# Guaranteed Recovery of One-Hidden-Layer Neural Networks via Cross Entropy

Haoyu Fu, *Student Member*, Yuejie Chi, *Senior Member*, Yingbin Liang, *Senior Member*

***Abstract*—We study model recovery for data classification, where the training labels are generated from a one-hidden-layer neural network with sigmoid activations, also known as a single-layer feedforward network, and the goal is to recover the weights of the neural network. We consider two network models, the fully-connected network (FCN) and the non-overlapping convolutional neural network (CNN). We prove that with Gaussian inputs, the empirical risk based on cross entropy exhibits strong convexity and smoothness *uniformly* in a local neighborhood of the ground truth, as soon as the sample complexity is sufficiently large. This implies that if initialized in this neighborhood, gradient descent converges linearly to a critical point that is provably close to the ground truth. Furthermore, we show such an initialization can be obtained via the tensor method. This establishes the global convergence guarantee for empirical risk minimization using cross entropy via gradient descent for learning one-hidden-layer neural networks, at the near-optimal sample and computational complexity with respect to the network input dimension without unrealistic assumptions such as requiring a fresh set of samples at each iteration.**

## I. Introduction

Neural networks have attracted a significant amount of research interest in recent years due to the success of deep neural networks [1] in practical domains such as computer vision and artificial intelligence [2], [3], [4]. However, the theoretical underpinnings behind such success remains mysterious to a large extent. Efforts have been taken to understand which classes of functions can be represented by deep neural networks [5], [6], [7], [8], when (stochastic) gradient descent is effective for optimizing a nonconvex loss function [9], and why these networks generalize well [10], [11], [12].

One important line of research that has attracted extensive attention is the model-recovery problem, which is important for the network to generalize well [13]. Specifically, it is shown in [13] that in a model-recovery setting, a network cannot generalize well if the underlying parameters cannot be recovered accurately, therefore linking model recovery to generalization. In addition, the problem of model recovery provides a framework to leverage statistical nature of the input data in an intuitive manner, which allows shedding more light to the understanding of optimization of complex neural networks.

H. Fu and Y. Liang are with Dept. of ECE, The Ohio State University, Columbus, OH 43210, USA. Emails: {fu.436, liang.889}osu.edu.

Y. Chi is with Dept. of ECE, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Email: yuejiechi@cmu.edu.

The work of H. Fu and Y. Liang is supported in part by U.S. National Science Foundation under the grants CCF-1761506, CCF-1801855 and CCF-1900145. The work of Y. Chi is supported in part by AFOSR under the grant FA9550-15-1-0205, by ONR under the grant N00014-18-1-2142, by ARO under the grant W911NF-18-1-0303, and by NSF under the grants CAREER ECCS-1818571, ECCS-1833553, CCF-1806154 and CCF-1901199.

Let the training samples $(\boldsymbol{x}_i, y_i) \sim (\boldsymbol{x}, y)$, $i = 1, \ldots, n$, be generated independently and identically distributed (i.i.d.) from a distribution $\mathcal{D}$ based on a neural network model with the ground truth parameter $\boldsymbol{W}^\star$, and the goal is to recover $\boldsymbol{W}^\star$ using the training samples given the network architecture. Consider a network whose output is given as $H(\boldsymbol{W}^\star, \boldsymbol{x})$. Previous studies along this topic can be mainly divided into two cases of data generations, with the input $\boldsymbol{x} \in \mathbb{R}^d$ being drawn from the Gaussian distribution.

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

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

  This type of regression problem has been studied in various settings. In particular, [14] studied the single-neuron model under the Rectified Linear Unit (ReLU) activation, [15] studied the one-hidden-layer multi-neuron network model, and [16] 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

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

  Such a problem has been studied in [17] 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. $\min_{\boldsymbol{W}} \frac{1}{n}\sum_{i=1}^n (y_i - H(\boldsymbol{W}, \boldsymbol{x}_i))^2$, given the training data. Two types of statistical guarantees were provided for such model recovery problems using the squared loss. More specifically, [15] 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 [17] 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 [15], compared to the uniform strong convexity in [17], is that independent fresh samples are required at each iteration to guarantee the linear convergence of gradient descent. Consequently, the sample complexity of [15] grows with respect to the recovery accuracy $\epsilon$, 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.

In this paper, we focus on the classification setting by minimizing the empirical loss using the cross entropy objective, which is a popular choice in training practical neural networks. The geometry as well as the optimization of the model recovery problem based on the cross-entropy loss function have not yet been understood even for one-hidden-layer networks. The main focus of this paper is to develop technical analysis for guaranteed model recovery under the challenging cross-entropy loss function for the classification problem for two types of one-hidden-layer network structures.

### A. Problem Formulation

We consider two popular types of one-hidden-layer nonlinear neural networks illustrated in Fig. 1, i.e., a Fully-Connected Network (FCN) [15] and a non-overlapping Convolutional Neural Network (CNN) [18]. 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

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

- *Non-overlapping CNN:* for simplicity we let $d = mK$ for some integers $m$. Let $\boldsymbol{w} \in \mathbb{R}^m$ be the network parameter, and the $k$th 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,

$$H_{\mathrm{CNN}}(\boldsymbol{w}, \boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K} \phi(\boldsymbol{w}^\top \boldsymbol{x}^{(k)}). \tag{2}$$

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

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

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 lot in previous literature [14], [19], [18], [20], 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.,

$$\mathbb{P}(y = 1|\boldsymbol{x}) = H(\boldsymbol{W}^\star, \boldsymbol{x}). \tag{3}$$

Our goal is to recover the network parameter, i.e., $\boldsymbol{W}^\star$. One natural choice is to maximize the log-likelihood function, which turns out to be equivalent to minimizing

$$f_n(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{W}; \boldsymbol{x}_i, y_i), \tag{4}$$

