12 \right)} \sup_{\boldsymbol{W} \in \mathcal{W}_\epsilon, \boldsymbol{v} \in V_{\frac{1}{4}}} \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^{n} G_i \right| \geq \frac{t}{6} \right), \end{aligned} \tag{B.56}$$

where $G_i = \langle \boldsymbol{v}, (\nabla^2 \ell(\boldsymbol{W}; \boldsymbol{x}_i) - \mathbb{E}[\nabla^2 \ell(\boldsymbol{W}; \boldsymbol{x})])\boldsymbol{v} \rangle$ and $\mathbb{E}[G_i] = 0$. Let $\boldsymbol{a} = \left[\boldsymbol{a}_1^\top, \cdots, \boldsymbol{a}_K^\top\right] \in \mathbb{R}^{dK}$. Then we can show that $\|G_i\|_{\psi_1}$ is upper bounded, which we summariz as follows, and whose proof is given in Appendix B.4.6.

Lemma 11. *Suppose the loss is associated with FCN. There exists some constant C such that*

$$\|G_i\|_{\psi_1} \leq C :\equiv \tau^2.$$

Applying the Bernstein inequality for sub-exponential random variables [50, Theorem 9] to (B.56), we have that for fixed $\boldsymbol{W} \in \mathcal{W}_\epsilon, \boldsymbol{v} \in V_{\frac{1}{4}}$,

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \left\langle \boldsymbol{v}, \left(\nabla^2 \ell(\boldsymbol{W}; \boldsymbol{x}_i) - \mathbb{E}\left[\nabla^2 \ell(\boldsymbol{W}; \boldsymbol{x})\right]\right)\boldsymbol{v}\right\rangle\right| \geq \frac{t}{6}\right) \\ &\leq 2\exp\left(-c \cdot n \cdot \min\left(\frac{t^2}{\tau^4}, \frac{t}{\tau^2}\right)\right), \end{aligned} \tag{B.57}$$

for some universal constant c. As a result, $\mathbb{P}(B_t)$ is upper bounded by

$$2\exp\left(-c \cdot n \cdot \min\left(\frac{t^2}{\tau^4}, \frac{t}{\tau^2}\right) + dK \log\frac{3r}{\epsilon} + dK \log 12\right).$$

Thus as long as

$$t > C \cdot \max\left\{\sqrt{\frac{\tau^4\left(dK \log\frac{36r}{\epsilon} + \log\frac{4}{\delta}\right)}{n}}, \frac{\tau^2\left(dK \log\frac{36r}{\epsilon} + \log\frac{4}{\delta}\right)}{n}\right\} \tag{B.58}$$

for some large enough constant C, we have $\mathbb{P}(B_t) \leq \frac{\delta}{2}$.

2. **Upper bound on $\mathbb{P}(A_t)$ and $\mathbb{P}(C_t)$.** These two events will be bounded in a similar way. We first present the following useful Lemma, whose proof is provided in Appendix B.4.8

Lemma 12. *Suppose the loss is associated with FCN. There exists some constant C such that*

$$\mathbb{E}\left[\sup_{\boldsymbol{W} \neq \boldsymbol{W}' \in \mathbb{B}(\boldsymbol{W}^\star, r)} \frac{\|\nabla^2 \ell(\boldsymbol{W}, \boldsymbol{x}) - \nabla^2 \ell(\boldsymbol{W}', \boldsymbol{x})\|}{\|\boldsymbol{W} - \boldsymbol{W}'\|_{\mathrm{F}}}\right] \leq C \cdot d\sqrt{K}. \tag{B.59}$$

Consider the event C_t first. We derive

$$\begin{aligned}
&\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)} \left\|\mathbb{E}\left[\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}\right)\right]-\mathbb{E}\left[\nabla^2\ell\left(\boldsymbol{W};\boldsymbol{x}\right)\right]\right\| \\
&\leq \sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)} \frac{\left\|\mathbb{E}\left[\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}\right)\right]-\mathbb{E}\left[\nabla^2\ell\left(\boldsymbol{W};\boldsymbol{x}\right)\right]\right\|}{\|\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}\|_{\mathrm{F}}} \\
&\quad\cdot \sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)} \|\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}\|_{\mathrm{F}} \\
&\leq C\cdot d\sqrt{K}\cdot\epsilon. \qquad \text{(B.60)}
\end{aligned}$$

Therefore, C_t holds as long as

$$t \geq C\cdot d\sqrt{K}\cdot\epsilon. \qquad \text{(B.61)}$$

We can bound the event A_t as below.

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)} \frac{1}{n}\left\|\sum_{i=1}^n\left[\nabla^2\ell\left(\boldsymbol{W};\boldsymbol{x}_i\right)-\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i\right)\right]\right\| \geq \frac{t}{3}\right) \\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)} \left\|\frac{1}{n}\sum_{i=1}^n\left[\nabla^2\ell\left(\boldsymbol{W};\boldsymbol{x}_i\right)-\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i\right)\right]\right\|\right] \qquad \text{(B.62)} \\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)} \left\|\nabla^2\ell\left(\boldsymbol{W};\boldsymbol{x}_i\right)-\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i\right)\right\|\right] \\
&\leq \frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)} \frac{\left\|\nabla^2\ell\left(\boldsymbol{W};\boldsymbol{x}_i\right)-\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i\right)\right\|}{\|\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}\|_{\mathrm{F}}}\right] \\
&\quad\cdot \sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)} \|\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}\|_{\mathrm{F}} \\
&\leq \frac{C\cdot d\sqrt{K}\cdot\epsilon}{t} \qquad \text{(B.63)}
\end{aligned}$$

where (B.62) follows from the Markov inequality. Thus, taking

$$t \geq \frac{6\epsilon\cdot C\cdot d\sqrt{K}}{\delta} \qquad \text{(B.64)}$$

ensures that $\mathbb{P}\left(A_t\right)\leq\frac{\delta}{2}$.

3. **Final step.** Let $\epsilon = \frac{\delta\tau^2}{C\cdot d\sqrt{K}\cdot ndK}$ and $\delta = d^{-10}$. Plugging ϵ and δ into (B.58) we need

$$t > \tau^2 \cdot \max\Bigg\{\frac{1}{ndK}, C\cdot\sqrt{\frac{\left(dK\log(36rnd^{11}K)+\log\frac{4}{\delta}\right)}{n}}, \frac{\left(dK\log(36rnd^{11}K)+\log\frac{4}{\delta}\right)}{n}\Bigg\}.$$

The middle term can be bounded as

$$\frac{dK\log(36rnd^{11}K)+10\log d}{n} \le \frac{dK\log n}{n}+\frac{dK\log 36r}{n}+\frac{11dK\log dK}{n} + \frac{10\log d}{n}.$$

If $n \ge C\cdot dK\log(dK)$ for some large enough constant C, the first term $dK\log n$ dominants and is on the order of $dK\log(dK)$. Moreover, it decreases as n increases when $n \ge 3$. Thus we can set

$$t \ge \tau^2\sqrt{\frac{\left(dK\log(36rnd^{11}K)+\log\frac{4}{\delta}\right)}{n}} \tag{B.65}$$

which holds as $t \ge C'\cdot\tau^2\sqrt{\frac{dK\log n}{n}}$ for some constant C'. By setting $t := C\tau^2\sqrt{\frac{dK\log n}{n}}$ for sufficiently large C, as long as $n \ge C'\cdot dK\log dK$,

$$\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\left\|\nabla^2 f_n(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W})\right\| \ge C\tau^2\sqrt{\frac{dK\log n}{n}}\right) \le d^{-10}.$$

- **The CNN case:** If the loss is associated with CNN, we redefine G_i as $G_i = \langle \boldsymbol{v}, (\nabla^2\ell(\boldsymbol{w};\boldsymbol{x}_i) - \mathbb{E}[\nabla^2\ell(\boldsymbol{w};\boldsymbol{x})])\,\boldsymbol{v}\rangle$ and we show the following Lemmas whose proof is given in Appendix B.4.7 and Appendix B.4.9 .

Lemma 13. *Suppose the loss is associated CNN. There exists some constant C such that*

$$\|G_i\|_{\psi_1} \le C\cdot K^2 :\equiv \tau^2. \tag{B.66}$$

Lemma 14. *Suppose the loss is associated with CNN. There exists some constant C such that*

$$\mathbb{E}\left[\sup_{\boldsymbol{W}\neq\boldsymbol{W}'\in\mathbb{B}(\boldsymbol{W}^\star,r)}\frac{\|\nabla^2\ell\left(\boldsymbol{W},\boldsymbol{x}\right)-\nabla^2\ell\left(\boldsymbol{W}',\boldsymbol{x}\right)\|}{\|\boldsymbol{W}-\boldsymbol{W}'\|_{\mathrm{F}}}\right]\leq C\cdot d\sqrt{K}. \tag{B.67}$$

Following argument similar to the proof of Lemma 7, we can obtain the following concentration inequality:

$$\sup_{\boldsymbol{w}\in\mathbb{B}(\boldsymbol{w}^\star,r)}\|\nabla^2 f_n\left(\boldsymbol{w}\right)-\nabla^2 f\left(\boldsymbol{w}\right)\|\leq C\cdot K^2\sqrt{\frac{\frac{d}{K}\cdot\log n}{n}}, \tag{B.68}$$

holds with probability at least $1-d^{-10}$, as long as the sample complexity $n\geq C\cdot\frac{d}{K}\log\left(\frac{d}{K}\right)$.

B.4.4 Proof of Lemma 8

In order to proceed we need the following Lemma 15 whose proof is given in Appendix B.4.10.

Lemma 15. *Suppose the loss is associated with FCN. Let $\boldsymbol{u}$ be a fixed unit norm vector $\boldsymbol{u}=\left[\boldsymbol{u}_1^\top,\cdots,\boldsymbol{u}_K^\top\right]\in\mathbb{R}^{dK}$ with $\|\boldsymbol{u}\|_2=1$. Then we have*

$$\|\boldsymbol{u}^\top\nabla\ell\left(\boldsymbol{W};\boldsymbol{x}\right)\|_{\psi_2}\leq\sqrt{K}.$$

Suppose the loss is associated with CNN. Let $\boldsymbol{u}$ be a fixed unit norm vector $\boldsymbol{u}\in\mathbb{R}^m$ with $\|\boldsymbol{u}\|_2=1$. Then

$$\|\left\langle\boldsymbol{u},\nabla\ell\left(\boldsymbol{w}\right)\right\rangle\|_{\psi_2}\leq C\cdot K.$$

Following argument (details omitted) similar to the proof of Lemma 7, and applies Lemma 15, for the loss associated with FCN, we can get the following concentration

inequality

$$\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r_{\mathrm{FCN}})} \|\nabla f_n(\boldsymbol{W}) - \nabla f(\boldsymbol{W})\|_2 \le C\cdot\sqrt{\frac{d\sqrt{K}\log n}{n}} \tag{B.69}$$

with probability at least $1-d^{-10}$, as long as the sample size $n \ge C\cdot dK\log(dK)$. For the loss associated with CNN, we obtain

$$\sup_{\boldsymbol{w}\in\mathbb{B}(\boldsymbol{w}^\star,r_{\mathrm{CNN}})} \|\nabla f_n(\boldsymbol{w}) - \nabla f(\boldsymbol{w})\| \le C\cdot\sqrt{K}\sqrt{\frac{\frac{d}{K}\log n}{n}} = C\cdot\sqrt{\frac{d\log n}{n}}, \tag{B.70}$$

with probability at least $1-d^{-10}$ as long as $n \ge C\cdot\frac{d}{K}\log\left(\frac{d}{K}\right)$.

B.4.5 Proof of Lemma 9

We take the first term in (B.26) as an example, since the second term follows exactly in the same way. We first derive

$$\mathbb{E}\left[\left(\frac{1}{K}\sum_{i=1}^{K}\phi\left(\boldsymbol{w}_i^\top\boldsymbol{x}\right)\right)^{-z}\right] \le \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\left(\phi\left(\boldsymbol{w}_i^\top\boldsymbol{x}\right)\right)^{-z}\right], \tag{B.71}$$

which follows from the fact that $f(x) = x^{-z}$ is convex for $x>0$ and $z\ge 1$. Further since $\frac{1}{\phi(x)} = 1+e^{-x}$, and $g = \boldsymbol{w}_i^\top\boldsymbol{x} \sim \mathcal{N}(0,\sigma_i^2 = \|\boldsymbol{w}_i\|_2^2)$, we can exactly calculate the summands in the above equation as follows:

$$\mathbb{E}\left[\phi(g)^{-z}\right] = \mathbb{E}\left[\sum_{l=0}^{z}\binom{z}{l}e^{-lg}\right] = \sum_{l=0}^{z}\binom{z}{l}e^{\left(\frac{\sigma_i^2 l^2}{2}\right)},$$

where we use the fact that g is a Gaussian random variable. Hence, we conclude that for $t = \max\left(\|\boldsymbol{w}_1\|_2,\cdots,\|\boldsymbol{w}_K\|_2\right)$ and $p\ge 1$,

$$\mathbb{E}\left[\left(\frac{1}{\frac{1}{K}\sum_{i=1}^{K}\phi\left(\boldsymbol{w}_i^\top\boldsymbol{x}\right)}\right)^{z}\right] \le C\cdot e^{t^2}, \tag{B.72}$$

holds for some constant C depending on z.

B.4.6 Proof of Lemma 11

The sub-exponential norm of G_i can be bounded as

$$\|G_i\|_{\psi_1} \leq \| \left\langle \boldsymbol{u}, \nabla^2 \ell \left(\boldsymbol{W}; z \right) \boldsymbol{u} \right\rangle \|_{\psi_1} + \| \nabla^2 f \left(\boldsymbol{W}; z \right) \|,$$

where $\| \nabla^2 f \left(\boldsymbol{W}; z \right) \|$ is upper bounded by C due to Lemma 6. Denote the (j,l)-th block of $\nabla^2 \ell \left(\boldsymbol{W}; z \right)$ as $\xi_{j,l} \cdot \boldsymbol{x}\boldsymbol{x}^\top$. We can derive

$$\begin{aligned} \| \left\langle \boldsymbol{u}, \nabla^2 \ell \left(\boldsymbol{W}; z \right) \boldsymbol{u} \right\rangle \|_{\psi_1} &\leq \sum_{j=1}^{K} \sum_{l=1}^{K} \| \xi_{j,l} \cdot \boldsymbol{u}_j^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u}_l \|_{\psi_1} \\ &\leq \sum_{j=1}^{K} \sum_{l=1}^{K} \sup_{t \geq 1} \quad t^{-1} \left(\mathbb{E} \left| \xi_{j,l} \cdot \boldsymbol{u}_j^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u}_l \right|^t \right)^{\frac{1}{t}}. \end{aligned} \tag{B.73}$$

Next we show that $\xi_{j,l}$ is upper bounded by some constant for all j and l.

- For $j \neq l$,

$$\begin{aligned} |\xi_{j,l}| &= \left| \frac{1}{K^2} \frac{\phi' \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \phi' \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right) \cdot \left(H \left(\boldsymbol{W} \right)^2 + y - 2y \cdot H \left(\boldsymbol{W} \right) \right)}{\left(H \left(\boldsymbol{W} \right) \left(1 - H \left(\boldsymbol{W} \right) \right) \right)^2} \right| \\ &= \begin{cases} \frac{1}{K^2} \frac{\phi' \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \phi' \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right)}{\left(1 - H(\boldsymbol{W}) \right)^2} & y = 0 \\ \frac{1}{K^2} \frac{\phi' \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \phi' \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right)}{H(\boldsymbol{W})^2} & y = 1 \end{cases}. \end{aligned} \tag{B.74}$$

  Moreover,

$$\begin{aligned} \frac{1}{K^2} \frac{\phi' \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \phi' \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right)}{\left(1 - H \left(\boldsymbol{W} \right) \right)^2} \leq \frac{\phi' \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \phi' \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right)}{\left(1 - \phi \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \right) \left(1 - \phi \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right) \right)} &\leq \phi \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \phi \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right) \\ &\leq 1, \end{aligned} \tag{B.75}$$

  where the first inequality holds due to the following fact,

$$\left(1 - H \left(\boldsymbol{W} \right) \right)^2 = \left(1 - \frac{1}{K} \sum_{j=1}^{K} \phi \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \right)^2 \geq \frac{1}{K^2} \left(1 - \phi \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \right) \left(1 - \phi \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right) \right),$$

  the second inequality follows because $\phi \left(x \right) \left(1 - \phi \left(x \right) \right) = \phi' \left(x \right)$. Similarly, we can show that

$$\frac{1}{K^2} \frac{\phi' \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right) \phi' \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right)}{H \left(\boldsymbol{W} \right)^2} \leq 1. \tag{B.76}$$

  Thus for $j \neq l$, $|\xi_{j,l}| \leq 1$ holds.