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

$$\begin{aligned} &\ell(\boldsymbol{W}; \boldsymbol{x}, y) \\ &= -y \cdot \log\left(H(\boldsymbol{W}, \boldsymbol{x})\right) - (1-y)\cdot \log\left(1 - H(\boldsymbol{W}, \boldsymbol{x})\right), \end{aligned} \tag{5}$$

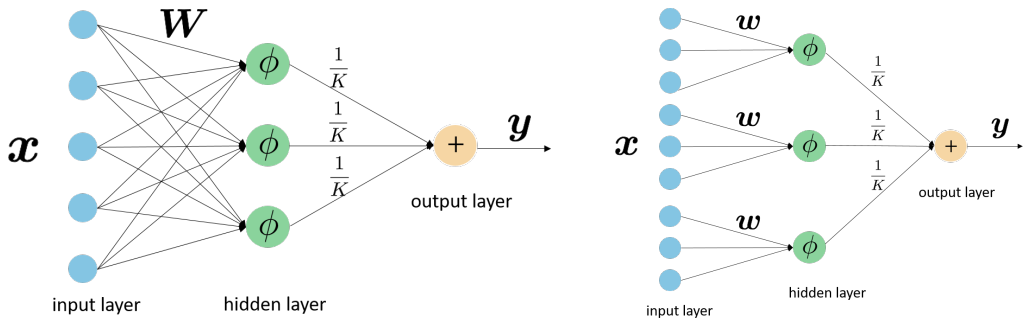
where $H(\boldsymbol{W}, \boldsymbol{x})$ can subsume either $H_{\mathrm{FCN}}$ or $H_{\mathrm{CNN}}$. Although the squared loss has been used in [17] to study the classification problem with a single neuron, the cross-entropy loss is a more natural and popular choice in practice for classification data, due to its natural connection to the principle of maximum likelihood estimation.

### B. Our Contributions

Considering the multi-neuron classification problem with either FCN or CNN, the main contributions of this work are summarized as follows. Throughout the discussions below, we assume the number $K$ of neurons is a constant, and state the scaling only in terms of the input dimension $d$ and the number $n$ of samples.

- *Uniform local strong convexity:* If the input is Gaussian, the empirical risk function $f_n(\boldsymbol{W})$ is *uniformly* strongly convex in a local neighborhood of the ground truth $\boldsymbol{W}^\star$ as soon as the sample size $n = O(d\log^2 d)$.
- *Statistical and computational rate of gradient descent:* consequently, 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. In particular, gradient descent finds the critical point $\widehat{\boldsymbol{W}}_n$ with a computation cost of $O(nd\log(1/\epsilon))$, where $\epsilon$ denotes the numerical accuracy and $\widehat{\boldsymbol{W}}_n$ converges to $\boldsymbol{W}^\star$ at a rate of $O(\sqrt{d\log n/n})$ in the Frobenius norm.
- *Tensor initialization:* We adopt the tensor method proposed in [15], and show that it provably provides an initialization in the neighborhood of the ground truth both for FCN and CNN. In particular, we strengthened the guarantee of the tensor method by replacing the homogeneous assumption on activation functions in [15] by a mild condition on the curvature of activation functions around $\boldsymbol{W}^\star$, which holds for a larger class of activation functions including *sigmoid* and *tanh*.

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 $H(\boldsymbol{W}, \boldsymbol{x})$ 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




(a) FCN (b) CNN

Fig. 1. Illustration of two types of one-hidden-layer neural networks considered in this paper: (a) a fully-connected network (FCN); (b) a non-overlapping convolutional neural network (CNN).

was also established in [15] 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 (Lemma 3) 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 [17], 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.

To the best of our knowledge, combining the analysis of gradient descent and initialization, this work provides the first globally convergent algorithm for the recovery of one-hidden-layer neural networks using the *cross-entropy* loss function.

### C. 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 [21]–[28], to name a few; see also the overview article [29]. 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 [30], [31], [32], [33]. For the case with multiple neurons ($2 \leq K \leq d$) in the under-parameterized setting, the work of Tian [34] studied the landscape of the population squared loss surface with ReLU activations. In particular, there exist spurious bad local minima in the optimization landscape [35], [36] even at the population level. Zhong et. al. [15] 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 ($K = 1$), [14] 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, [37] shows that randomly initialized gradient descent converges fast to the global optimum at a near-optimal sample complexity. On the other hand, [17] 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, [19] showed that projected gradient descent with a local initialization converges linearly for smooth activations with bounded second derivatives for the regression problem, [38] showed that gradient descent with tensor initialization converges linearly to a neighborhood of the ground truth using ReLU activations, and [39] showed the linear convergence of gradient descent with the spectral initialization using quadratic activations. For CNN with ReLU activations, [18] 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, [20] 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.

Finally, we note that several papers study one-hidden-layer or two-layer neural networks with different structures under Gaussian input. For example, [40] studied the overlapping convolutional neural network, [16] studied a two-layer feedforward networks with ReLU activations and identity mapping, and [41] introduced the Porcupine Neural Network.

### D. Paper Organization and Notations

The rest of the paper is organized as follows. Section II presents the main results on local geometry and local linear convergence of gradient descent. Section III discusses the initialization based on the tensor method. Numerical examples are demonstrated in Section IV, and finally, conclusions are drawn in Section V. Details of the technical proofs are delayed in the supplemental materials.

Throughout this paper, we use boldface letters to denote vectors and matrices, e.g. $\boldsymbol{w}$ and $\boldsymbol{W}$. The transpose of $\boldsymbol{W}$ is denoted by $\boldsymbol{W}^\top$, and $\|\boldsymbol{W}\|$, $\|\boldsymbol{W}\|_{\mathrm{F}}$ denote the spectral norm and the Frobenius norm. For a positive semidefinite (PSD) matrix $\boldsymbol{A}$, we write $\boldsymbol{A} \succeq 0$. 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.

Denote $\|\cdot\|_{\psi_1}$ as the sub-exponential norm of a random variable. We use $c, C, C_1, \ldots$ to denote constants whose values may vary from place to place. For nonnegative functions $f(x)$ and $g(x)$, $f(x) = O(g(x))$ means there exist positive constants $c$ and $a$ such that $f(x) \le cg(x)$ for all $x \ge a$; $f(x) = \Omega(g(x))$ means there exist positive constants $c$ and $a$ such that $f(x) \ge cg(x)$ for all $x \ge a$.

## II. Gradient Descent and its Performance Guarantee

To estimate the network parameter $\boldsymbol{W}^\star$, since (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 III. In this section, we focus on the performance of the local update rule

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

where $\eta$ is the constant step size. The algorithm is summarized in Algorithm 1.

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.

### A. 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

**Algorithm 1** Gradient Descent (GD)

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

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

**Output**: $\boldsymbol{W}_T$

to denote the local neighborhood of $\boldsymbol{W}^\star$ for FCN or of $\boldsymbol{w}^\star$ for CNN.

$$\mathbb{B}(\boldsymbol{W}^\star, r) = \left\{\boldsymbol{W} \in \mathbb{R}^{d\times K} : \|\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}} \le r\right\}, \quad (6a)$$

$$\mathbb{B}(\boldsymbol{w}^\star, r) = \left\{\boldsymbol{w} \in \mathbb{R}^m : \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \le r\right\}, \quad (6b)$$

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, $\sigma_i(\boldsymbol{W})$ denotes the $i$-th largest singular value of $\boldsymbol{W}^\star$. Let the condition number be $\kappa = \sigma_1/\sigma_K$, and $\lambda = \prod_{i=1}^K (\sigma_i/\sigma_K)$. Moreover, we introduce an important quantity $\rho(\sigma)$ regarding $\phi(z)$, the sigmoid activation function, that captures the geometric properties of the loss function for neural networks (1) and (2).

**Definition 1** (Key quantity for FCN)**.** *Let $z \sim \mathcal{N}(0,1)$ 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*

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

Note that Definition 1 for FCN is different from that in [15, Property 3.2] but consistent with [15, Lemma D.4] which removes the third term in [15, Property 3.2]. For the activation function considered in this paper, the first two terms suffice. Definition 2 for CNN is a newly distilled quantity in this paper 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 1** (Local Strong Convexity)**.** *Consider the classification model with FCN* (1) *or CNN* (2) *and the sigmoid activation function.*

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

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

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

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

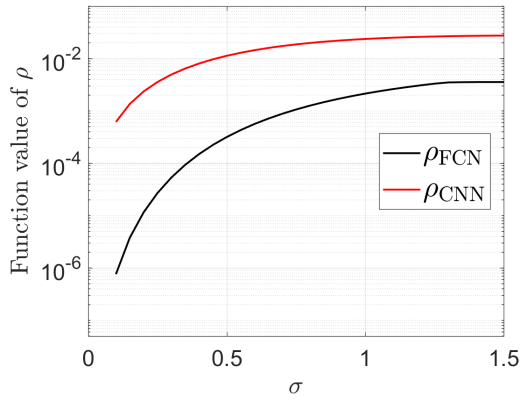



Fig. 2. Illustration $\rho(\sigma)$ for both FCN and CNN with the sigmoid activation.

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

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

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

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

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

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

We note that for FCN (1), all column permutations of $\boldsymbol{W}^\star$ are equivalent global minimum of the loss function, and Theorem 1 applies to all such permutation matrices of $\boldsymbol{W}^\star$. The proof of Theorem 1 is outlined in Appendix B.

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 $K, \kappa, \lambda$ are all positive. We depict $\rho(\sigma)$ as a function of $\sigma$ in a certain range for the sigmoid activation in Fig. 2. It can be seen from Fig. 2 that $\rho(\sigma)$ is monotonic increasing when $\sigma$ increases, and we have $\rho(\sigma) > 0$ as long as $\sigma > 0$. When $\boldsymbol{W}^\star$ is orthogonal, $\kappa$ and $\lambda$ 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 $\kappa$, $\lambda$, $\rho(\sigma)$ 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 1 guarantees that for both FCN (1) and CNN (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 1 depend on the dimension parameters of the network ($n$ and $K$), as well as the ground truth ($\rho_{\mathrm{FCN}}(\sigma_K)$, $\lambda$, $\rho_{\mathrm{CNN}}\left(\|\boldsymbol{w}^\star\|_2\right)$).

### B. 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 $\widehat{\boldsymbol{W}}_n$.

**Theorem 2** (Performance Guarantees of Gradient Descent)**.** *Assume the assumptions in Theorem 1 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*

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

*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.*

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

*for* $\eta \le c_3$, *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*

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

*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.*

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

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

Similarly to Theorem 1, for FCN (1) Theorem 2 also holds for all column permutations of $\boldsymbol{W}^\star$. The proof can be found in Appendix C. Theorem 2 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 (1) and $O\left(K\sqrt{d\log n/n}\right)$ for CNN(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 (1) and CNN (2) gradient descent converges linearly to $\widehat{\boldsymbol{W}}_n$ (or resp. $\widehat{\boldsymbol{w}}_n$) at a linear rate, as long as it is initialized in the basin of attraction. To achieve $\epsilon$-accuracy, i.e. $\left\|\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n\right\|_{\mathrm{F}} \le \epsilon$ (or resp. $\|\boldsymbol{w}_t - \widehat{\boldsymbol{w}}_n\|_2 \le \epsilon$), it requires a computational complexity of $O\left(ndK^4\log(1/\epsilon)\right)$ (or resp. $O\left(ndK^2\log(1/\epsilon)\right)$), which is linear in $n$, $d$ and $\log(1/\epsilon)$.