- For $j = l$,

$$|\xi_{j,j}| \leq \left| \frac{1}{K^2} \frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right) \phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right) \cdot \left(H\left(\boldsymbol{W}\right)^2 + y - 2y \cdot H\left(\boldsymbol{W}\right)\right)}{\left(H\left(\boldsymbol{W}\right)\left(1 - H\left(\boldsymbol{W}\right)\right)\right)^2} \right| + \left| \frac{1}{K} \frac{\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\left(y - H\left(\boldsymbol{W}\right)\right)}{H\left(\boldsymbol{W}\right)\left(1 - H\left(\boldsymbol{W}\right)\right)} \right|.$$

  For the second term in the above equation, we have

$$\left| \frac{1}{K} \frac{\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\left(y - H\left(\boldsymbol{W}\right)\right)}{H\left(\boldsymbol{W}\right)\left(1 - H\left(\boldsymbol{W}\right)\right)} \right| = \begin{cases} \frac{1}{K} \frac{\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{\left(1 - H(\boldsymbol{W})\right)} \leq 1 & y = 0 \\ \frac{1}{K} \frac{\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})} \leq 1 & y = 1 \end{cases},$$

  which follows from the fact that the second derivative is

$$\phi''\left(x\right) = \phi\left(x\right)\left(1 - \phi\left(x\right)\right)\left(1 - 2\phi\left(x\right)\right),$$

  the absolute value of which can be upper bounded by $\phi\left(x\right)$ or $1 - \phi\left(x\right)$.

Hence,

$$\begin{aligned} &\left\| \left\langle \boldsymbol{u}, \nabla^2 \ell\left(\boldsymbol{W}; z\right) \boldsymbol{u} \right\rangle \right\|_{\psi_1} \\ &\leq C \cdot \sum_{j=1}^{K} \sum_{l=1}^{K} \sup_{t \geq 1} \quad t^{-1} \left(\sqrt{\mathbb{E}\left[\left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^{2t}\right]} \cdot \sqrt{\mathbb{E}\left[\left(\boldsymbol{u}_l^\top \boldsymbol{x}\right)^{2t}\right]} \right)^{\frac{1}{t}} \\ &\leq C \cdot \sum_{j=1}^{K} \sum_{l=1}^{K} \|\boldsymbol{u}_j\|_2 \|\boldsymbol{u}_l\|_2 \cdot \sup_{t \geq 1} \quad t^{-1} \left((2t-1)!!\right)^{\frac{1}{t}} \\ &\leq C :\equiv \tau^2, \end{aligned} \tag{B.77}$$

where the last inequality holds because

$$\begin{aligned} &\sup_{t \geq 1} \quad t^{-1} \left((2t-1)!!\right)^{\frac{1}{t}} \leq \sup_{t \geq 1} \quad t^{-1} \left((2t)^t\right)^{\frac{1}{t}} \leq 2, \\ &\sum_{j=1}^{K} \sum_{l=1}^{K} \|\boldsymbol{u}_j\|_2 \|\boldsymbol{u}_l\|_2 \leq \sum_{j=1}^{K} \sum_{l=1}^{K} \frac{\|\boldsymbol{u}_j\|_2^2 + \|\boldsymbol{u}_l\|_2^2}{2} = \frac{1}{2}. \end{aligned} \tag{B.78}$$

Thus, we conclude

$$\|G_i\|_{\psi_1} \leq C :\equiv \tau^2.$$

B.4.7 Proof of Lemma 13

Again the sub-exponential norm of G_i can be bounded as

$$\|G_i\|_{\psi_1} \leq \| \left\langle \boldsymbol{u}, \nabla^2 \ell \left(\boldsymbol{w}; z \right) \boldsymbol{u} \right\rangle \|_{\psi_1} + \| \nabla^2 f \left(\boldsymbol{w}; z \right) \|,$$

where $\| \nabla^2 f \left(\boldsymbol{W}; z \right) \|$ is upper bounded by $C \cdot K$ due to Lemma 6. Applying the triangle inequality, the sub-exponential norm of $\left\langle \boldsymbol{u}, \nabla^2 \ell \left(\boldsymbol{w} \right) \boldsymbol{u} \right\rangle$ can be bounded as

$$\| \left\langle \boldsymbol{u}, \nabla^2 \ell \left(\boldsymbol{w} \right) \boldsymbol{u} \right\rangle \|_{\psi_1} \leq \sum_{j \neq l} \| g_{j,l} \left(\boldsymbol{w} \right) \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \boldsymbol{u}^\top \boldsymbol{x}^{(l)} \|_{\psi_1} + \sum_{j=l} \| g_{j,l} \left(\boldsymbol{w} \right) \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \boldsymbol{u}^\top \boldsymbol{x}^{(l)} \|_{\psi_1}. \tag{B.79}$$

Hence, we have

$$\begin{aligned} & \left| \frac{1}{K^2} \frac{H \left(\boldsymbol{w} \right)^2 + y - 2y \cdot H \left(\boldsymbol{w} \right)}{\left(H \left(\boldsymbol{w} \right) \left(1 - H \left(\boldsymbol{w} \right) \right) \right)^2} \phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right) \phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(l)} \right) \right| \\ = & \begin{cases} \frac{1}{K^2} \frac{\phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right) \phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(l)} \right)}{H(\boldsymbol{w})^2} \leq 1 & y = 1 \\ \frac{1}{K^2} \frac{\phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right) \phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(l)} \right)}{(1 - H(\boldsymbol{w}))^2} \leq 1 & y = 0 \end{cases}, \end{aligned}$$

and

$$\left| \frac{1}{K} \frac{y - H \left(\boldsymbol{w} \right)}{H \left(\boldsymbol{w} \right) \left(1 - H \left(\boldsymbol{w} \right) \right)} \phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right) \right| = \begin{cases} \frac{1}{K} \frac{\phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right)}{H(\boldsymbol{w})} \leq 1 & y = 1 \\ \frac{1}{K} \frac{\phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right)}{1 - H(\boldsymbol{w})} \leq 1 & y = 0 \end{cases}.$$

Plugging it back to (B.79), we obtain

$$\| \left\langle \boldsymbol{u}, \nabla^2 \ell \left(\boldsymbol{w} \right) \boldsymbol{u} \right\rangle \|_{\psi_1} \leq \sum_{j \neq l} \| \left(\boldsymbol{u}^\top \boldsymbol{x}^{(j)} \right) \left(\boldsymbol{u}^\top \boldsymbol{x}^{(l)} \right) \|_{\psi_1} + \sum_{j=1}^{K} \| \left(\boldsymbol{u}^\top \boldsymbol{x}^{(j)} \right)^2 \|_{\psi_1} \leq C \cdot K^2. \tag{B.80}$$

B.4.8 Proof of Lemma 12

As noted before, we can write the (j,l)-th block of $\nabla^2 \ell \left(\boldsymbol{W}; \boldsymbol{z} \right)$ as $\xi_{j,l} \left(\boldsymbol{W} \right) \boldsymbol{x} \boldsymbol{x}^\top$. Then we can obtain the following bound,

$$\| \nabla^2 \ell \left(\boldsymbol{W}; z \right) - \nabla^2 \ell \left(\boldsymbol{W}'; z \right) \| \leq \sum_{j=1}^{K} \sum_{l=1}^{K} | \xi_{j,l} \left(\boldsymbol{W} \right) - \xi_{j,l} \left(\boldsymbol{W}' \right) | \cdot \| \boldsymbol{x} \boldsymbol{x}^\top \|. \tag{B.81}$$

Using the same method as shown in the proof of Lemma 5, we can upper bound $|\xi_{j,l}(\boldsymbol{W}) - \xi_{j,l}(\boldsymbol{W}')|$ as

$$|\xi_{j,l}(\boldsymbol{W}) - \xi_{j,l}(\boldsymbol{W}')| \leq \left(\max_k |T_{j,l,k}|\right) \cdot \|\boldsymbol{x}\|_2 \cdot \sqrt{K} \cdot \|\boldsymbol{W} - \boldsymbol{W}'\|_F,$$

where following from (B.23),

$$|T_{j,l,k}| \leq \max\left\{\frac{2}{K^3}\frac{1}{H(\boldsymbol{W})^3}, \frac{1}{K^2}\frac{1}{H(\boldsymbol{W})^2}, \frac{2}{K^3}\frac{1}{(1-H(\boldsymbol{W}))^3}, \frac{1}{K^2}\frac{1}{(1-H(\boldsymbol{W}))^2}\right\}. \tag{B.82}$$

And thus, if $\|\boldsymbol{W} - \boldsymbol{W}'\|_F \leq 0.7$ we have

$$\begin{aligned}\mathbb{E}\left[\sup_{\boldsymbol{W}\neq\boldsymbol{W}'} \frac{\|\nabla^2\ell(\boldsymbol{W}) - \nabla^2\ell(\boldsymbol{W}')\|}{\|\boldsymbol{W} - \boldsymbol{W}'\|_F}\right] &\leq \sqrt{K}\cdot K^2 \cdot \mathbb{E}\left[\left(\max_{j,l,k}|T_{j,l,k}|\right)\cdot\|\boldsymbol{x}\|_2\cdot\|\boldsymbol{x}\boldsymbol{x}^\top\|\right] \\ &\leq C\cdot d\sqrt{K}. \end{aligned} \tag{B.83}$$

Thus we only need to set $J^\star \geq C\cdot d\sqrt{K}$ for some large enough C.

B.4.9 Proof of Lemma 14

Following from (B.4) we can write

$$\|\nabla^2\ell(\boldsymbol{w}) - \nabla^2\ell(\boldsymbol{w}')\| \leq \sum_{j=1}^K\sum_{l=1}^K |g_{j,l}(\boldsymbol{w}) - g_{j,l}(\boldsymbol{w}')|\cdot\|\boldsymbol{x}^{(j)}\boldsymbol{x}^{(l)\top}\|. \tag{B.84}$$

Similarly, the analysis in the proof of Lemma 5 implies that

$$|g_{j,l}(\boldsymbol{w}) - g_{j,l}(\boldsymbol{w}')| \leq \left(\max_k |S_{j,l,k}|\right)\cdot\sqrt{K}\|\boldsymbol{x}\|_2\cdot\|\boldsymbol{w} - \boldsymbol{w}'\|_2, \tag{B.85}$$

where we upper-bound $S_{j,l,k}$ in (B.34) as

$$|S_{j,l,k}| \leq \begin{cases}\max\left\{\frac{1}{K^2}\frac{1}{(1-H(\boldsymbol{w}))^3}, \frac{1}{K^2}\frac{1}{(H(\boldsymbol{w}))^3}\right\} & j\neq l \\ \max\left\{\frac{1}{K}\frac{1}{(1-H(\boldsymbol{w}))^2}, \frac{1}{K}\frac{1}{(H(\boldsymbol{w}))^2}\right\} & j = l\end{cases}. \tag{B.86}$$

Hence, if $\|\boldsymbol{w} - \boldsymbol{w}'\|_2 \leq 0.7$, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{\boldsymbol{w}\neq\boldsymbol{w}'} \frac{\|\nabla^2\ell(\boldsymbol{w}) - \nabla^2\ell(\boldsymbol{w}')\|}{\|\boldsymbol{w} - \boldsymbol{w}'\|_F}\right] &\leq \sqrt{K}\cdot\sum_{j=1}^K\sum_{l=1}^K \mathbb{E}\left[\left(\max_k|S_{j,l,k}|\right)\cdot\|\boldsymbol{x}\|_2\cdot\|\boldsymbol{x}^{(j)}\boldsymbol{x}^{(l)\top}\|\right] \\ &\leq C\cdot d\sqrt{K}. \end{aligned} \tag{B.87}$$

Thus, in this case we can set $J^\star \geq C\cdot d\sqrt{K}$ as well.

B.4.10 Proof of Lemma 15

- **The FCN case:** Following from (B.1), we have

$$\langle \nabla \ell \left(\boldsymbol{W} \right), \boldsymbol{u} \rangle = \frac{1}{K} \sum_{j=1}^{K} \left(\frac{\left(y - H\left(\boldsymbol{W} \right) \right) \cdot \phi' \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right)}{H\left(\boldsymbol{W} \right) \left(1 - H\left(\boldsymbol{W} \right) \right)} \right) \left(\boldsymbol{u}_j^\top \boldsymbol{x} \right),$$

and by definition, we can upper-bound the sub-Gaussian norm as

$$\begin{aligned} & \| \langle \nabla \ell \left(\boldsymbol{W} \right), \boldsymbol{u} \rangle \|_{\psi_2} \\ & \leq \begin{cases} \frac{1}{K} \sum_{j=1}^{K} \left\| \frac{\phi' \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right)}{\left(1 - \frac{1}{K} \sum_{l=1}^{K} \phi \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right) \right)} \boldsymbol{u}_j^\top \boldsymbol{x} \right\|_{\psi_2} \leq \sum_{j=1}^{K} \| \boldsymbol{u}_j^\top \boldsymbol{x} \|_{\psi_2} & y = 0 \\ \frac{1}{K} \sum_{j=1}^{K} \left\| \frac{\phi' \left(\boldsymbol{w}_j^\top \boldsymbol{x} \right)}{\frac{1}{K} \sum_{l=1}^{K} \phi \left(\boldsymbol{w}_l^\top \boldsymbol{x} \right)} \boldsymbol{u}_j^\top \boldsymbol{x} \right\|_{\psi_2} \leq \sum_{j=1}^{K} \| \boldsymbol{u}_j^\top \boldsymbol{x} \|_{\psi_2} & y = 1 \end{cases}. \end{aligned}$$

Thus we conclude that

$$\| \langle \nabla \ell \left(\boldsymbol{W} \right), \boldsymbol{u} \rangle \|_{\psi_2} \leq \sum_{j=1}^{K} \| \boldsymbol{u}_j \|_2 \leq \sqrt{K}, \tag{B.88}$$

and the directional gradient is $\sqrt{K}$-sub-Gaussian.

- **The CNN case:** Following from (B.3), we have

$$\langle \nabla \ell \left(\boldsymbol{w} \right), \boldsymbol{u} \rangle = - \sum_{j=1}^{K} \frac{1}{K} \phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right) \frac{y - H\left(\boldsymbol{w} \right)}{H\left(\boldsymbol{w} \right) \left(1 - H\left(\boldsymbol{w} \right) \right)} \cdot \left(\boldsymbol{u}^\top \boldsymbol{x}^{(j)} \right),$$

where

$$\left| \phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right) \frac{y - H\left(\boldsymbol{w} \right)}{H\left(\boldsymbol{w} \right) \left(1 - H\left(\boldsymbol{w} \right) \right)} \right| = \begin{cases} \frac{\phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right)}{\sum_{j=1}^{K} \frac{1}{K} \phi \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right)} \leq K & y = 1 \\ \frac{\phi' \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right)}{\sum_{j=1}^{K} \frac{1}{K} \left(1 - \phi \left(\boldsymbol{w}^\top \boldsymbol{x}^{(j)} \right) \right)} \leq K & y = 0 \end{cases}.$$

Then the sub-Gaussian norm of $\langle \nabla \ell \left(\boldsymbol{w} \right), \boldsymbol{u} \rangle$ is upper bounded as

$$\| \langle \nabla \ell \left(\boldsymbol{w} \right), \boldsymbol{u} \rangle \|_{\psi_2} \leq K \cdot \frac{1}{K} \sum_{j=1}^{K} \| \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \|_{\psi_2} \leq C \cdot K. \tag{B.89}$$

Hence, the directional gradient is K-sub-Gaussian.