## III. Initialization via Tensor Method

Our initialization adopts the tensor method proposed in [15]. The initialization method works for the FCN model directly, and works for the CNN model with slight modification as presented in [42]. 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 [15], [42] for more details.

### A. Preliminary and Algorithm

We start with introducing the necessary definitions which can be found in [15]. 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

$$\boldsymbol{M}\widetilde{\otimes}\boldsymbol{I} = \sum_{i=1}^{r}\boldsymbol{s}_i\sum_{j=1}^{d}\sum_{l=1}^{6}\boldsymbol{A}_{l,i,j}, \tag{7}$$

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:*
$\boldsymbol{M}_1 = \mathbb{E}[y\cdot\boldsymbol{x}]$,
$\boldsymbol{M}_2 = \mathbb{E}[y\cdot(\boldsymbol{x}\otimes\boldsymbol{x}-\boldsymbol{I})]$,
$\boldsymbol{M}_3 = \mathbb{E}[y\cdot(\boldsymbol{x}^{\otimes 3}-\boldsymbol{x}\widetilde{\otimes}\boldsymbol{I})]$,
$\boldsymbol{M}_4 = \mathbb{E}[y\cdot(\boldsymbol{x}^{\otimes 4}-(\boldsymbol{x}\otimes\boldsymbol{x})\widetilde{\otimes}\boldsymbol{I}+\boldsymbol{I}\widetilde{\otimes}\boldsymbol{I})]$,
$m_{l,i} = g_{l,i}\left(\|\boldsymbol{w}_i^\star\|\right), \forall l = 0,1,2,3,4$,
*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

$$\sum_{i'}^{n_1}\sum_{j'}^{n_2}\sum_{k'}^{n_3}\boldsymbol{T}_{i',j',k'}\boldsymbol{A}_{i',i}\boldsymbol{B}_{j',j}\boldsymbol{C}_{k',k}. \tag{8}$$

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[15, 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.

Step 1 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.

Step 2 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 [15].

**Algorithm 2** Initialization via Tensor Method

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

### B. Performance Guarantee of Initialization

For the classification problem, we make the following technical assumptions, similarly to [15, 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*

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

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

*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., $\phi(az)=a^pz$ for an integer $p$) required in [15], 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 $\delta$ 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 3.** *For the classification model* (1)*, under Assumptions 1 and 2, for any $0<\epsilon<1$ and $\zeta>1$, if the sample size $n\geq d\cdot\mathrm{poly}(K,\kappa,\zeta,\log d,1/\epsilon)$, then the output $\boldsymbol{W}_0\in\mathbb{R}^{d\times K}$ of Algorithm 2 satisfies*

$$\|\boldsymbol{W}_0-\boldsymbol{W}^\star\|_{\mathrm{F}}\leq\epsilon\mathrm{poly}(K,\kappa)\|\boldsymbol{W}^\star\|_{\mathrm{F}}, \tag{9}$$

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

The proof of Theorem 3 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 [15], 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 E.

## IV. 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 $\ell$th run as $\widehat{\boldsymbol{w}}_n^{(\ell)} = \mathrm{vec}(\widehat{\boldsymbol{W}}_n^{(\ell)})$ and the mean of the runs as $\bar{\boldsymbol{w}}$. The error is calculated as $\mathrm{SD}_n = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L} \|\widehat{\boldsymbol{w}}_n^{(\ell)} - \bar{\boldsymbol{w}}\|^2}$, where $L = 20$ is the total number of random initializations. Adopted from [17], 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 $\mathrm{SD}_n \le 10^{-4}$. We generate the ground truth $\boldsymbol{W}^\star$ from Gaussian matrices, and the training samples are generated using the FCN (1). Figure 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.

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 the ground truth such that $\|\boldsymbol{W}_0 - \boldsymbol{W}^\star\|_{\mathrm{F}} \le 0.1 \cdot \|\boldsymbol{W}^\star\|_{\mathrm{F}}$. 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 (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 (2). In Fig. 4 (a), we say an experiment is successful if $\mathrm{SD}_n \le 10^{-14}$, 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. 4 (b). Suppose we initialized in the neighborhood of $\boldsymbol{w}^\star$, i.e., $\|\boldsymbol{w}_0 - \boldsymbol{w}^\star\|_2 \le 0.9 \cdot \|\boldsymbol{w}^\star\|_2$, 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.

## V. Conclusions

In this paper, we have studied the model recovery problem of a 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. In the future, it will be interesting to extend the analysis in this paper to more general class of activation functions, particularly ReLU-like activations.

## Appendix A
## Gradient and Hessian of 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.

### A. 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

$$\mathbb{E}\left[\frac{\partial \ell\left(\boldsymbol{W}\right)}{\partial \boldsymbol{w}_j}\right] = \mathbb{E}\left[-\frac{1}{K}\frac{\left(y - H(\boldsymbol{W})\right)}{H(\boldsymbol{W})\left(1 - H(\boldsymbol{W})\right)}\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\boldsymbol{x}\right], \tag{10}$$

$$\mathbb{E}\left[\frac{\nabla^2 \ell\left(\boldsymbol{W}\right)}{\partial \boldsymbol{w}_j \partial \boldsymbol{w}_l}\right] = \mathbb{E}\left[\xi_{j,l}\left(\boldsymbol{W}\right) \cdot \boldsymbol{x}\boldsymbol{x}^\top\right], \tag{11}$$

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

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

and when $j = l$,

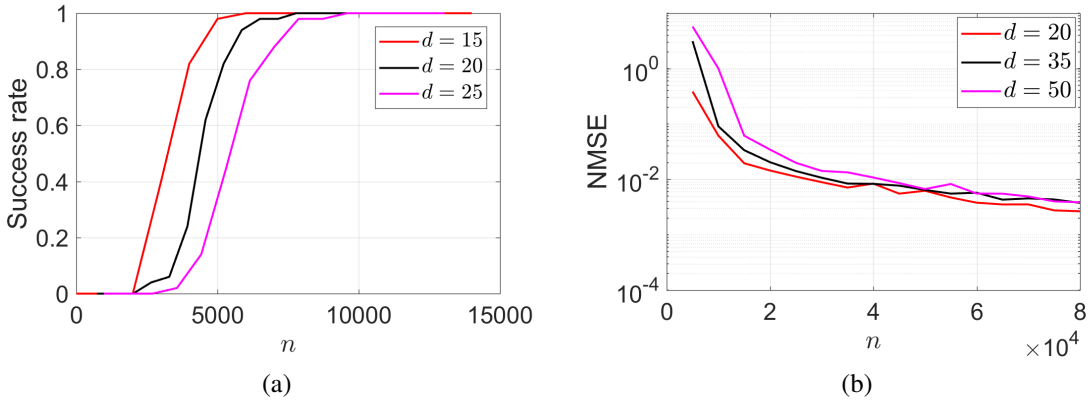
$$\xi_{j,j}\left(\boldsymbol{W}\right) = \frac{1}{K^2}\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)^2 \cdot \frac{H\left(\boldsymbol{W}\right)^2 + y - 2y \cdot H\left(\boldsymbol{W}\right)}{H^2(\boldsymbol{W})\left(1 - H(\boldsymbol{W})\right)^2} - \frac{1}{K}\phi''\left(\boldsymbol{w}_j^\top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

### B. The CNN case

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

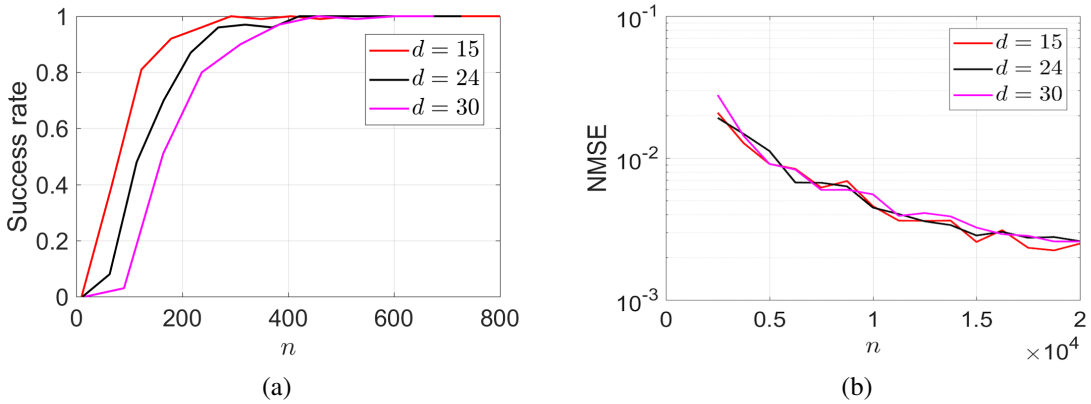
$$\mathbb{E}\left[\frac{\partial \ell\left(\boldsymbol{w}\right)}{\partial \boldsymbol{w}}\right] = \mathbb{E}\left[-\phi'(\boldsymbol{w}^\top \boldsymbol{x}^{(1)}) \cdot \frac{y - H\left(\boldsymbol{w}\right)}{H\left(\boldsymbol{w}\right)\left(1 - H\left(\boldsymbol{w}\right)\right)} \cdot \boldsymbol{x}^{(1)}\right], \tag{12}$$

$$\mathbb{E}\left[\frac{\nabla^2 \ell\left(\boldsymbol{w}\right)}{\partial \boldsymbol{w}^2}\right] = \mathbb{E}\left[\sum_{j=1}^{K}\sum_{l=1}^{K} g_{j,l}\left(\boldsymbol{w}\right)\boldsymbol{x}^{(j)}\boldsymbol{x}^{(l)\top}\right], \tag{13}$$




(a) (b)

Fig. 3. For FCN (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 $d$.




(a) (b)

Fig. 4. For CNN (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 $d$.

where when $j \neq l$,

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

and when $j = l$,

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

## Appendix B
## Proof of Theorem 1

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 1);
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 2);
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 3).
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 1** (Hessian Smoothness of Population Loss)**.** *Suppose the loss $\ell(\cdot)$ associates with FCN (1), and assume $\|\boldsymbol{w}_k^\star\|_2 \leq 1$ for all $k$ and $\|\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}} \leq 0.7$. Then we have*

$$\|\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)\| \leq \frac{C_1}{K^{\frac{3}{2}}} \cdot \|\boldsymbol{W} - \boldsymbol{W}^\star\|_{\mathrm{F}}, \quad (14)$$

*holds. Similarly, suppose the loss $\ell(\cdot)$ associates with CNN (2), and assume $\|\boldsymbol{w}^\star\|_2 \leq 1$ and $\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \leq 0.7$. We have*

$$\|\nabla^2 f(\boldsymbol{w}) - \nabla^2 f(\boldsymbol{w}^\star)\| \leq C_2 \cdot K \cdot \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2, \quad (15)$$

*holds. Here $C_1$ and $C_2$ denote some large constants.*

The proof is provided in Appendix D-A. Together with the fact that $\nabla^2 f(\boldsymbol{W}^\star)$ be lower and upper bounded, Lemma 1

allows us to bound $\nabla^2 f(\boldsymbol{W})$ in a neighborhood around ground truth, given below.