B.5 Proof of Theorem 5

We first define a product $\widetilde{\otimes}$ as follows. If $\boldsymbol{v} \in \mathbb{R}^d$ is a vector and $\boldsymbol{I}$ is the identity matrix, then $\boldsymbol{v}\widetilde{\otimes}\boldsymbol{I} = \sum_{j=1}^{d}[\boldsymbol{v}\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_j + \boldsymbol{e}_j\otimes\boldsymbol{v}\otimes\boldsymbol{e}_j + \boldsymbol{e}_j\otimes\boldsymbol{e}_j\otimes\boldsymbol{v}]$. If $\boldsymbol{M}$ is a symmetric rank-r matrix factorized as $\boldsymbol{M} = \sum_{i=1}^{r}\boldsymbol{s}_i\boldsymbol{v}_i\boldsymbol{v}_i^\top$ and $\boldsymbol{I}$ is the identity matrix, then

$$\boldsymbol{M}\widetilde{\otimes}\boldsymbol{I} = \sum_{i=1}^{r}\boldsymbol{s}_i\sum_{j=1}^{d}\sum_{l=1}^{6}\boldsymbol{A}_{l,i,j}, \tag{B.90}$$

where $\boldsymbol{A}_{1,i,j} = \boldsymbol{v}_i\otimes\boldsymbol{v}_i\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_j$, $\boldsymbol{A}_{2,i,j} = \boldsymbol{v}_i\otimes\boldsymbol{e}_j\otimes\boldsymbol{v}_i\otimes\boldsymbol{e}_j$, $\boldsymbol{A}_{3,i,j} = \boldsymbol{e}_j\otimes\boldsymbol{v}_i\otimes\boldsymbol{v}_i\otimes\boldsymbol{e}_j$, $\boldsymbol{A}_{4,i,j} = \boldsymbol{v}_i\otimes\boldsymbol{e}_j\otimes\boldsymbol{e}_j\otimes\boldsymbol{v}_i$, $\boldsymbol{A}_{5,i,j} = \boldsymbol{e}_j\otimes\boldsymbol{v}_i\otimes\boldsymbol{e}_j\otimes\boldsymbol{v}_i$ and $\boldsymbol{A}_{6,i,j} = \boldsymbol{e}_j\otimes\boldsymbol{e}_j\otimes\boldsymbol{v}_i\otimes\boldsymbol{v}_i$.

And We further define a tensor operation as follows. For a tensor $\boldsymbol{T} \in \mathbb{R}^{n_1\times n_2\times n_3}$ and three matrices $\boldsymbol{A} \in \mathbb{R}^{n_1\times d_1}, \boldsymbol{B} \in \mathbb{R}^{n_2\times d_2}, \boldsymbol{C} \in \mathbb{R}^{n_3\times d_3}$, the (i,j,k)-th entry of the tensor $\boldsymbol{T}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C})$ is given by

$$\sum_{i'}^{n_1}\sum_{j'}^{n_2}\sum_{k'}^{n_3}\boldsymbol{T}_{i',j',k'}\boldsymbol{A}_{i',i}\boldsymbol{B}_{j',j}\boldsymbol{C}_{k',k}. \tag{B.91}$$

The proof contains two parts. Part (a) proves that the estimation of the direction of $\boldsymbol{W}^\star$ is sufficiently accurate, which follows from the arguments similar to those in [7] and is only briefly summarized below. Part (b) is different, where we do not require the homogeneous condition for the activation function, and instead, our proof is based on a mild condition in Assumption 2. We detail our proof in part (b).

(a) In order to estimate the direction of each $\boldsymbol{w}_i$ for $i = 1,\ldots,K$, [7] showed that for the regression problem, if the sample size $n \geq d\text{poly}(K,\kappa,\zeta,\log d)$, where $\zeta > 1$ is any constant, then

$$\|\overline{\boldsymbol{w}_i}^\star - s_i\boldsymbol{V}\widehat{\boldsymbol{u}}_i\| \leq \epsilon\text{poly}(\mathrm{K},\kappa) \tag{B.92}$$

holds with probability at least $1 - d^{-\Omega(\zeta)}$. Such a result also holds for the classification problem with only slight difference in the proof as we describe as follows. The main

idea of the proof is to bound the estimation error of $\boldsymbol{P}_2$ and $\boldsymbol{R}_3$ via Bernstein inequality. For the regression problem, Bernstein inequality was applied to terms associated with each neuron individually, and the bounds were then put together via the triangle inequality in [7]. However, for the classification problem here, we apply Bernstein inequality to the terms associated with all neurons together. Another difference is that the label y_i of the classification model is bounded by nature, whereas the output y_i in the regression model needs to be upper-bounded via homogeneously bounded conditions of the activation function. A reader can refer to [7] for the details of the proof for this part.

(b) In order to estimate $\|\boldsymbol{w}_i\|$ for $i = 1, \ldots, K$, we provide a different proof from [7], which does not require the homogeneous condition on the activation function, but assumes a more relaxed condition in Assumption 2.

We define a quantity Q_1 as follows:

$$Q_1 = \boldsymbol{M}_{l_1}(\boldsymbol{I}, \underbrace{\boldsymbol{\alpha}, \cdots, \boldsymbol{\alpha}}_{(l_1-1)}), \tag{B.93}$$

where l_1 is the first non-zero index such that $\boldsymbol{M}_{l_1} \neq 0$. For example, if $l_1 = 3$, then Q_1 takes the following form

$$Q_1 = \boldsymbol{M}_3(\boldsymbol{I}, \boldsymbol{\alpha}, \boldsymbol{\alpha}) = \frac{1}{K}\sum_{i=1}^{K} m_{3,i}(\|\boldsymbol{w}_i^\star\|)\left(\boldsymbol{\alpha}^\top \overline{\boldsymbol{w}}_i^\star\right)^2 \overline{\boldsymbol{w}}_i^\star, \tag{B.94}$$

where $\overline{\boldsymbol{w}} = \boldsymbol{w}/\|\boldsymbol{w}\|$ and by definition

$$m_{3,i}(\|\boldsymbol{w}_i^\star\|) = \mathbb{E}\left[\phi\left(\|\boldsymbol{w}_i^\star\| \cdot z\right) z^3\right] - 3\mathbb{E}\left[\phi\left(\|\boldsymbol{w}_i^\star\| \cdot z\right) z\right]. \tag{B.95}$$

Clearly, Q_1 has information of $\|\boldsymbol{w}_i^\star\|$, which can be estimated by solving the following optimization problem:

$$\beta^\star = \mathrm{argmin}_{\beta \in \mathbb{R}^K} \left\| \frac{1}{K}\sum_{i=1}^{K} \beta_i s_i \overline{\boldsymbol{w}_i}^\star - Q_1 \right\|, \tag{B.96}$$

where each entry of the solution takes the form

$$\beta_i^\star = s_i^3 m_{3,i}(\|\boldsymbol{w}_i^\star\|) \left(\boldsymbol{\alpha}^T s_i \overline{\boldsymbol{w}_i}^\star\right)^2. \tag{B.97}$$

In the initialization, we substitute $\widehat{Q}_1$ (estimated from training data) for Q_1, $\boldsymbol{V}\hat{u}_i$ (estimated in part (a)) for $s_i\overline{\boldsymbol{w}_i}^\star$ into (B.96), and obtain an estimate $\hat{\beta}$ of $\beta^\star$. We then substitute $\hat{\beta}$ for $\beta^\star$ and $\boldsymbol{V}\hat{u}_i$ for $s_i\overline{\boldsymbol{w}_i}^\star$ into (B.97) to obtain an estimate $\hat{a}_i$ of $\|\boldsymbol{w}_i^\star\|$ via the following equation

$$\hat{\beta}_i = s_i^3 m_{3,i}(\hat{a}_i) \left(\boldsymbol{\alpha}^\top \boldsymbol{V}\hat{u}_i\right)^2. \tag{B.98}$$

Furthermore, since $m_{l_1,i}(x)$ has fixed sign for $x > 0$ and for $l_1 \geq 1$, s_i can be estimated correctly from the sign of $\hat{\beta}_i$ for $i = 1, \ldots, K$.

For notational simplicity, let $\beta_{1,i}^\star := \frac{\beta_i^\star}{s_i^3\left(\boldsymbol{\alpha}^\top s_i \overline{\boldsymbol{w}_i}^\star\right)^2}$ and $\hat{\beta}_{1,i} := \frac{\hat{\beta}_i}{s_i^3\left(\boldsymbol{\alpha}^\top \boldsymbol{V}\hat{u}_i\right)^2}$, and then (B.97) and (B.98) become

$$\hat{\beta}_{1,i} = m_{3,i}(\hat{a}_i), \quad \beta_{1,i}^\star = m_{3,i}(\|\boldsymbol{w}_i^\star\|). \tag{B.99}$$

By Assumption 2 and (B.97), there exists a constant $\delta' > 0$ such that the inverse function $g(\cdot)$ of $m_{3,1}(\cdot)$ has upper-bounded derivative in the interval $(\beta_{1,i}^\star - \delta', \beta_{1,i}^\star + \delta')$, i.e., $|g'(x)| < \Gamma$ for a constant Γ. By employing the result in [7], if the sample size $n \geq d\text{poly}\,(K, \kappa, t, \log d)$, then $\widehat{Q}_1$ and Q_1, $\boldsymbol{V}\hat{u}_i$ and $s_i\overline{\boldsymbol{w}_i}^\star$ can be arbitrarily close so that $|\beta_{1,i}^\star - \hat{\beta}_{1,i}| < \min\{\delta', \frac{r}{\sqrt{K}\Gamma}\}$.

Thus, by (B.99) and the mean value theorem, we obtain

$$|\hat{a}_i - \|\boldsymbol{w}_i^\star\|| = |g'(\xi)||\beta_{1,i}^\star - \hat{\beta}_{1,i}|, \tag{B.100}$$

where ξ is between $\beta_{1,i}^\star$ and $\hat{\beta}_{1,i}$, and hence $|g'(\xi)| < \Gamma$. Therefore, $|\hat{a}_i - \|\boldsymbol{w}_i^\star\|| \leq \frac{r}{\sqrt{K}}$, which is the desired result.

Appendix C: Proofs for Chapter 4

C.1 Preliminary

For convenience, we introduce some useful results, for which the proofs of our main result will rely on.

Lemma 16. *Let $\boldsymbol{z} \in \mathbb{R}^d$ be a standard Gaussian random vector, then for any fixed unit vector $\boldsymbol{u} \in \mathbb{R}^d$, i.e., $\|\boldsymbol{u}\|_2 = 1$ and a non-zero fixed vector $\boldsymbol{w} \in \mathbb{R}^d$, we have the following identities,*

$$\mathbb{E}\left[\boldsymbol{u}^\top \boldsymbol{z} \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] = \frac{1}{\sqrt{2\pi}} \frac{\boldsymbol{u}^\top \boldsymbol{w}}{\|\boldsymbol{w}\|_2}, \tag{C.1}$$

$$\mathbb{E}\left[\left(\boldsymbol{u}^\top \boldsymbol{z}\right)^2 \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] = \frac{1}{2}. \tag{C.2}$$

Proof. The basic idea is to decompose $\boldsymbol{u}$ into an orthogonal pair, i.e., $\boldsymbol{u} = \boldsymbol{u}^\perp + \boldsymbol{u}^\parallel$, where $\boldsymbol{u}^\parallel = \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|_2^2}\boldsymbol{u}$ is the projection of $\boldsymbol{u}$ onto $\boldsymbol{w}$, and $\boldsymbol{u}^\perp = \left(\boldsymbol{I} - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|_2^2}\right)\boldsymbol{u}$ is the projection of $\boldsymbol{u}$ onto the corresponding complementary subspace. Then for (C.1) we have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{u}^\top \boldsymbol{z} \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] &= \mathbb{E}\left[\boldsymbol{z}^\top \left(\boldsymbol{u}^\perp + \boldsymbol{u}^\parallel\right) \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] \\
&= \mathbb{E}\left[\boldsymbol{z}^\top \boldsymbol{u}^\parallel \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] \\
&= \frac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{w}\|_2^2} \cdot \mathbb{E}\left[\boldsymbol{z}^\top \boldsymbol{w} \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right],
\end{aligned}$$

where for the second equality we have used the fact that $\boldsymbol{z}^\top \boldsymbol{u}^\perp$ is a zero mean Gaussian random variable and it is independent of $\boldsymbol{z}^\top \boldsymbol{w}$. Let $x = \boldsymbol{w}^\top \boldsymbol{z} \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \|\boldsymbol{w}\|_2^2$, then the result follows by evaluating the expectation:

$$\mathbb{E}\left[\boldsymbol{z}^\top \boldsymbol{w} \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] = \int_0^\infty x \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \frac{\sigma}{\sqrt{2\pi}} = \frac{\|\boldsymbol{w}\|_2}{\sqrt{2\pi}}.$$

Similarly, we can calculate (C.2) as

$$\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{u}^\top \boldsymbol{z}\right)^2 \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] &= \mathbb{E}\left[\left(\left(\boldsymbol{z}^\top \boldsymbol{u}^{\|}\right)^2 + \left(\boldsymbol{z}^\top \boldsymbol{u}^\perp\right)^2\right) \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{z}^\top \boldsymbol{u}^{\|}\right)^2 \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] + \mathbb{E}\left[\left(\boldsymbol{z}^\top \boldsymbol{u}^\perp\right)^2 \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] \quad \text{(C.3)} \\
&= \left(\frac{\boldsymbol{w}^\top \boldsymbol{u}}{\|\boldsymbol{w}\|_2^2}\right)^2 \cdot \mathbb{E}\left[\left(\boldsymbol{z}^\top \boldsymbol{w}\right)^2 \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] \\
&\quad + \mathbb{E}\left[\left(\boldsymbol{z}^\top \boldsymbol{u}^\perp\right)^2\right] \cdot \mathbb{E}\left[1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right], \quad \text{(C.4)}
\end{aligned}$$

and calculate the three expectations as follows:

$$\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{z}^\top \boldsymbol{w}\right)^2 \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] &= \int_0^\infty x^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \cdot \sigma^2 \cdot \int_0^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\
&= \frac{\sigma^2}{2} \\
&= \frac{\|\boldsymbol{w}\|_2^2}{2},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[1_{\left\{\boldsymbol{w}^\top \boldsymbol{z} > 0\right\}}\right] &= \mathbb{P}\left(\boldsymbol{w}^\top \boldsymbol{z} > 0\right) = \frac{1}{2}, \\
\mathbb{E}\left[\left(\boldsymbol{z}^\top \boldsymbol{u}^\perp\right)^2\right] &= \|\boldsymbol{u}^\perp\|_2^2 = 1 - \frac{\left(\boldsymbol{u}^\top \boldsymbol{w}\right)^2}{\|\boldsymbol{w}\|_2^2}.
\end{aligned}$$

Hence, the result follows by plugging the above results to (C.4). □

Lemma 17 (Lemma 3.1 in [86]). *Let $\boldsymbol{x}_i \in \mathbb{R}^m$ be i.i.d. Gaussian random vectors, then for any $0 < \epsilon < 1$, if $n > c_0 m \epsilon^{-2}$,*

$$(1-\epsilon)\left\|\boldsymbol{h}\right\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{x}_i^\top \boldsymbol{h}\right)^2 \leq (1+\epsilon)\left\|\boldsymbol{h}\right\|_2^2 \tag{C.5}$$

holds for all non-zero vectors $\boldsymbol{h} \in \mathbb{R}^m$ with probability at least $1 - 2\exp\left(-c_1\epsilon^2 n\right)$, and c_0, c_1 are some constants.

Lemma 18. *Denote $\boldsymbol{h} = \boldsymbol{w} - \boldsymbol{w}^\star$, and let $\rho = \frac{\boldsymbol{w}^\top \boldsymbol{w}^\star}{\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2}$. If $\|\boldsymbol{h}\|_2 = \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 = \nu \cdot \|\boldsymbol{w}^\star\|_2$ where $0 \leq \nu < 1$, then we have*

$$\rho \geq 1 - \nu^2 \tag{C.6}$$

holds.