**Lemma 2** (Local Strong Convexity and Smoothness of Population Loss)**.** *If the loss $\ell(\cdot)$ associates with FCN* (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* (2)*, then we have*

$$C_3\cdot\frac{\rho_{\mathrm{CNN}}(\|\boldsymbol{w}^\star\|_2)}{K}\cdot\boldsymbol{I}\preceq\nabla^2 f(\boldsymbol{w})\preceq C_4\cdot K\cdot\boldsymbol{I},\qquad(16)$$

*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 D-B. 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 3.** *If the loss $\ell(\cdot)$ associates with FCN* (1)*, then 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*

$$\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r_{\mathrm{FCN}})}\|\nabla^2 f_n(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W})\|\leq C\sqrt{\frac{dK\log n}{n}},\qquad(17)$$

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

$$\sup_{\boldsymbol{w}\in\mathbb{B}(\boldsymbol{w}^\star,r_{\mathrm{CNN}})}\|\nabla^2 f_n(\boldsymbol{w})-\nabla^2 f(\boldsymbol{w})\|\leq CK^2\sqrt{\frac{\frac{d}{K}\cdot\log(n)}{n}},\qquad(18)$$

*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 D-C. 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 1.* 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 $\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r_{\mathrm{FCN}})$. □

## Appendix C
## Proof of Theorem 2

We have established that $f_n(\boldsymbol{W})$ is strongly convex in $\mathbb{B}(\boldsymbol{W}^\star,r)$ in Theorem 1. Thus there exists at most one critical point in $\mathbb{B}(\boldsymbol{W}^\star,r)$. The proof of Theorem 2 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 4), and then invoke [17, 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 $\widehat{\boldsymbol{W}}_n$ 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 4.** *If the loss $\ell(\cdot)$ associates with FCN* (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* (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*

$$\sup_{\boldsymbol{w}\in\mathbb{B}(\boldsymbol{w}^\star,r_{\mathrm{CNN}})}\|\nabla f_n(\boldsymbol{w})-\nabla f(\boldsymbol{w})\|\leq C\cdot\sqrt{\frac{d\log n}{n}}\qquad(19)$$

*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 D-D. 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 3 and Lemma 4, we can invoke [17, 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 mean value theorem, there exists $\boldsymbol{W}'$ one the straight line connecting $\boldsymbol{W}^\star$ and $\widehat{\boldsymbol{W}}_n$ such that

$$\begin{aligned}f_n\left(\widehat{\boldsymbol{W}}_n\right)&=f_n(\boldsymbol{W}^\star)+\left\langle\nabla f_n(\boldsymbol{W}^\star),\mathrm{vec}\left(\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\right)\right\rangle\\&\quad+\frac{1}{2}\mathrm{vec}\left(\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\right)^\top\nabla^2 f_n(\boldsymbol{W}')\,\mathrm{vec}\left(\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\right)\\&\leq f_n(\boldsymbol{W}^\star),\end{aligned}\qquad(20)$$

where the last inequality follows from the optimality of $\widehat{\boldsymbol{W}}_n$. By Theorem 1, we have

$$\begin{aligned}&\frac{1}{2}\mathrm{vec}\left(\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\right)^\top\nabla^2 f_n\left(\boldsymbol{W}'\right)\mathrm{vec}\left(\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\right)\\&\geq\Omega\left(\frac{1}{K^2}\cdot\frac{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}{\kappa^2\lambda}\right)\left\|\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\right\|_{\mathrm{F}}^2.\end{aligned}\tag{21}$$

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

$$\begin{aligned}&\left|\left\langle\nabla f_n\left(\boldsymbol{W}^\star\right),\mathrm{vec}\left(\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\right)\right\rangle\right|\\&\leq\|\nabla f_n\left(\boldsymbol{W}^\star\right)\|_2\|\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\|_{\mathrm{F}}\\&\leq\Omega\left(\sqrt{\frac{dK^{1/2}\log n}{n}}\right)\|\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\|_{\mathrm{F}},\end{aligned}\tag{22}$$

where the last line follows from Lemma 4. Plugging (21) and (22) into (20), we have

$$\|\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^\star\|_{\mathrm{F}}\leq\Omega\left(\frac{K^{\frac{9}{4}}\kappa^2\lambda}{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}\sqrt{\frac{d\log n}{n}}\right).\tag{23}$$

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 1, 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

$$\begin{aligned}\|\boldsymbol{W}_{t+1}-\widehat{\boldsymbol{W}}_n\|_{\mathrm{F}}&\leq\|\boldsymbol{I}-\eta\int_0^1\nabla^2 f_n\left(\boldsymbol{W}\left(\gamma\right)\right)\|\|\boldsymbol{W}_t-\widehat{\boldsymbol{W}}_n\|_{\mathrm{F}}\\&\leq\left(1-\eta H_{\min}\right)\|\boldsymbol{W}_t-\widehat{\boldsymbol{W}}_n\|_{\mathrm{F}}.\end{aligned}\tag{24}$$

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

$$\|\boldsymbol{W}_{t+1}-\widehat{\boldsymbol{W}}_n\|_{\mathrm{F}}\leq\left(1-\frac{H_{\min}}{H_{\max}}\right)\|\boldsymbol{W}_t-\widehat{\boldsymbol{W}}_n\|_{\mathrm{F}},\tag{25}$$

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

## REFERENCES


[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] 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.

[3] 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.

[4] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[5] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[6] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[7] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information theory*, vol. 39, no. 3, pp. 930–945, 1993.

[8] M. Telgarsky, "benefits of depth in neural networks," in *Conference on Learning Theory*, 2016, pp. 1517–1539.

[9] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Advances in neural information processing systems*, 2014, pp. 2933–2941.

[10] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[11] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 6241–6250.

[12] A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz, "SGD learns over-parameterized networks that provably generalize on linearly separable data," in *International Conference on Learning Representations*, 2018.

[13] M. Mondelli and A. Montanari, "On the connection between learning two-layer neural networks and tensor decomposition," in *Proceedings of Machine Learning Research*, 2019, pp. 1051–1060.

[14] M. Soltanolkotabi, "Learning relus via gradient descent," in *Advances in Neural Information Processing Systems*, 2017, pp. 2007–2017.

[15] 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.

[16] 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.