Proof. We first show that the assumption $\nu < 1$ implies $\rho > 0$ since

$$\|\boldsymbol{h}\|_2^2 = \|\boldsymbol{w}^\star\|_2^2 + \|\boldsymbol{w}\|_2^2 - 2\rho\|\boldsymbol{w}^\star\|_2\|\boldsymbol{w}\|_2 < \|\boldsymbol{w}^\star\|_2^2, \tag{C.7}$$

gives us

$$\rho > \frac{\|\boldsymbol{w}\|_2}{2\|\boldsymbol{w}^\star\|_2} > 0. \tag{C.8}$$

We next show a tighter lower bound of ρ. Let α be the angle between $\boldsymbol{h}$ and $\boldsymbol{w}^\star$, i.e., $\cos(\alpha) = \frac{\boldsymbol{h}^\top \boldsymbol{w}^\star}{\|\boldsymbol{h}\|_2\|\boldsymbol{w}^\star\|_2}$. We can calculate ρ as

$$\rho = \frac{\boldsymbol{w}^\top \boldsymbol{w}^\star}{\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2} = \frac{\boldsymbol{h}^\top \boldsymbol{w}^\star + \|\boldsymbol{w}^\star\|_2^2}{\|\boldsymbol{w}^\star + \boldsymbol{h}\|_2\|\boldsymbol{w}^\star\|_2} = \frac{\cos(\alpha)\|\boldsymbol{h}\|_2 + \|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{h} + \boldsymbol{w}^\star\|_2},$$

square the two sides will give us,

$$\rho^2 = \frac{\cos(\alpha)^2 \cdot \|\boldsymbol{h}\|_2^2 + \|\boldsymbol{w}^\star\|_2^2 + 2\cos(\alpha)\|\boldsymbol{h}\|_2\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}^\star\|_2^2 + \|\boldsymbol{h}\|_2^2 + 2\cos(\alpha)\|\boldsymbol{w}^\star\|_2\|\boldsymbol{h}\|_2} = 1 - \frac{(1-\cos^2(\alpha))\cdot\nu^2}{1+\nu^2+2\nu\cos(\alpha)}. \tag{C.9}$$

Moreover, when $\cos(\alpha)^2 \neq 1$,

$$\frac{(1-\cos^2(\alpha))\cdot\nu^2}{1+\nu^2+2\nu\cos(\alpha)} = \frac{\nu^2}{1+\frac{(\nu+\cos(\alpha))^2}{1-\cos^2(\alpha)}} \leq \nu^2,$$

where the equality holds when $\cos(\alpha) = -\nu$, and when $\cos(\alpha)^2 = 1$, $\rho^2 = 1$. Together we can conclude that

$$1 \geq \rho \geq \rho^2 \geq 1-\nu^2. \tag{C.10}$$

□

C.2 Proof of GD on the Population Risk

C.2.1 Proof of Theorem 6

Proof. According to the definition of strong convexity, we can calculate the Hessian of the population risk function and verify it is positive definite in certain region. We next derive the Hessian and check its spectrum. First of all, according to [71], we can write the population risk (up to additive factors in $\boldsymbol{w}^\star$) and its gradient as

$$\begin{aligned}
L(\boldsymbol{w}) &= \frac{1}{K^2}\Bigg[\left(\frac{K^2-K}{2\pi}+\frac{K}{2}\right)\|\boldsymbol{w}\|_2^2 - \frac{K}{\pi}\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2\left(\sin(\theta)+(\pi-\theta)\cos(\theta)\right) \\
&\quad - \frac{K^2-K}{\pi}\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2\Bigg], \\
\nabla L(\boldsymbol{w}) &= \left(\frac{1}{K}+\frac{1-\frac{1}{K}}{\pi}\right)\cdot\boldsymbol{w} - \frac{\|\boldsymbol{w}^\star\|_2}{\pi K}\cdot\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}\cdot\sin(\theta) - \frac{1-\frac{1}{K}}{\pi}\cdot\|\boldsymbol{w}^\star\|_2\cdot\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \\
&\quad - \frac{\pi-\theta}{K\pi}\cdot\boldsymbol{w}^\star.
\end{aligned}$$

When $\theta \neq 0$ or π and $\boldsymbol{w} \neq \mathbf{0}$, we can calculate the Hessian as

$$
\begin{aligned}
&\nabla^2 L(\boldsymbol{w}) \\
&= \left(\frac{1}{K} + \frac{1-\frac{1}{K}}{\pi}\right) \cdot \boldsymbol{I} - \frac{\|\boldsymbol{w}^\star\|_2}{\pi K} \cdot \left[\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \cdot \cos\theta \cdot \frac{\partial \theta}{\partial \boldsymbol{w}} + \sin\theta \left(\frac{1}{\|\boldsymbol{w}\|_2} \cdot \boldsymbol{I} - \frac{1}{\|\boldsymbol{w}\|_2^3} \cdot \boldsymbol{w}\boldsymbol{w}^\top\right)\right] \\
&\quad - \frac{1-\frac{1}{K}}{\pi} \cdot \|\boldsymbol{w}^\star\|_2 \left(\frac{1}{\|\boldsymbol{w}\|_2} \cdot \boldsymbol{I} - \frac{1}{\|\boldsymbol{w}\|_2^3} \cdot \boldsymbol{w}\boldsymbol{w}^\top\right) + \frac{1}{K\pi} \cdot \boldsymbol{w}^\star \cdot \frac{\partial \theta}{\partial \boldsymbol{w}} \\
&= \underbrace{\left(\frac{1}{K} + \frac{1-\frac{1}{K}}{\pi} - \frac{\sin(\theta)}{\pi K} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} - \frac{1-\frac{1}{K}}{\pi} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2}\right) \cdot \boldsymbol{I}}_{\boldsymbol{H}_1} \\
&\quad + \underbrace{\frac{1}{\pi K} \cdot \frac{\cos(\theta)}{\sin(\theta)} \cdot \frac{1}{\|\boldsymbol{w}\|_2^2} \cdot \left(\boldsymbol{w}^\star \boldsymbol{w}^\top + \boldsymbol{w}\boldsymbol{w}^{\star\top}\right)}_{\boldsymbol{H}_2} \\
&\quad + \underbrace{\left(\frac{1-\frac{1}{K}}{\pi} - \frac{1}{\pi K} \cdot \frac{\cos(2\theta)}{\sin(\theta)}\right) \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2^3} \cdot \boldsymbol{w}\boldsymbol{w}^\top}_{\boldsymbol{H}_3} \\
&\quad \underbrace{- \frac{1}{K\pi} \frac{1}{\sin(\theta)} \cdot \frac{1}{\|\boldsymbol{w}\|_2 \|\boldsymbol{w}^\star\|_2} \boldsymbol{w}^\star \boldsymbol{w}^{\star\top}}_{\boldsymbol{H}_4}.
\end{aligned}
\tag{C.11}
$$

It is easy to show that the first term $\boldsymbol{H}_1$ is positively definite under some mild conditions. The difficulty actually lies in lower bounding the spectrum of the last three terms. We then lower bound the smallest eigenvalue of $\boldsymbol{H}_2 + \boldsymbol{H}_3 + \boldsymbol{H}_4$ as

following,

$$
\begin{aligned}
&\min_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top \left(\boldsymbol{H}_2+\boldsymbol{H}_3+\boldsymbol{H}_4\right)\boldsymbol{u} \\
&= \min_{\|\boldsymbol{u}\|_2=1} \frac{2}{\pi K}\cdot\frac{\cos(\theta)}{\sin(\theta)}\cdot\frac{1}{\|\boldsymbol{w}\|_2^2}\cdot\boldsymbol{u}^\top\boldsymbol{w}\cdot\boldsymbol{u}^\top\boldsymbol{w}^\star - \frac{1}{K\pi}\frac{1}{\sin(\theta)}\cdot\frac{\left(\boldsymbol{u}^\top\boldsymbol{w}^\star\right)^2}{\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2} \\
&\quad + \left(\frac{1-\frac{1}{K}}{\pi} - \frac{1}{\pi K}\cdot\frac{\cos(2\theta)}{\sin(\theta)}\right)\cdot\frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2^3}\cdot\left(\boldsymbol{u}^\top\boldsymbol{w}\right)^2 \\
&= \min_{\|\boldsymbol{u}\|_2=1} \frac{2}{\pi K}\cdot\frac{\cos(\theta)}{\sin(\theta)}\cdot\frac{1}{\|\boldsymbol{w}\|_2^2}\cdot\boldsymbol{u}^\top\boldsymbol{w}\cdot\boldsymbol{u}^\top\boldsymbol{w}^\star - \frac{1}{\pi K}\cdot\frac{\cos(2\theta)}{\sin(\theta)}\cdot\frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2^3}\cdot\left(\boldsymbol{u}^\top\boldsymbol{w}\right)^2 \\
&\quad - \frac{1}{K\pi}\frac{1}{\sin(\theta)}\cdot\frac{\left(\boldsymbol{u}^\top\boldsymbol{w}^\star\right)^2}{\|\boldsymbol{w}\|_2\|\boldsymbol{w}^\star\|_2} + \frac{1-\frac{1}{K}}{\pi}\frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2^3}\cdot\left(\boldsymbol{u}^\top\boldsymbol{w}\right)^2 \\
&= \min_{\|\boldsymbol{u}\|_2=1} \frac{1}{\pi K}\cdot\frac{1}{\sin(\theta)}\cdot\frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2}\Big[2\cos(\theta)\,\boldsymbol{u}^\top\bar{\boldsymbol{w}}\cdot\boldsymbol{u}^\top\bar{\boldsymbol{w}^\star} - \cos(\theta)^2\left(\boldsymbol{u}^\top\bar{\boldsymbol{w}}\right)^2 \\
&\quad + \sin(\theta)^2\left(\boldsymbol{u}^\top\bar{\boldsymbol{w}}\right)^2 - \left(\boldsymbol{u}^\top\bar{\boldsymbol{w}^\star}\right)^2\Big] + \frac{1-\frac{1}{K}}{\pi}\cdot\frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2}\left(\boldsymbol{u}^\top\bar{\boldsymbol{w}}\right)^2 \\
&= \min_{\|\boldsymbol{u}\|_2=1} \frac{1}{\pi K}\cdot\frac{1}{\sin(\theta)}\cdot\frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2}\Big[\sin(\theta)^2\left(\boldsymbol{u}^\top\bar{\boldsymbol{w}}\right)^2 - \left(\boldsymbol{u}^\top\left(\boldsymbol{I}-\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^\top\right)\bar{\boldsymbol{w}^\star}\right)^2 \\
&+ (K-1)\sin(\theta)\left(\boldsymbol{u}^\top\bar{\boldsymbol{w}}\right)^2\Big] \qquad \text{(C.12)} \\
&\geq \frac{1}{\pi K}\cdot\frac{1}{\sin(\theta)}\cdot\frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2}\cdot\left(-\sin(\theta)^2\right) \\
&= -\frac{\sin(\theta)}{\pi K}\cdot\frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2}
\end{aligned}
$$

where we use $\bar{\boldsymbol{w}}$ to denote $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, and note that $\bar{\boldsymbol{w}} \perp \left(\boldsymbol{I}-\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^\top\right)\bar{\boldsymbol{w}^\star}$, hence setting $\boldsymbol{u}$ in the direction of $\left(\boldsymbol{I}-\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^\top\right)\bar{\boldsymbol{w}^\star}$ will achieve the minimum of (C.12). Moreover,

$$
\max_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top\left(\boldsymbol{H}_2+\boldsymbol{H}_3+\boldsymbol{H}_4\right)\boldsymbol{u} \leq \frac{1}{\pi}\frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} \qquad \text{(C.13)}
$$

holds by setting $\boldsymbol{u}$ in the direction of $\bar{\boldsymbol{w}}$. Hence if the condition $\|\boldsymbol{w}\|_2 \geq \frac{K+1}{K-1+\frac{3}{4}\pi}\|\boldsymbol{w}^\star\|_2$ holds, we have the eigenvalue of the Hessian is lower bounded by some positive value,

$$\begin{aligned}
&\min_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top \boldsymbol{H}_1 \boldsymbol{u} + \boldsymbol{u}^\top \left(\boldsymbol{H}_2 + \boldsymbol{H}_3 + \boldsymbol{H}_4\right) \boldsymbol{u} \\
&\geq \frac{1}{K} + \frac{1-\frac{1}{K}}{\pi} - \frac{\sin\left(\theta\right)}{\pi K} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} - \frac{1-\frac{1}{K}}{\pi} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} - \frac{\sin\left(\theta\right)}{\pi K} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} \\
&\geq \frac{1}{K} + \frac{1-\frac{1}{K}}{\pi} - \left(\frac{2}{\pi K} + \frac{K-1}{K\pi}\right) \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} \\
&\geq \frac{1}{4K}, \qquad \text{(C.14)}
\end{aligned}$$

and we can also obtain the upper bound as

$$\begin{aligned}
&\max_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top \left(\boldsymbol{H}_1\right) \boldsymbol{u} + \boldsymbol{u}^\top \left(\boldsymbol{H}_2 + \boldsymbol{H}_3 + \boldsymbol{H}_4\right) \boldsymbol{u} \\
&\leq \frac{1}{K} + \frac{1-\frac{1}{K}}{\pi} - \frac{\sin\left(\theta\right)}{\pi K} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} - \frac{1-\frac{1}{K}}{\pi} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} + \frac{1}{\pi} \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} \\
&\leq \frac{K-1+\pi}{\pi K} + \frac{1}{\pi K} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2} \\
&\leq 3. \qquad \text{(C.15)}
\end{aligned}$$

As a summary, when $\theta \neq 0$ or π and $\boldsymbol{w} \neq 0$ we have shown that

$$\frac{1}{4K} \cdot \boldsymbol{I} \preceq \nabla^2 L\left(\boldsymbol{w}\right) \preceq 3 \cdot \boldsymbol{I} \qquad \text{(C.16)}$$

holds when $\|\boldsymbol{w}\|_2 \geq \frac{K+1}{K-1+\frac{3}{4}\pi}\|\boldsymbol{w}^\star\|_2$.

Next we will consider the case $\theta = 0$ and calculate the Hessian i.e., the derivative of $\nabla L\left(\boldsymbol{w}\right)$ at $\lambda \cdot \boldsymbol{w}^\star$ where $\lambda \in \mathbb{R}^+$. We decompose $\nabla L\left(\boldsymbol{w}\right)$ as two terms

$$\nabla L\left(\boldsymbol{w}\right) = g_1\left(\boldsymbol{w}\right) + g_2\left(\boldsymbol{w}\right), \qquad \text{(C.17)}$$

where

$$g_1\left(\boldsymbol{w}\right) = \frac{\|\boldsymbol{w}^\star\|_2}{K\pi} \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \cdot \sin\left(\theta\right) - \frac{\theta}{K\pi} \cdot \boldsymbol{w}^\star, \qquad \text{(C.18)}$$

$$g_2\left(\boldsymbol{w}\right) = \left(\frac{1}{K} + \frac{1-\frac{1}{K}}{\pi}\right) \cdot \boldsymbol{w} - \frac{1-\frac{1}{K}}{\pi} \cdot \|\boldsymbol{w}^\star\|_2 \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}. \qquad \text{(C.19)}$$

For $g_1(\boldsymbol{w})$, according to the definition [87, Definition 9.11], if there exist $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ such that

$$\lim_{\epsilon\to 0} \frac{\|g_1(\lambda\boldsymbol{w}^\star + \epsilon\boldsymbol{u}) - \boldsymbol{A}\cdot\epsilon\cdot\boldsymbol{u}\|_2}{\|\epsilon\boldsymbol{u}\|_2} = 0, \tag{C.20}$$

holds for all $\|\boldsymbol{u}\|_2 = 1$, then we say that $g_1(\boldsymbol{w})$ is differentiable at $\lambda\boldsymbol{w}^\star$ and we write $\nabla g_1(\lambda\boldsymbol{w}^\star) = \boldsymbol{A}$. Plugging $g_1(\boldsymbol{w})$ into the left-hand side of (C.20) we obtain

$$\begin{aligned}
&\lim_{\epsilon\to 0} \left\| \frac{\frac{\|\boldsymbol{w}^\star\|_2}{K\pi}\cdot\frac{\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u}}{\|\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u}\|_2}\cdot\sin(\theta_\epsilon) - \frac{\theta_\epsilon}{\pi K}\cdot\boldsymbol{w}^\star}{\epsilon} - \boldsymbol{A}\boldsymbol{u} \right\|_2 \\
&= \lim_{\epsilon\to 0} \left\| \frac{\|\boldsymbol{w}^\star\|_2}{K\pi}\cdot\frac{\lambda\boldsymbol{w}^\star}{\|\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u}\|_2}\cdot\frac{\sin(\theta_\epsilon)}{\epsilon} + \frac{\|\boldsymbol{w}^\star\|_2}{K\pi}\cdot\frac{\boldsymbol{u}}{\|\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u}\|_2}\cdot\sin(\theta_\epsilon) \right. \\
&\qquad \left. - \frac{\theta_\epsilon}{\epsilon}\cdot\frac{\boldsymbol{w}^\star}{\pi K} - \boldsymbol{A}\cdot\boldsymbol{u} \right\|_2
\end{aligned}$$

where θ_ϵ is the angle between $\lambda\boldsymbol{w}^\star + \epsilon\boldsymbol{u}$ and $\lambda\boldsymbol{w}^\star$, and we have

$$\begin{aligned}
\lim_{\epsilon\to 0}\frac{\sin(\theta_\epsilon)}{\epsilon} = \lim_{\epsilon\to 0}\frac{\sqrt{1-\left(\frac{\lambda\boldsymbol{w}^{\star\top}(\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u})}{\|\lambda\boldsymbol{w}^\star\|_2\|\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u}\|_2}\right)^2}}{\epsilon} &= \lim_{\epsilon\to 0}\sqrt{\frac{\|\lambda\boldsymbol{w}^\star\|_2^2 - \left(\lambda\boldsymbol{w}^{\star\top}\boldsymbol{u}\right)^2}{\|\lambda\boldsymbol{w}^\star\|_2^2\|\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u}\|_2^2}} \\
&= \sqrt{\frac{\|\lambda\boldsymbol{w}^\star\|_2^2 - \left(\lambda\boldsymbol{w}^{\star\top}\boldsymbol{u}\right)^2}{\|\lambda\boldsymbol{w}^\star\|_2^2\|\lambda\boldsymbol{w}^\star\|_2^2}}
\end{aligned} \tag{C.21}$$

exists. Further since,

$$\lim_{\epsilon\to 0}\frac{\theta_\epsilon}{\epsilon} = \lim_{\epsilon\to 0}\left(\frac{\theta_\epsilon}{\sin(\theta_\epsilon)}\cdot\frac{\sin(\theta_\epsilon)}{\epsilon}\right) \tag{C.22}$$

and

$$\lim_{\epsilon\to 0}\frac{\theta_\epsilon}{\sin(\theta_\epsilon)} = \lim_{\epsilon\to 0}\frac{1}{\cos(\theta_\epsilon)} = \lim_{\epsilon\to 0}\frac{\|\lambda\boldsymbol{w}^\star\|_2\|\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u}\|_2}{\lambda\boldsymbol{w}^{\star\top}(\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u})} = 1, \quad \lim_{\epsilon\to 0}\frac{\theta_\epsilon}{\epsilon} = \lim_{\epsilon\to 0}\frac{\sin(\theta_\epsilon)}{\epsilon}, \tag{C.23}$$

both exist, hence

$$\lim_{\epsilon\to 0}\frac{\|\boldsymbol{w}^\star\|_2}{K\pi}\cdot\frac{\lambda\boldsymbol{w}^\star}{\|\lambda\boldsymbol{w}^\star+\epsilon\boldsymbol{u}\|_2}\cdot\frac{\sin(\theta_\epsilon)}{\epsilon} - \frac{\theta_\epsilon}{\epsilon}\cdot\frac{\boldsymbol{w}^\star}{\pi K} = 0. \tag{C.24}$$

Together with

$$\lim_{\epsilon \to 0} \frac{\|\boldsymbol{w}^\star\|_2}{K\pi} \cdot \frac{\boldsymbol{u}}{\|\lambda \boldsymbol{w}^\star + \epsilon \boldsymbol{u}\|_2} \cdot \sin(\theta_\epsilon) = 0, \tag{C.25}$$

we will have (C.20) hold when $\boldsymbol{A} = \boldsymbol{0}$, i.e.,

$$\nabla g_1(\lambda \boldsymbol{w}^\star) = \boldsymbol{0}. \tag{C.26}$$

Next for $g_2(\boldsymbol{w}) = \left(\frac{1}{K} + \frac{1-\frac{1}{K}}{\pi}\right) \cdot \boldsymbol{w} - \frac{1-\frac{1}{K}}{\pi} \cdot \|\boldsymbol{w}^\star\|_2 \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the derivative can be calculated as

$$\nabla g_2(\boldsymbol{w}) = \left(\frac{1}{K} + \frac{1-\frac{1}{K}}{\pi} - \frac{1-\frac{1}{K}}{\pi} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2}\right) \cdot \boldsymbol{I} + \frac{1-\frac{1}{K}}{\pi} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2^3} \cdot \boldsymbol{w}\boldsymbol{w}^\top, \tag{C.27}$$

and since $\nabla L(\boldsymbol{w}) = g_1(\boldsymbol{w}) + g_2(\boldsymbol{w})$, we can conclude that when $\theta = 0$,

$$\begin{aligned} \nabla^2 L(\boldsymbol{w}) &= \left(\frac{1}{K} + \frac{1-\frac{1}{K}}{\pi} - \frac{1-\frac{1}{K}}{\pi} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2}\right) \cdot \boldsymbol{I} + \frac{1-\frac{1}{K}}{\pi} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2^3} \cdot \boldsymbol{w}\boldsymbol{w}^\top \\ &\succeq \left(\frac{1}{K} + \frac{1-\frac{1}{K}}{\pi} - \frac{1-\frac{1}{K}}{\pi} \cdot \frac{\|\boldsymbol{w}^\star\|_2}{\|\boldsymbol{w}\|_2}\right) \cdot \boldsymbol{I}, \end{aligned} \tag{C.28}$$

and

$$\nabla^2 L(\boldsymbol{w}) \preceq \left(\frac{1}{K} + \frac{1-\frac{1}{K}}{\pi}\right) \cdot \boldsymbol{I}. \tag{C.29}$$

Hence when $\theta = 0$, the result still holds.

□

C.3 Proof of GD on the Empirical Risk

C.3.1 Proof of Lemma 1

Proof. Recall that the empirical risk function (4.3) is

$$L_n(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{K} \sum_{j=1}^{K} \phi\left(\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)}\right) - y_i\right)^2, \tag{C.30}$$

and we can calculate its gradient as

$$\begin{aligned}\nabla L_n(\boldsymbol{w}) &= 2\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{K}\sum_{j=1}^{K}\phi\left(\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}\right)-y_i\right)\cdot\left(\frac{1}{K}\sum_{j=1}^{K}\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}\right)\boldsymbol{x}_i^{(j)}\right)\\ &= 2\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\left(\phi\left(\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}\right)-\phi\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right)\right)\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}\right)\boldsymbol{x}_i^{(l)},\end{aligned} \tag{C.31}$$

where ϕ' denotes the derivative of ϕ, and following the convention we define it as

$$\phi'(z)=1_{\{z>0\}}=\begin{cases}1 & z>0\\ 0 & z\le 0\end{cases}. \tag{C.32}$$

Further denote $\boldsymbol{h}=\boldsymbol{w}-\boldsymbol{w}^\star$, then the left-hand side of (4.15) can be written as

$$\begin{aligned}\left\langle\nabla L_n(\boldsymbol{w}),\boldsymbol{h}\right\rangle = 2\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\Bigg[&\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}\cdot\boldsymbol{h}^\top\boldsymbol{x}_i^{(l)}\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}>0\right\}}\\ &+\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\cdot\boldsymbol{h}^\top\boldsymbol{x}_i^{(l)}\cdot\left(1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}>0\right\}}-1_{\left\{\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}>0\right\}}\right)\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}>0\right\}}\Bigg].\end{aligned} \tag{C.33}$$

One pivotal observation of (C.33) is that when $\boldsymbol{w}$ is close to $\boldsymbol{w}^\star$, the first term can be lower bounded by the distance $\|\boldsymbol{h}\|_2$, and the second term is very small since $\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}$ and $\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}$ will have the same sign with high probability. We summarize the observation in the following two lemmas.

Lemma 19. *When the sample complexity satisfies $n\ge c\cdot mK^2\cdot\log(n)$ for some sufficiently large constant c, with probability at least $1-d^{-10}$, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}\cdot\boldsymbol{h}^\top\boldsymbol{x}_i^{(l)}\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}>0\right\}}\ge\frac{1}{4K}\cdot\|\boldsymbol{h}\|_2^2 \tag{C.34}$$

holds for all $\boldsymbol{w}\in\mathbb{B}\left(\boldsymbol{w}^\star,\frac{1}{K}\|\boldsymbol{w}^\star\|_2\right)$.

Lemma 20. *For some sufficiently large constant c_1, c_2, if $n \geq c_1 \cdot mK^2 \log(K)$ and $\|\boldsymbol{h}\|_2 \leq \frac{1}{K}\|\boldsymbol{w}^\star\|_2$ hold, then with probability at least $1 - c_2 \exp(-K^2 n)$,*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\cdot\boldsymbol{h}^\top\boldsymbol{x}_i^{(l)}\cdot\left(1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}>0\right\}}-1_{\left\{\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}>0\right\}}\right)\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}>0\right\}}$$
$$\leq \frac{1.2}{K^{\frac{3}{2}}}\cdot\|\boldsymbol{h}\|_2^2 \tag{C.35}$$

holds for all $\boldsymbol{w} \in \mathbb{B}\left(\boldsymbol{w}^\star, \frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2\right)$.

The proof of Lemma 19 and Lemma 20 is provided in Section C.3.2. Together with the above two lemmas we can conclude that as long as $n \geq c \cdot mK^2 \cdot \log(n)$ for some large constant c and $\|\boldsymbol{h}\|_2 \leq \frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2$,

$$\langle \nabla L_n(\boldsymbol{w}), \boldsymbol{h}\rangle \geq \frac{\sqrt{K}-5.8}{4K^{\frac{3}{2}}}\cdot\|\boldsymbol{h}\|_2^2 \tag{C.36}$$

holds with probability at least $1-\frac{1}{d^{10}}$ for all $\boldsymbol{w}$. For the right-hand side of RC (4.15) we successfully upper bound $\|\nabla L_n(\boldsymbol{w})\|_2^2$, and the result is summarized as follows.

Lemma 21. *If $n \geq c_1 m\epsilon^{-2}$, then with probability at least $1 - 2\exp(-c_2\epsilon^2 n) - 2\exp(-n/2)$ we have*

$$\|\nabla L_n(\boldsymbol{w})\|_2^2 \leq 18(1+\epsilon)\|\boldsymbol{h}\|_2^2 \tag{C.37}$$

holds for all $\boldsymbol{w}$.

Setting $\epsilon = \frac{1}{K}$ in lemma 21 we have the following inequality

$$\|\nabla L_n(\boldsymbol{w})\|_2^2 \leq 18\left(1+\frac{1}{K}\right)\|\boldsymbol{h}\|_2^2 \tag{C.38}$$

holds with probability at least $1 - c\cdot\exp(-mK^2)$ for some large constant c as long as $n \geq c_4 \cdot mK^2$. Hence, further take $\lambda = \frac{1}{72K^2}$ and $\mu = \frac{1}{4K^2}$, then we have

$$\langle \nabla L_n(\boldsymbol{w}), \boldsymbol{w}-\boldsymbol{w}^\star\rangle \geq \lambda\|\nabla L_n(\boldsymbol{w})\|_2^2 + \mu\|\boldsymbol{w}-\boldsymbol{w}^\star\|_2^2 \tag{C.39}$$

holds with probability at least $1 - \frac{1}{d^{10}}$ as long as the sample complexity satisfies $n \geq c \cdot mK^2 \cdot \log(n)$ for some large constant c and $\|\boldsymbol{h}\|_2 \leq \frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2$. □

C.3.2 Proof of Auxiliary Lemmas

Proof of Lemma 19

Proof. Lower bounding (C.34) can be boiled down to lower bound the smallest eigenvalue of the symmetric matrix $\boldsymbol{A}_n(\boldsymbol{w}) \in \mathbb{R}^{m\times m}$, where

$$\boldsymbol{A}_n(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{x}_i^{(j)}\cdot\boldsymbol{x}_i^{(l)\top}\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}>0\right\}}. \tag{C.40}$$

Due to the randomness of $\boldsymbol{A}_n(\boldsymbol{w})$, it's hard to lower bound $\boldsymbol{A}_n(\boldsymbol{w})$ directly, however it is easy to calculate and lower bound $\mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]$. Hence, we first lower bound $\mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]$, and then show that $\boldsymbol{A}_n(\boldsymbol{w})$ concentrates around its expectation.

With the two identities from Lemma 16, we can lower bound $\mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]$ as

$$\begin{aligned}
&\min_{\|\boldsymbol{u}\|_2=1}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{u}^\top\boldsymbol{x}_i^{(j)}\cdot\boldsymbol{x}_i^{(l)\top}\boldsymbol{u}\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}>0\right\}}\right]\cdot\boldsymbol{I}\\
&=\min_{\|\boldsymbol{u}\|_2=1}\frac{1}{K^2}\Bigg(\sum_{j\neq l}\mathbb{E}\left[\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0\right\}}\right]\cdot\mathbb{E}\left[\boldsymbol{u}^\top\boldsymbol{x}^{(l)}\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}}\right]\\
&\qquad+\sum_{j=1}^{K}\mathbb{E}\left[\left(\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right)^2\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0\right\}}\right]\Bigg)\cdot\boldsymbol{I}\\
&=\min_{\|\boldsymbol{u}\|_2=1}\frac{1}{K^2}\left(\left(K^2-K\right)\cdot\frac{1}{2\pi}\frac{\left(\boldsymbol{u}^\top\boldsymbol{w}\right)^2}{\|\boldsymbol{w}\|_2^2}+\frac{K}{2}\right)\cdot\boldsymbol{I}\\
&\succeq\frac{1}{2K}\cdot\boldsymbol{I}
\end{aligned}\tag{C.41}$$

where in the first equality we have used the fact that $\boldsymbol{x}^{(j)}$ is independent of $\boldsymbol{x}^{(l)}$ when $j\neq l$, and for the last inequality, the equality holds when $\langle\boldsymbol{u},\boldsymbol{w}\rangle = 0$. In a summary, we have shown that

$$\mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]\succeq\frac{1}{2K}\cdot\boldsymbol{I} \tag{C.42}$$

holds. Next we are going to apply a covering argument to show that with high probability the perturbation $\|\boldsymbol{A}_n(\boldsymbol{w}) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]\|$ is small for all $\boldsymbol{w} \in \mathbb{B}\left(\boldsymbol{w}^\star, \frac{1}{K}\|\boldsymbol{w}^\star\|_2\right)$. Notice that only the sign of $\boldsymbol{w}^\top \boldsymbol{x}$ matters for $\boldsymbol{A}_n(\boldsymbol{w})$, hence considering $\boldsymbol{w} \in \mathbb{S}^{m-1}$ will be enough. Denote $\mathcal{N}_\epsilon$ as the ϵ-net of the sphere $\mathbb{S}^{m-1}$, i.e. for any $\boldsymbol{w} \in \mathbb{S}^{m-1}$, there exist a $\boldsymbol{w}_\epsilon \in \mathcal{N}_\epsilon$ corresponding to $\boldsymbol{w}$ such that $\|\boldsymbol{w} - \boldsymbol{w}_\epsilon\|_2 \leq \epsilon$. By triangle inequality, we can write

$$\begin{aligned}\|\boldsymbol{A}_n(\boldsymbol{w}) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]\| \leq \|\boldsymbol{A}_n(\boldsymbol{w}) - \boldsymbol{A}_n(\boldsymbol{w}_\epsilon)\| + \|\boldsymbol{A}_n(\boldsymbol{w}_\epsilon) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)]\| \\ + \|\mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)] - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]\|, \end{aligned} \tag{C.43}$$

and hence

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}} \|\boldsymbol{A}_n(\boldsymbol{w}) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]\| \geq t\right) \\ &\leq \mathbb{P}\left(\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}} \|\boldsymbol{A}_n(\boldsymbol{w}) - \boldsymbol{A}_n(\boldsymbol{w}_\epsilon)\| \geq \frac{t}{3}\right) \\ &\quad + \mathbb{P}\left(\sup_{\boldsymbol{w}_\epsilon\in\mathcal{N}_\epsilon} \|\boldsymbol{A}_n(\boldsymbol{w}_\epsilon) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)]\| \geq \frac{t}{3}\right) \\ &\quad + \mathbb{P}\left(\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}} \|\mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)] - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})\|] \geq \frac{t}{3}\right). \end{aligned} \tag{C.44}$$

Next we will deal with the above terms step by step.