[17] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for nonconvex losses," *Ann. Statist.*, vol. 46, no. 6A, pp. 2747–2774, 12 2018.

[18] 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.

[19] S. Oymak, "Learning compact neural networks with regularization," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 3966–3975.

[20] S. S. Du, J. D. Lee, and Y. Tian, "When is a convolutional filter easy to learn?" in *International Conference on Learning Representations*, 2018.

[21] 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.

[22] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation," *IEEE Signal Processing Magazine*, 2018.

[23] 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.

[24] R. Ge and T. Ma, "On the optimization landscape of tensor decompositions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 3653–3663.

[25] 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.

[26] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery using nonconvex optimization," *International Conference on Machine Learning*, pp. 2351–2360, 2015.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] D. Boob and G. Lan, "Theoretical properties of the global optimizer of two layer neural network," *arXiv preprint arXiv:1710.11241*, 2017.

[32] 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.

[33] Q. Nguyen and M. Hein, "The loss surface of deep and wide neural networks," in *International Conference on Machine Learning*, 2017, pp. 2603–2612.

[34] 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.

[35] R. Ge, J. D. Lee, and T. Ma, "Learning one-hidden-layer neural networks with landscape design," in *International Conference on Learning Representations*, 2018.

[36] I. Safran and O. Shamir, "Spurious local minima are common in two-layer ReLU neural networks," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 4433–4441.

[37] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Math. Program. 176, 5–37*, 2019.

[38] X. Zhang, Y. Yu, L. Wang, and Q. Gu, "Learning one-hidden-layer relu networks via gradient descent," *International Conference on Artificial Intelligence and Statistics*, 2019.

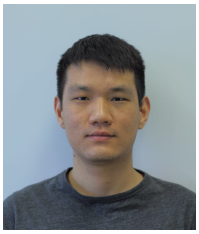
[39] Y. Li, C. Ma, Y. Chen, and Y. Chi, "Nonconvex matrix factorization from rank-one measurements," *arXiv preprint arXiv:1802.06286*, 2018.

[40] S. Goel, A. Klivans, and R. Meka, "Learning one convolutional layer with overlapping patches," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1783–1791.

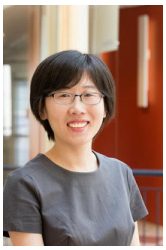
[41] S. Feizi, H. Javadi, J. Zhang, and D. Tse, "Porcupine neural networks: Approximating neural network landscapes," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 4836–4846.

[42] K. Zhong, Z. Song, and I. S. Dhillon, "Learning non-overlapping convolutional neural networks with multiple kernels," *arXiv preprint arXiv:1711.03440*, 2017.

[43] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Compressed Sensing, Theory and Applications*, pp. 210 – 268, 2012.

**Haoyu Fu** received the Ph.D. degree in Electrical and Computer Engineering from The Ohio State University, Columbus, OH, USA, in 2019. His research interests include compressed sensing, statistical signal processing, optimization, and machine learning.



**Yuejie Chi** (S'09-M'12-SM'17) received the Ph.D. degree in Electrical Engineering from Princeton University in 2012, and the B.E. (Hon.) degree in Electrical Engineering from Tsinghua University, Beijing, China, in 2007. She was with The Ohio State University from 2012 to 2017. Since 2018, she is an Associate Professor with the department of Electrical and Computer Engineering at Carnegie Mellon University, where she holds the Robert E. Doherty Early Career Development Professorship. Her research interests include signal processing, statistical inference, machine learning, large-scale optimization, and their applications in data science, inverse problems, imaging, and sensing systems. She is a recipient of the PECASE Award, NSF CAREER Award, AFOSR and ONR Young Investigator Program Awards, Ralph E. Powe Junior Faculty Enhancement Award, and Google Faculty Research Award. She received IEEE SPS Early Career Technical Achievement Award and Young Author Best Paper Award from IEEE Signal Processing Society, and Best Paper Award at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). She has served as an Elected Member of the SPTM, SAM and MLSP Technical Committees of the IEEE Signal Processing Society. She currently serves as an Associate Editor of IEEE Trans. on Signal Processing.

**Dr. Yingbin Liang** (S'01–M'05-SM'16) is currently a Professor at the Department of Electrical and Computer Engineering at the Ohio State University (OSU). She received the Ph.D. degree in Electrical Engineering from the University of Illinois at Urbana-Champaign in 2005, and served on the faculty of University of Hawaii and Syracuse University before she joined OSU. Dr. Liang's research interests include machine learning, optimization, information theory and statistical signal processing. Dr. Liang received the National Science Foundation CAREER Award and the State of Hawaii Governor Innovation Award in 2009. She also received EURASIP Best Paper Award for the EURASIP Journal on Wireless Communications and Networking in 2014. She served as an Associate Editor for the Shannon Theory of the IEEE Transactions on Information Theory during 2013-2015.

# Supplementary Materials: Additional Proofs

## APPENDIX D
## PROOF OF AUXILIARY LEMMAS

### A. Proof of Lemma 1.

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,

$$\|\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)\| = \max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^\top (\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star))\boldsymbol{a} = \max_{\|\boldsymbol{a}\|=1} \sum_{j=1}^K \sum_{l=1}^K \boldsymbol{a}_j^\top \boldsymbol{\Delta}_{j,l} \boldsymbol{a}_l. \tag{26}$$

From (11) we know that

$$\boldsymbol{\Delta}_{j,l} = \frac{\partial^2 f(\boldsymbol{W})}{\partial \boldsymbol{w}_j \partial \boldsymbol{w}_l} - \frac{\partial^2 f(\boldsymbol{W}^\star)}{\partial \boldsymbol{w}_j^\star \partial \boldsymbol{w}_l^\star} = \mathbb{E}\left[ \left( \xi_{j,l}(\boldsymbol{W}) - \xi_{j,l}(\boldsymbol{W}^\star) \right) \cdot \boldsymbol{x}\boldsymbol{x}^\top \right], \tag{27}$$