Firstly, let $V_{\frac{1}{4}}$ be a $\frac{1}{4}$-net of $\mathbb{B}(0,1)$ with $\log|V_{\frac{1}{4}}| \leq m \log 12$. We have

$$\|\boldsymbol{A}_n(\boldsymbol{w}_\epsilon) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)]\| \leq 2 \sup_{\boldsymbol{u}\in V_{\frac{1}{4}}} |\langle \boldsymbol{u}, (\boldsymbol{A}_n(\boldsymbol{w}_\epsilon) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)])\,\boldsymbol{u}\rangle|, \tag{C.45}$$

and then applying union bound over N_ϵ and $V_{\frac{1}{4}}$ will give us

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{w}_\epsilon \in N_\epsilon} \|\boldsymbol{A}_n(\boldsymbol{w}_\epsilon) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)]\| \geq \frac{t}{3}\right) \\
&\leq |N_\epsilon| \cdot |V_{\frac{1}{4}}| \cdot \sup_{\boldsymbol{w}_\epsilon \in N_\epsilon} \sup_{\boldsymbol{u} \in V_{\frac{1}{4}}} \mathbb{P}\left(|\langle \boldsymbol{u}, (\boldsymbol{A}_n(\boldsymbol{w}_\epsilon) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)])\boldsymbol{u}\rangle| \geq \frac{t}{6}\right) \\
&\leq \exp\left(m \log\left(1 + \frac{2}{\epsilon}\right) + m\log(12)\right) \\
&\quad \cdot \sup_{\boldsymbol{w}_\epsilon \in N_\epsilon} \sup_{\boldsymbol{u} \in V_{\frac{1}{4}}} \mathbb{P}\left(|\langle \boldsymbol{u}, (\boldsymbol{A}_n(\boldsymbol{w}_\epsilon) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)])\boldsymbol{u}\rangle| \geq \frac{t}{6}\right).
\end{aligned} \tag{C.46}$$

Further since the sub-exponential norm of $\boldsymbol{A}_n(\boldsymbol{w})$ can be upper bounded as

$$\begin{aligned}
&\left\| \frac{1}{K^2} \sum_{j=1}^{K} \sum_{l=1}^{K} \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \cdot \boldsymbol{x}^{(l)\top} \boldsymbol{u} \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}^{(j)} > 0, \boldsymbol{w}^\top \boldsymbol{x}^{(l)} > 0\right\}} \right\|_{\psi_1} \\
&\leq \frac{1}{K^2} \sum_{j=1}^{K} \sum_{l=1}^{K} \|\boldsymbol{u}^\top \boldsymbol{x}^{(j)} \cdot \boldsymbol{x}^{(l)\top} \boldsymbol{u}\|_{\psi_1} \\
&\leq C,
\end{aligned}$$

where C is some constant. Hence, applying the Bernstein inequality will give us

$$\mathbb{P}\left(|\langle \boldsymbol{u}, (\boldsymbol{A}_n(\boldsymbol{w}_\epsilon) - \mathbb{E}[\boldsymbol{A}(\boldsymbol{w}_\epsilon)])\boldsymbol{u}\rangle| \geq \frac{t}{6}\right) \leq 2\exp\left(-C \cdot n \cdot \min\left\{c_1 \cdot t^2, c_2 \cdot t\right\}\right), \tag{C.47}$$

together we will obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{w}_\epsilon \in N_\epsilon} \|\boldsymbol{A}_n(\boldsymbol{w}_\epsilon) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)]\| \geq \frac{t}{3}\right) \\
&\quad \leq \exp\left(m \log\left(12 + \frac{24}{\epsilon}\right)\right) \cdot 2\exp\left(-C \cdot n \cdot \min\left\{c_1 \cdot t^2, c_2 \cdot t\right\}\right).
\end{aligned}$$

Secondly, by the definition of spectral norm and apply Lemma 16 will give us

$$\begin{aligned}
\|\mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)] - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]\| &= \sup_{\|\boldsymbol{u}\|_2 = 1} \boldsymbol{u}^\top (\mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w}_\epsilon)] - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})])\boldsymbol{u} \\
&= \sup_{\|\boldsymbol{u}\|_2 = 1} \frac{K-1}{2\pi K} \left(\frac{\left(\boldsymbol{u}^\top \boldsymbol{w}_\epsilon\right)^2}{\|\boldsymbol{w}_\epsilon\|_2^2} - \frac{\left(\boldsymbol{u}^\top \boldsymbol{w}\right)^2}{\|\boldsymbol{w}\|_2^2}\right),
\end{aligned} \tag{C.48}$$

recall that $\|\boldsymbol{w}\|_2 = \|\boldsymbol{w}_\epsilon\|_2 = 1$ and plugging it back to the above equation,

$$\begin{aligned}\|\mathbb{E}\left[\boldsymbol{A}_n\left(\boldsymbol{w}_\epsilon\right)\right] - \mathbb{E}\left[\boldsymbol{A}_n\left(\boldsymbol{w}\right)\right]\| &= \sup_{\|\boldsymbol{u}\|_2=1} \frac{K-1}{2\pi K}\left(\boldsymbol{u}^\top\left(\boldsymbol{w}_\epsilon - \boldsymbol{w}\right)\cdot \boldsymbol{u}^\top\left(\boldsymbol{w}_\epsilon + \boldsymbol{w}\right)\right) \\ &< \frac{K-1}{2\pi K}\|\boldsymbol{w}_\epsilon - \boldsymbol{w}\|_2 \cdot \|\boldsymbol{w}_\epsilon + \boldsymbol{w}\|_2, \end{aligned} \tag{C.49}$$

where in the last step we apply Cauchy-Schwarz inequality. Hence by the definition of $\boldsymbol{w}_\epsilon$ we have the following holds

$$\|\mathbb{E}\left[\boldsymbol{A}_n\left(\boldsymbol{w}_\epsilon\right)\right] - \mathbb{E}\left[\boldsymbol{A}_n\left(\boldsymbol{w}\right)\right]\| < \frac{K-1}{\pi K}\cdot\epsilon. \tag{C.50}$$

Thus if we set $\frac{t}{3} \geq \frac{K-1}{\pi K}\cdot\epsilon$ we will have

$$\mathbb{P}\left(\sup_{\boldsymbol{w}\in N_\epsilon} \|\mathbb{E}\left[\boldsymbol{A}_n\left(\boldsymbol{w}_\epsilon\right)\right] - \mathbb{E}\left[\boldsymbol{A}_n\left(\boldsymbol{w}\right)\right]\| \geq \frac{t}{3}\right) = 0. \tag{C.51}$$

Finally, by Markov's inequality we will obtain

$$\mathbb{P}\left(\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}} \|\boldsymbol{A}_n\left(\boldsymbol{w}\right) - \boldsymbol{A}_n\left(\boldsymbol{w}_\epsilon\right)\| \geq \frac{t}{3}\right) \leq \frac{3}{t}\cdot\mathbb{E}\left[\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}} \|\boldsymbol{A}_n\left(\boldsymbol{w}\right) - \boldsymbol{A}_n\left(\boldsymbol{w}_\epsilon\right)\|\right], \tag{C.52}$$

and we can upper bound the right-hand side as following

$$\begin{aligned}&\mathbb{E}\left[\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}} \|\boldsymbol{A}_n\left(\boldsymbol{w}\right) - \boldsymbol{A}_n\left(\boldsymbol{w}_\epsilon\right)\|\right] \\ &= \mathbb{E}\Bigg[\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}} \|\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{x}^{(j)}\cdot\boldsymbol{x}^{(l)\top} \\ &\quad\cdot\left(1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}} - 1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}\right)\|\Bigg] \\ &\leq \mathbb{E}\Bigg[\|\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{x}^{(j)}\cdot\boldsymbol{x}^{(l)\top}\| \qquad\qquad (C.53) \\ &\quad\cdot\left(\max_{j,l}\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}} |1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}} - 1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}|\right)\Bigg] \\ &\leq \sqrt{\mathbb{E}\left[\|\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{x}^{(j)}\cdot\boldsymbol{x}^{(l)\top}\|^2\right]} \qquad\qquad (C.54) \\ &\quad\cdot\sqrt{\mathbb{E}\left[\left(\max_{j,l}\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}} |1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}} - 1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}|\right)^2\right]}, \qquad (C.55)\end{aligned}$$

the last inequality follows from Cauchy-Schwarz inequality. Notice that

$$\begin{aligned}
&\left(\max_{j,l}\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}}|1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}}-1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}|\right)^2\\
&=\max_{j,l}\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}}|1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}}-1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}|^2,\\
&=\max_{j,l}\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}}|1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}}-1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}|
\end{aligned}\tag{C.56}$$

since $|1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}}-1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}|$ is either 0 or 1. Applying triangle inequality we have

$$\begin{aligned}
&\max_{j,l}\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}}|1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}}-1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}|\\
&\le\max_{j,l}\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}}|1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0\right\}}-1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0\right\}}|+|1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}}-1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}|
\end{aligned}\tag{C.57}$$

hold. Furthermore, we have

$$|1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0\right\}}-1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0\right\}}|=\begin{cases}1 & \boldsymbol{w}^\top\boldsymbol{x}^{(j)}\cdot\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}\le 0\\ 0 & \text{otherwise}\end{cases}.\tag{C.58}$$

Next we are going to control $\mathbb{P}\left(\boldsymbol{w}^\top\boldsymbol{x}^{(j)}\cdot\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}\le 0\right)$. Denote $z_1=\bar{\boldsymbol{w}}^\top\boldsymbol{x}^{(j)}$ and $z_2=\bar{\boldsymbol{w}}_\epsilon^\top\boldsymbol{x}^{(j)}$, then z_1 and z_2 follow a joint Gaussian distribution and

$$\begin{aligned}
&\mathbb{P}\left(z_1z_2<0\right)\\
&=\mathbb{P}\left(z_1<0,z_2>0\right)+\mathbb{P}\left(z_1>0,z_2<0\right)\\
&=\frac{2}{2\pi\sqrt{1-\rho^2}}\int_0^{+\infty}\int_{-\infty}^0\exp\left(-\frac{1}{2(1-\rho^2)}\left(z_1^2-2\rho z_1z_2+z_2^2\right)\right)dz_1dz_2\\
&\le\frac{2}{2\pi\sqrt{1-\rho^2}}\int_0^{+\infty}\int_{-\infty}^0\exp\left(-\frac{1}{2(1-\rho^2)}\left(z_1^2+z_2^2\right)\right)dz_1dz_2\\
&=\frac{2}{2\pi\sqrt{1-\rho^2}}\int_0^{+\infty}\exp\left(-\frac{z_1^2}{2(1-\rho^2)}\right)dz_1\cdot\int_{-\infty}^0\exp\left(-\frac{z_2^2}{2(1-\rho^2)}\right)dz_2\\
&=\frac{2}{2\pi\sqrt{1-\rho^2}}\cdot\left(\frac{1}{2}\cdot\sqrt{2\pi\left(1-\rho^2\right)}\right)^2\\
&=\frac{1}{2}\sqrt{1-\rho^2},
\end{aligned}\tag{C.59}$$

where $\rho = \cos\theta_\epsilon$ denotes the correlation between z_1 and z_2, and the inequality follows since $\rho z_1 z_2 < 0$. Hence, by the definition of θ_ϵ and ρ we have

$$\mathbb{E}\left[\left(\max_{j,l}\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}}|\left(1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}^{(l)}>0\right\}}-1_{\left\{\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(j)}>0,\boldsymbol{w}_\epsilon^\top\boldsymbol{x}^{(l)}>0\right\}}\right)|\right)^2\right]\leq\sup_{\boldsymbol{w}\in\mathbb{S}^{m-1}}\frac{\sin\theta_\epsilon}{2}. \tag{C.60}$$

And for the first term in (C.55), denote $\boldsymbol{X}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_K]^\top\in\mathbb{R}^{K\times m}$, we have

$$\mathbb{E}\left[\|\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{x}^{(j)}\cdot\boldsymbol{x}^{(l)\top}\|^2\right]=\mathbb{E}\left[\|\frac{1}{K^2}\boldsymbol{X}\boldsymbol{X}^\top\|^2\right]=\frac{1}{K^4}\mathbb{E}\left[\mathrm{s}_{\max}\left(\boldsymbol{X}\right)^4\right] \tag{C.61}$$

from [88, Corollary 5.35] we know that

$$\mathrm{s}_{\max}\left(\boldsymbol{X}\right)\geq\sqrt{K}+\sqrt{m}+t \tag{C.62}$$

holds with probability less or equal than $2\exp\left(-\frac{t^2}{2}\right)$, in other words,

$$\mathbb{P}\left(\mathrm{s}_{\max}\left(\boldsymbol{X}\right)^4\geq t\right)\leq 2\exp\left(-\frac{\left(t^{\frac{1}{4}}-\sqrt{K}-\sqrt{m}\right)^2}{2}\right), \tag{C.63}$$

and then applying the following fact, $\mathbb{E}\left[Z\right]=\int_0^\infty\mathbb{P}\left(Z\geq t\right)dt$ holds for a positive random variable, we will obtain

$$\mathbb{E}\left[\mathrm{s}_{\max}\left(\boldsymbol{X}\right)^4\right]=\int_0^\infty\mathbb{P}\left(\mathrm{s}_{\max}\left(\boldsymbol{X}\right)^4\geq t\right)dt\leq\int_0^\infty 2\exp\left(-\frac{\left(t^{\frac{1}{4}}-\sqrt{K}-\sqrt{m}\right)^2}{2}\right)dt, \tag{C.64}$$

and by changing variable we can write

$$\begin{aligned}\mathbb{E}\left[\mathrm{s}_{\max}\left(\boldsymbol{X}\right)^4\right]&\leq\int_0^\infty 8\left(t+\sqrt{K}+\sqrt{m}\right)^3\exp\left(-\frac{t^2}{2}\right)dt\\&=\int_0^\infty 8\left(t^3+\left(\sqrt{K}+\sqrt{m}\right)^3+3t\left(\sqrt{K}+\sqrt{m}\right)^2+3t^2\left(\sqrt{K}+\sqrt{m}\right)\right)\\&\qquad\cdot\exp\left(-\frac{t^2}{2}\right)dt\\&\leq C\cdot\left(\sqrt{K}+\sqrt{m}\right)^3.\end{aligned} \tag{C.65}$$

Combining results, we can show that

$$\mathbb{E}\left[\sup_{\boldsymbol{w}} \|\boldsymbol{A}_n(\boldsymbol{w}) - \boldsymbol{A}_n(\boldsymbol{w}_\epsilon)\|\right] \leq \frac{\left(\sqrt{K}+\sqrt{m}\right)^{\frac{3}{2}}}{K^2} \cdot \sqrt{\sup_{\boldsymbol{w}} \frac{\sin\theta_\epsilon}{2}}, \tag{C.66}$$

where from Lemma 18 we know that $(\sin\theta_\epsilon)^2 < (\epsilon)^2$ holds when ϵ is small, hence

$$\mathbb{P}\left(\sup_{\boldsymbol{w}} \|\boldsymbol{A}_n(\boldsymbol{w}) - \boldsymbol{A}_n(\boldsymbol{w}_\epsilon)\| \geq \frac{t}{3}\right) \leq \frac{C}{t} \cdot \frac{\left(\sqrt{K}+\sqrt{m}\right)^{\frac{3}{2}}}{K^2} \cdot \sqrt{\epsilon} \tag{C.67}$$

Thus if we let

$$\exp\left(m\log\left(12+\frac{24}{\epsilon}\right)\right) \cdot 2\exp\left(-C\cdot n\cdot \min\left\{c_1\cdot t^2, c_2\cdot t\right\}\right) \leq \frac{\delta}{2}, \tag{C.68}$$

$$\frac{C}{t} \cdot \frac{\left(\sqrt{K}+\sqrt{m}\right)^{\frac{3}{2}}}{K^2} \cdot \sqrt{\epsilon} \leq \frac{\delta}{2}, \tag{C.69}$$

and then when

$$t \geq \max\left\{C\cdot\frac{K-1}{K}\cdot\epsilon, \frac{C\cdot\sqrt{\epsilon}\cdot\left(\sqrt{K}+\sqrt{m}\right)^{\frac{3}{2}}}{\delta K^2}, \sqrt{\frac{m\log(\frac{1}{\epsilon})-\log(\delta)}{n}}\right\}, \tag{C.70}$$

set $\epsilon = \frac{\delta}{n}$, $\delta = \frac{1}{d^{-10}}$, we will obtain that with probability at least $1-\frac{1}{d^{10}}$,

$$\|\boldsymbol{A}_n(\boldsymbol{w}) - \mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})]\| \leq \sqrt{\frac{C\cdot m\cdot\log(n)}{n}} \tag{C.71}$$

holds for all $\boldsymbol{w}\in\mathbb{S}^{m-1}$ and remember that only the direction of $\boldsymbol{w}$ matters, hence the claim also holds for all $\boldsymbol{w}\in\mathbb{B}\left(\boldsymbol{w}^\star, \frac{1}{K}\|\boldsymbol{w}^\star\|_2\right)$.

Together with the result that

$$\mathbb{E}[\boldsymbol{A}_n(\boldsymbol{w})] \succeq \frac{1}{2K}\cdot\boldsymbol{I}, \tag{C.72}$$

we can conclude that when the sample size $n \geq mK^2\cdot\log(n)$ with probability at least $1-\frac{1}{d^{10}}$,

$$\boldsymbol{A}_n(\boldsymbol{w}) \succeq \frac{1}{4K}\cdot\boldsymbol{I} \tag{C.73}$$

holds for all $\boldsymbol{w} \in \mathbb{B}\left(\boldsymbol{w}^\star, \frac{1}{K}\|\boldsymbol{w}^\star\|_2\right)$.

□

Proof of Lemma 20

Proof. Next we will show that the second term in (C.33) is uniformly upper bounded for all $\boldsymbol{w}$ in a local neighborhood of $\boldsymbol{w}^\star$. With some algebra we can have the following deterministic upper bound,

$$\begin{aligned}
&\left|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)} \cdot \boldsymbol{h}^\top \boldsymbol{x}_i^{(l)} \cdot \left(1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)} > 0\right\}} - 1_{\left\{\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)} > 0\right\}}\right) \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(l)} > 0\right\}}\right| \\
&\quad \le \left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(l)}\right| \cdot \left|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right| \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)} \cdot \boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)} < 0\right\}}, \qquad \text{(C.74)}
\end{aligned}$$

further since $1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)} \cdot \boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)} < 0\right\}} = 1_{\left\{(\boldsymbol{w}^\star + \boldsymbol{h})^\top \boldsymbol{x}_i^{(j)} \cdot \boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)} < 0\right\}} \le 1_{\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}| \le |\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|\right\}}$ holds, we will obtain the following bound,

$$\begin{aligned}
\left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(l)}\right| \cdot \left|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right| \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)} \cdot \boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)} < 0\right\}} &\le \left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(l)}\right| \cdot \left|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right| \cdot 1_{\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}| \le |\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|\right\}} \\
&\le \left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(l)}\right| \cdot \left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}\right| \cdot 1_{\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}| \le |\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|\right\}},
\end{aligned}$$

i.e.,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \frac{1}{K^2}\sum_{j=1}^K\sum_{l=1}^K \boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)} \cdot \boldsymbol{h}^\top \boldsymbol{x}_i^{(l)} \cdot \left(1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)} > 0\right\}} - 1_{\left\{\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)} > 0\right\}}\right) \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(l)} > 0\right\}} \\
&\le \frac{1}{n}\sum_{i=1}^n \frac{1}{K^2}\sum_{j=1}^K\sum_{l=1}^K \left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(l)}\right| \cdot \left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}\right| \cdot 1_{\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}| \le |\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|\right\}} \\
&= \frac{1}{K^2}\sum_{j\neq l} \frac{1}{n}\sum_{i=1}^n \left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(l)}\right| \cdot \left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}\right| \cdot 1_{\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}| \le |\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|\right\}} \\
&\quad + \frac{1}{K^2}\sum_{j=1}^K \frac{1}{n}\sum_{i=1}^n \left|\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}\right|^2 \cdot 1_{\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}| \le |\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|\right\}}. \qquad \text{(C.75)}
\end{aligned}$$

The key observation here is that $\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}| \le |\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|\right\}$ is an event with small probability when $\boldsymbol{w}$ is close to $\boldsymbol{w}^\star$, we first show the second term in (C.75) can be upper bounded by $\|\boldsymbol{h}\|_2^2$ up to some scaling factor and the result is summarized as follows.

Lemma 22. *For any $\epsilon > 0$ and some large enough constants c_1, c_2, if $n \geq c_1 \cdot m\epsilon^{-2}\log(\epsilon^{-1})$, then with probability at least $1 - c_2\exp(-\epsilon^2 n)$,*

$$\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|^2 \cdot 1_{\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}|\leq|\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}|\right\}} \leq \left(\frac{1.18}{K^{\frac{3}{2}}}+\epsilon\right)\cdot\|\boldsymbol{h}\|_2^2, \tag{C.76}$$

holds for all non-zeros $\boldsymbol{h}\in\mathbb{R}^m$ satisfying $\|\boldsymbol{h}\|_2 \leq \frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2$, by setting $\epsilon = \frac{0.02}{K^{\frac{3}{2}}}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|^2 \cdot 1_{\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}|\leq|\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}|\right\}} \leq \frac{1.2}{K^{\frac{3}{2}}}\cdot\|\boldsymbol{h}\|_2^2 \tag{C.77}$$

holds with probability at least $1-c_2\exp(-K^3 n)$, as long as $n \geq c_1\cdot mK^3\log\left(K^{\frac{3}{2}}\right)$ and $\|\boldsymbol{h}\|_2 \leq \frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2$ hold.

Moreover, with slight adaption we can directly show a similar upper bound holds for the fist term in (C.75). Hence applying Lemma 22 we have that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\cdot\boldsymbol{h}^\top\boldsymbol{x}_i^{(l)}\cdot\left(1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}>0\right\}}-1_{\left\{\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}>0\right\}}\right)\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}>0\right\}}\\
&\leq \frac{K^2-K}{K^2}\cdot\frac{1.2}{K^{\frac{3}{2}}}\|\boldsymbol{h}\|_2^2+\frac{K}{K^2}\cdot\frac{1.2}{K^{\frac{3}{2}}}\|\boldsymbol{h}\|_2^2 \leq \frac{1.2}{K^{\frac{3}{2}}}\cdot\|\boldsymbol{h}\|_2^2,
\end{aligned} \tag{C.78}$$

holds with probability at least $1-c_2\exp(-K^3 n)$ as long as $n\geq c_1\cdot mK^3\log\left(K^{\frac{3}{2}}\right)$ and $\|\boldsymbol{h}\|_2\leq\frac{1}{K^{\frac{3}{2}}}\|\boldsymbol{w}^\star\|_2$ hold. □

Proof of Lemma 21

Proof. Recall that the gradient is

$$\begin{aligned}
\nabla L_n(\boldsymbol{w}) &= 2\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{1}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\left(\phi\left(\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}\right)-\phi\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right)\right)\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}\right)\boldsymbol{x}_i^{(l)}\\
&= \frac{2}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}\cdot 1_{\left\{\boldsymbol{w}^\top\boldsymbol{x}_i^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}>0\right\}}\right.\\
&\qquad\left.-\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\cdot 1_{\left\{\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}>0,\boldsymbol{w}^\top\boldsymbol{x}_i^{(l)}>0\right\}}\right)\boldsymbol{x}_i^{(l)}
\end{aligned} \tag{C.79}$$

Let $\boldsymbol{X}^{(l)} = \left[\boldsymbol{x}_1^{(l)}, \cdots, \boldsymbol{x}_n^{(l)}\right] \in \mathbb{R}^{m\times n}$ and $\boldsymbol{a}^{(j,l)} = \left[a_1^{(j,l)}, \cdots, a_n^{(j,l)}\right]^\top \in \mathbb{R}^n$ where $a_i^{(j,l)} = \boldsymbol{w}^\top \boldsymbol{x}_i^{(j)} \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)}>0, \boldsymbol{w}^\top \boldsymbol{x}_i^{(l)}>0\right\}} - \boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)} \cdot 1_{\left\{\boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)}>0, \boldsymbol{w}^\top \boldsymbol{x}_i^{(l)}>0\right\}}$, then we can rewrite the gradient as

$$\nabla L_n(\boldsymbol{w}) = \frac{2}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\frac{1}{n}\boldsymbol{X}^{(l)}\boldsymbol{a}^{(j,l)} \tag{C.80}$$

since the spectrum of $\boldsymbol{X}^{(l)}$ is well controlled when n is large enough, hence in order to upper bound the norm of $\nabla L_n(\boldsymbol{w})$, we will analyze the upper bound of $\|\boldsymbol{a}^{(j,l)}\|_2^2$. We can further rewrite $a_i^{(j,l)}$ as

$$\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)} \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)}>0, \boldsymbol{w}^\top \boldsymbol{x}_i^{(l)}>0\right\}} + \boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)} \cdot \left(1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)}>0\right\}} - 1_{\left\{\boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)}>0\right\}}\right) \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(l)}>0\right\}}$$

by triangle inequality we have that

$$\begin{aligned}
&\|\boldsymbol{a}^{(j,l)}\|_2 \\
&\le \sqrt{\sum_{i=1}^{n}\left(\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)} \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)}>0, \boldsymbol{w}^\top \boldsymbol{x}_i^{(l)}>0\right\}}\right)^2} \\
&\quad + \sqrt{\sum_{i=1}^{n}\left(\boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)} \cdot \left(1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)}>0\right\}} - 1_{\left\{\boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)}>0\right\}}\right) \cdot 1_{\left\{\boldsymbol{w}^\top \boldsymbol{x}_i^{(l)}>0\right\}}\right)^2} \\
&\le \sqrt{\sum_{i=1}^{n}\left(\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}\right)^2} + \sqrt{\sum_{i=1}^{n}\left(\boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)}\right)^2 \cdot 1_{\left\{\left(\boldsymbol{w}^\top \boldsymbol{x}_i^{(j)}\right)\cdot\left(\boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)}\right)<0\right\}}} \\
&\le \sqrt{\sum_{i=1}^{n}\left(\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}\right)^2} + \sqrt{\sum_{i=1}^{n}\left(\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}\right)^2 \cdot 1_{\left\{|\boldsymbol{w}^{\star\top} \boldsymbol{x}_i^{(j)}|<|\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|\right\}}} \\
&\le 2\sqrt{\sum_{i=1}^{n}\left(\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}\right)^2}
\end{aligned} \tag{C.81}$$

Applying Lemma 17 to the above term will give us

$$\mathbb{P}\left(\sqrt{\sum_{i=1}^{n}\left(\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}\right)^2} \le \sqrt{n(1+\epsilon)}\|\boldsymbol{h}\|_2\right) \ge 1 - 2\exp\left(-c\epsilon^2 n\right), \tag{C.82}$$

as long as $n \geq cm\epsilon^{-2}$. Finally we can upper bound the norm of the gradient as

$$\|\nabla L_n(\boldsymbol{w})\|_2^2 \leq \frac{2}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\|\frac{1}{n}\boldsymbol{X}^{(l)}\boldsymbol{a}^{(j,l)}\|_2^2 \leq \frac{2}{K^2}\sum_{j=1}^{K}\sum_{l=1}^{K}\frac{1}{n^2}\|\boldsymbol{X}^{(l)}\|^2\cdot\|\boldsymbol{a}^{(j,l)}\|_2^2,$$

from [88, Corollary 5.35] we know that $\mathbb{P}\left(\|\boldsymbol{X}^{(l)}\|^2 \leq 9n\right) \geq 1-2\exp(-n/2)$ as long as $n \geq m$, hence by union bound we can conclude that

$$\mathbb{P}\left(\|\nabla L_n(\boldsymbol{w})\|_2^2 \leq 18(1+\epsilon)\|\boldsymbol{h}\|_2^2\right) \geq 1-2\exp\left(-c\epsilon^2 n\right)-2\exp(-n/2) \tag{C.83}$$

holds as long as $n \geq cm\epsilon^{-2}$. □

Proof of Lemma 22

Proof. We will follow the same idea in the proof of [74, Lemma 7]. Firstly, for a fixed $\boldsymbol{h}$ we will apply Bernstein type concentration inequality to show the result holds, and then apply a covering argument to generalize the result. We define the following Lipschitz function,

$$F_i(t) = \begin{cases} t, & t > \left(\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right)^2 \\ \frac{1}{\delta}\left(t-\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right)^2\right)+\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right)^2 & (1-\delta)\cdot\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right)^2 \leq t \leq \left(\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}\right)^2 \\ 0, & \text{otherwise} \end{cases}$$

for $i=1,\cdots,n$, and it's easy to verify that

$$|\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}|^2\cdot 1_{\left\{|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}|\leq|\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}|\right\}} \leq F_i\left(|\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}|^2\right) \leq |\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}|^2\cdot 1_{\left\{(1-\delta)|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}|^2\leq|\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}|^2\right\}}.$$

Next we are going to upper bound the expectation of the right-hand side term above, for simplicity we let $\gamma_i = \frac{|\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}|^2}{\|\boldsymbol{h}\|_2^2}\cdot 1_{\left\{(1-\delta)|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}|^2\leq|\boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}|^2\right\}}$ and $r=\frac{\|\boldsymbol{h}\|_2}{\|\boldsymbol{w}^\star\|_2}$, then

$$\mathbb{E}[\gamma_i] = \int\int_{-\infty}^{\infty}\mathbb{E}\left[\gamma_i|\boldsymbol{w}^{\star\top}\boldsymbol{x}_i^{(j)}=\tau_1\|\boldsymbol{w}^\star\|_2, \boldsymbol{h}^\top\boldsymbol{x}_i^{(j)}=\tau_2\|\boldsymbol{h}\|_2\right]\cdot f(\tau_1,\tau_2)\,d\tau_1 d\tau_2, \tag{C.84}$$

where $f(\tau_1, \tau_2)$ is the joint density of two Gaussian random variable with correlation $\rho = \frac{\boldsymbol{h}^\top \boldsymbol{w}^\star}{\|\boldsymbol{h}\|_2 \|\boldsymbol{w}^\star\|_2}$. Continuing evaluate (C.84) we will obtain

$$
\begin{aligned}
&\mathbb{E}[\gamma_i] \\
&= \int \int_{-\infty}^{\infty} \tau_2^2 \cdot 1_{\left\{\sqrt{1-\delta}|\tau_1| \leq |\tau_2| \cdot r\right\}} \cdot f(\tau_1, \tau_2)\, d\tau_1 d\tau_2 \\
&= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \left(\operatorname{erf}\left(\frac{\left(\frac{r}{\sqrt{1-\delta}} - \rho\right)\tau_2}{\sqrt{1-\rho^2}} \right) + \operatorname{erf}\left(\frac{\left(\frac{r}{\sqrt{1-\delta}} + \rho\right)\tau_2}{\sqrt{1-\rho^2}} \right) \right) d\tau_2,
\end{aligned}
$$

we omit calculations for the second equality here, further since

$$
\begin{aligned}
&\operatorname{erf}\left(\frac{\left(\frac{r}{\sqrt{1-\delta}} - \rho\right)\tau_2}{\sqrt{1-\rho^2}} \right) + \operatorname{erf}\left(\frac{\left(\frac{r}{\sqrt{1-\delta}} + \rho\right)\tau_2}{\sqrt{1-\rho^2}} \right) \\
&= \operatorname{erf}\left(\frac{\left(\frac{r}{\sqrt{1-\delta}} + \rho\right)\tau_2}{\sqrt{1-\rho^2}} \right) - \operatorname{erf}\left(\frac{\left(\rho - \frac{r}{\sqrt{1-\delta}}\right)\tau_2}{\sqrt{1-\rho^2}} \right) \\
&= \left(\frac{2\frac{r}{\sqrt{1-\delta}} \cdot \tau_2}{\sqrt{1-\rho^2}} \right) \cdot \frac{2}{\sqrt{\pi}} \exp\left(-z^2\right), \qquad \text{(C.85)}
\end{aligned}
$$

where the first equality holds since $\operatorname{erf}(\cdot)$ is an odd function, the second equality holds by applying the mean value theorem, and $z = \lambda \frac{\left(\frac{r}{\sqrt{1-\delta}} + \rho\right)\tau_2}{\sqrt{1-\rho^2}} + (1-\lambda) \frac{\left(\rho - \frac{r}{\sqrt{1-\delta}}\right)\tau_2}{\sqrt{1-\rho^2}}$ for some $\lambda \in (0,1)$.

From Lemma 18 we know that

$$
\rho > \rho^2 \geq 1 - r^2 \qquad \text{(C.86)}
$$

and we have that $r = \frac{\|\boldsymbol{h}\|_2}{\|\boldsymbol{w}^\star\|_2} \leq \frac{1}{K^{\frac{3}{2}}}$. When the number of neurons $K \geq 2$ and δ is small, e.g. $\delta = 0.01$, then $\frac{r}{\sqrt{1-\delta}} \approx r$. Hence, $0 < \rho - \frac{r}{\sqrt{1-\delta}} < \rho + \frac{r}{\sqrt{1-\delta}}$, notice that $\exp(-z^2)$ is monotonic decreasing with respect to z, thus we have

$$
\begin{aligned}
\left(\frac{2\frac{r}{\sqrt{1-\delta}} \cdot \tau_2}{\sqrt{1-\rho^2}} \right) \cdot \frac{2}{\sqrt{\pi}} \exp\left(-z^2\right) &\leq 2.27 \cdot r \cdot \frac{\tau_2}{\sqrt{1-\rho^2}} \cdot \exp\left(- \left(\frac{\left(\rho - \frac{r}{\sqrt{1-\delta}}\right)\tau_2}{\sqrt{1-\rho^2}} \right)^2 \right) \\
&\leq 2.27 \cdot r \cdot \frac{\tau_2}{\sqrt{1-\rho^2}} \cdot \exp\left(-\frac{0.12\tau_2^2}{1-\rho^2} \right), \qquad \text{(C.87)}
\end{aligned}
$$

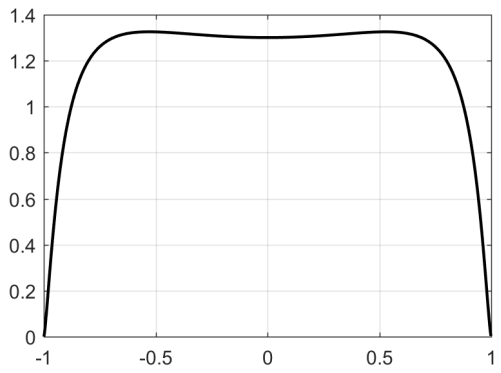


Figure C.1: Numerical integral with respect to ρ

holds, the last step follows from the fact that $\rho - \frac{r}{\sqrt{1-\delta}} \geq \sqrt{1-r^2} - \frac{r}{\sqrt{1-\delta}} \geq 0.36$, and notice that the right-hand side is monotonic decreasing, and $r \leq \frac{1}{2}$. Together we can upper bound the expectation as

$$
\begin{aligned}
&\mathbb{E}\left[\gamma_i\right] \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \left(\operatorname{erf}\left(\frac{\left(\frac{r}{\sqrt{1-\delta}} - \rho\right)\tau_2}{\sqrt{1-\rho^2}} \right) + \operatorname{erf}\left(\frac{\left(\frac{r}{\sqrt{1-\delta}} + \rho\right)\tau_2}{\sqrt{1-\rho^2}} \right) \right) d\tau_2 \\
&\leq \frac{2.27}{\sqrt{2\pi}} \cdot r \cdot \int_0^\infty \tau_2^3 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{0.12\tau_2^2}{1-\rho^2}\right) d\tau_2, \\
&\leq \frac{1.18}{K^{\frac{3}{2}}},
\end{aligned}
$$

where we evaluate the integral in the first inequality over different choices of ρ in Fig C.1, we see that the integral is upper bounded by a constant, hence we can obtain the last inequality. Moreover, $\|F_i\left(|\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|^2\right)\|_{\psi_1} \leq C \cdot \|\boldsymbol{h}\|_2^2$ for all $i = 1, \cdots, n$, hence when $r = \frac{\|\boldsymbol{h}\|_2}{\|\boldsymbol{w}^\star\|_2} \leq \frac{1}{K}$ we can have

$$
\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \frac{F_i\left(|\boldsymbol{h}^\top \boldsymbol{x}_i^{(j)}|^2\right)}{\|\boldsymbol{h}\|_2^2} \geq \left(\frac{1.18}{K^{\frac{3}{2}}} + \epsilon \right) \right) < \exp\left(-c \cdot n \cdot \epsilon^2\right), \tag{C.88}
$$

holds for some constant c. Since the covering argument will be the same as the proof of [74, Lemma 7], we omit the repetition and conclude.

□

Bibliography

[1] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on.* IEEE, 2008, pp. 16–21.

[2] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[3] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[4] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[5] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[6] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Advances in Neural Information Processing Systems*, 2015, pp. 739–747.

[7] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon, "Recovery guarantees for one-hidden-layer neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 4140–4149.

[8] S. Lloyd, "Least squares quantization in PCM," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[9] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289 –1306, April 2006.

[10] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[11] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *Information Theory, IEEE Transactions on*, vol. 56, no. 1, pp. 520–544, 2010.

[12] M. Mishali, Y. C. Eldar, and A. J. Elron, "Xampling: Signal acquisition and processing in union of subspaces," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4719–4734, 2011.

[13] M. Wakin, S. Becker, E. Nakamura, M. Grant, E. Sovero, D. Ching, J. Yoo, J. Romberg, A. Emami-Neyestanak, and E. Candes, "A nonuniform sampler for wideband spectrally-sparse environments," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 2, no. 3, pp. 516–529, 2012.

[14] F. Zeng, C. Li, and Z. Tian, "Distributed compressive spectrum sensing in cooperative multihop cognitive networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 1, pp. 37–48, 2011.

[15] Y. L. Polo, Y. Wang, A. Pandharipande, and G. Leus, "Compressive wide-band spectrum sensing," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.* IEEE, 2009, pp. 2337–2340.

[16] O. Mehanna and N. Sidiropoulos, "Frugal sensing: Wideband power spectrum sensing from few bits," *Signal Processing, IEEE Transactions on*, vol. 61, no. 10, pp. 2693–2703, May 2013.

[17] Y. Chi and H. Fu, "Subspace learning from bits," *IEEE Transactions on Signal Processing*, vol. 65, no. 17, pp. 4429–4442, Sept 2017.

[18] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on selected areas in communications*, vol. 17, no. 4, pp. 539–550, 1999.

[19] J. N. Laska, P. T. Boufounos, M. A. Davenport, and R. G. Baraniuk, "Democracy in action: Quantization, saturation, and compressive sensing," *Applied and Computational Harmonic Analysis*, vol. 31, no. 3, pp. 429–443, 2011.

[20] M. Slawski and P. Li, "Linear signal recovery from b-bit-quantized linear measurements: precise analysis of the trade-off between bit depth and number of measurements," *arXiv preprint arXiv:1607.02649*, 2016.

[21] H. L. Van Trees, *Detection, estimation, and modulation theory.* John Wiley & Sons, 2004.

[22] Y. Chi, L. Scharf, A. Pezeshki, and A. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2182–2195, May 2011.

[23] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE transactions on information theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

[24] Y. Chi and Y. Chen, "Compressive two-dimensional harmonic retrieval via atomic norm minimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 1030–1042, 2015.

[25] Y. Li and Y. Chi, "Off-the-grid line spectrum denoising and estimation with multiple measurement vectors," *Signal Processing, IEEE Transactions on*, vol. 64, no. 5, pp. 1257–1269.

[26] R. Heckel and M. Soltanolkotabi, "Generalized line spectral estimation via convex optimization," *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4001–4023, 2017.

[27] Z. Yang and L. Xie, "Exact joint sparse frequency recovery via optimization methods," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5145–5157, 2016.

[28] G. Tang, B. N. Bhaskar, and B. Recht, "Near minimax line spectral estimation," *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 499–512, 2015.

[29] B. N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5987–5999, 2013.

[30] Y. Chi and M. F. Da Costa, "Harnessing sparsity over the continuum: Atomic norm minimization for super resolution," *arXiv preprint arXiv:1904.04283*, 2019.

[31] G. Gray and G. Zeoli, "Quantization and saturation noise due to analog-to-digital conversion," *IEEE Transactions on Aerospace and Electronic Systems*, no. 1, pp. 222–223, 1971.

[32] A. Gupta, R. Nowak, and B. Recht, "Sample complexity for 1-bit compressed sensing and sparse classification," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1553–1557.

[33] M. Yan, Y. Yang, and S. Osher, "Robust 1-bit compressive sensing using adaptive outlier pursuit," *Signal Processing, IEEE Transactions on*, vol. 60, no. 7, pp. 3868–3875, 2012.

[34] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1275–1297, 2013.

[35] ——, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 482–494, 2013.

[36] ——, "The generalized Lasso with non-linear observations," *IEEE Transactions on Information Theory*, vol. 62, no. 3, pp. 1528–1537, 2016.

[37] Y. Plan, R. Vershynin, and E. Yudovina, "High-dimensional estimation with geometric constraints," *Information and Inference: A Journal of the IMA*, vol. 6, no. 1, pp. 1–40, 2016.

[38] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *Information Theory, IEEE Transactions on*, vol. 59, no. 4, pp. 2082–2102, 2013.

[39] S. Oymak and M. Soltanolkotabi, "Fast and reliable parameter estimation from nonlinear observations," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2276–2300, 2017.

[40] P. T. Boufounos, L. Jacques, F. Krahmer, and R. Saab, "Quantization and compressive sensing," in *Compressed Sensing and its Applications*. Springer, 2015, pp. 193–237.

[41] J. N. Laska and R. G. Baraniuk, "Regime change: Bit-depth versus measurement-rate in compressive sensing," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3496–3505, 2012.

[42] A. Host-Madsen and P. Handel, "Effects of sampling and quantization on single-tone frequency estimation," *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 650–662, 2000.

[43] P. Pakrooh, L. L. Scharf, A. Pezeshki, and Y. Chi, "Analysis of fisher information and the Cramér-Rao bound for nonlinear parameter estimation after compressed sensing," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6630–6634.

[44] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[45] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[48] M. Soltanolkotabi, "Learning ReLUs via gradient descent," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 2007–2017.

[49] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with ReLU activation," in *Advances in Neural Information Processing Systems*, 2017, pp. 597–607.

[50] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for nonconvex losses," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2747–2774, 2018.

[51] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.

[52] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 14–31, 2018.

[53] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, April 2015.

[54] R. Ge and T. Ma, "On the optimization landscape of tensor decompositions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 3653–3663.

[55] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[56] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery using nonconvex optimization," *International Conference on Machine Learning*, pp. 2351–2360, 2015.

[57] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[58] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 3345–3354.

[59] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, 2019.

[60] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Transactions on Information Theory*, 2018.

[61] D. Boob and G. Lan, "Theoretical properties of the global optimizer of two layer neural network," *arXiv preprint arXiv:1710.11241*, 2017.

[62] I. Safran and O. Shamir, "On the quality of the initial basin in overspecified neural networks," in *International Conference on Machine Learning*, 2016, pp. 774–782.

[63] Q. Nguyen and M. Hein, "The loss surface of deep and wide neural networks," in *International Conference on Machine Learning*, 2017, pp. 2603–2612.

[64] Y. Tian, "An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 3404–3413.

[65] R. Ge, J. D. Lee, and T. Ma, "Learning one-hidden-layer neural networks with landscape design," in *International Conference on Learning Representations*, 2018.

[66] I. Safran and O. Shamir, "Spurious local minima are common in two-layer relu neural networks," *arXiv preprint arXiv:1712.08968*, 2017.

[67] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval," *Mathematical Programming*, vol. 176, no. 1-2, pp. 5–37, 2019.

[68] S. Oymak, "Learning compact neural networks with regularization," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 3966–3975.

[69] X. Zhang, Y. Yu, L. Wang, and Q. Gu, "Learning One-hidden-layer ReLU networks via Gradient Descent," in *Proceedings of Machine Learning Research*, ser.

Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 1524–1534.

[70] Y. Li, C. Ma, Y. Chen, and Y. Chi, "Nonconvex matrix factorization from rank-one measurements," in *Proceedings of Machine Learning Research*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 1496–1505.

[71] A. Brutzkus and A. Globerson, "Globally optimal gradient descent for a ConvNet with Gaussian inputs," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 605–614.

[72] S. S. Du, J. D. Lee, and Y. Tian, "When is a convolutional filter easy to learn?" in *International Conference on Learning Representations*, 2018.

[73] S. Du, J. Lee, Y. Tian, A. Singh, and B. Poczos, "Gradient descent learns one-hidden-layer CNN: Dont be afraid of spurious local minima," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1339–1348.

[74] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 5164–5198, 2017.

[75] T. Adali, P. J. Schreier, and L. L. Scharf, "Complex-valued signal processing: The proper way to deal with impropriety," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5101–5125, 2011.

[76] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming."

[77] N. Boyd, G. Schiebinger, and B. Recht, "The alternating descent conditional gradient method for sparse inverse problems," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 616–639, 2017.

[78] N. Rao, P. Shah, and S. Wright, "Forward–backward greedy algorithms for atomic norm regularization," *IEEE Transactions on Signal Processing*, vol. 63, no. 21, pp. 5798–5811, 2015.

[79] K. Zhong, Z. Song, and I. S. Dhillon, "Learning non-overlapping convolutional neural networks with multiple kernels," *arXiv preprint arXiv:1711.03440*, 2017.

[80] H. Fu and Y. Chi, "Quantized spectral compressed sensing: Cramér-Rao bounds and recovery algorithms," *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3268–3279, 2018.

[81] ——, "Principal subspace estimation for low-rank toeplitz covariance matrices with binary sensing," in *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 1344–1348.

[82] H. Fu, Y. Chi, and Y. Liang, "Guaranteed recovery of one-hidden-layer neural networks via cross entropy," *arXiv preprint arXiv:1802.06463*, 2018.

[83] Y. Chen, Y. Chi, and A. J. Goldsmith, "Robust and universal covariance estimation from quadratic measurements via convex programming," in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 2017–2021.

[84] J.-F. Cai, X. Qu, W. Xu, and G.-B. Ye, "Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction," *Applied and computational harmonic analysis*, vol. 41, no. 2, pp. 470–490, 2016.

[85] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Compressed Sensing, Theory and Applications*, pp. 210 – 268, 2012.

[86] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

[87] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.

[88] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012, p. 210268.