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

$$\xi_{j,l}(\boldsymbol{W}) = \xi_{j,l}(\boldsymbol{W}^\star) + \sum_{k=1}^K \left\langle \frac{\partial \xi_{j,l}\left(\widetilde{\boldsymbol{W}}\right)}{\partial \widetilde{\boldsymbol{w}}_k}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle, \tag{28}$$

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

$$\boldsymbol{\Delta}_{j,l} = \mathbb{E}\left[ \left( \sum_{k=1}^K \left\langle \frac{\partial \xi_{j,l}\left(\widetilde{\boldsymbol{W}}\right)}{\partial \widetilde{\boldsymbol{w}}_k}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle \right) \cdot \boldsymbol{x}\boldsymbol{x}^\top \right], \tag{29}$$

which can be further simplified as

$$\boldsymbol{\Delta}_{j,l} = \mathbb{E}\left[ \left( \sum_{k=1}^K T_{j,l,k} \left\langle \boldsymbol{x}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle \right) \cdot \boldsymbol{x}\boldsymbol{x}^\top \right], \tag{30}$$

by the fact that $\frac{\partial \xi_{j,l}(\widetilde{\boldsymbol{W}})}{\partial \widetilde{\boldsymbol{w}}_k}$ 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 $\frac{\partial \xi_{j,l}(\widetilde{\boldsymbol{W}})}{\partial \widetilde{\boldsymbol{w}}_k}$ for illustration,

$$\frac{\partial \xi_{j,j}(\boldsymbol{W})}{\partial \boldsymbol{w}_k} = \begin{cases} \left( -\frac{2}{K^2} \frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)^2}{H(\boldsymbol{W})^3} + \frac{1}{K} \frac{\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})^2} \right) \frac{1}{K} \phi\left(\boldsymbol{w}_k^\top \boldsymbol{x}\right) \boldsymbol{x} & k \neq j \\ \left( \frac{2}{K^2} \left( \frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})^2} - \frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)^2}{H(\boldsymbol{W})^3} \right) + \frac{1}{K} \left( \frac{\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})^2} - \frac{\phi'''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})} \right) \right) \frac{1}{K} \phi\left(\boldsymbol{w}_k^\top \boldsymbol{x}\right) \boldsymbol{x} & k = j \end{cases}, \tag{31}$$

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

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

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

$$|T_{j,l,k}| \leq \max\left\{ \frac{2}{K^3} \frac{1}{H(\widetilde{\boldsymbol{W}})^3}, \frac{1}{K^2} \frac{1}{H(\widetilde{\boldsymbol{W}})^2}, \frac{2}{K^3} \frac{1}{(1 - H(\widetilde{\boldsymbol{W}}))^3}, \frac{1}{K^2} \frac{1}{(1 - H(\widetilde{\boldsymbol{W}}))^2} \right\}, \tag{32}$$

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,

$$\begin{aligned} \boldsymbol{a}_j^\top \boldsymbol{\Delta}_{j,l} \boldsymbol{a}_l &= \mathbb{E}\left[ \left( \sum_{k=1}^K T_{j,l,k} \left\langle \boldsymbol{x}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle \right) \cdot \left( \boldsymbol{a}_j^\top \boldsymbol{x} \right) \left( \boldsymbol{a}_l^\top \boldsymbol{x} \right) \right] \\ &\leq \sqrt{ \mathbb{E}\left[ \sum_{k=1}^K T_{j,l,k}^2 \right] \cdot \mathbb{E}\left[ \sum_{k=1}^K \left( \left\langle \boldsymbol{x}, \boldsymbol{w}_k - \boldsymbol{w}_k^\star \right\rangle \left( \boldsymbol{a}_j^\top \boldsymbol{x} \right) \left( \boldsymbol{a}_l^\top \boldsymbol{x} \right) \right)^2 \right] } \\ &\leq \sqrt{ \sum_{k=1}^K \mathbb{E}\left[ T_{j,l,k}^2 \right] } \cdot \sqrt{ \sum_{k=1}^K \|\boldsymbol{w}_k - \boldsymbol{w}_k^\star\|_2^2 \cdot \|\boldsymbol{a}_j\|_2^2 \cdot \|\boldsymbol{a}_l\|_2^2 }. \end{aligned} \tag{33}$$

Plug it back to (26) we can obtain the following inequality,

$$\|\nabla^2 f(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W}^\star)\| \leq \max_{\|\boldsymbol{a}\|=1} \sum_{j=1}^{K}\sum_{l=1}^{K} \sqrt{\sum_{k=1}^{K}\mathbb{E}\left[T_{j,l,k}^2\right]} \cdot \sqrt{\sum_{k=1}^{K}\|\boldsymbol{w}_k-\boldsymbol{w}_k^\star\|_2^2\cdot\|\boldsymbol{a}_j\|_2^2\cdot\|\boldsymbol{a}_l\|_2^2}. \tag{34}$$

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 D-E.

**Lemma 5.** *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*

$$\mathbb{E}\left[\left(\frac{1}{\frac{1}{K}\sum_{j=1}^{K}\phi\left(\boldsymbol{w}_j^\top\boldsymbol{x}\right)}\right)^z\right]\leq C_1\cdot e^{t^2},\quad \mathbb{E}\left[\left(\frac{1}{\left(1-\frac{1}{K}\sum_{j=1}^{K}\phi\left(\boldsymbol{w}_j^\top\boldsymbol{x}\right)\right)}\right)^z\right]\leq C_2\cdot e^{t^2} \tag{35}$$

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

Setting $z=4$ and $z=6$ in Lemma 5, together with (32) we obtain that

$$\mathbb{E}\left[T_{j,l,k}^2\right]\leq\frac{C}{K^4}\cdot e^{\max_{1\leq i\leq k}\|\widetilde{\boldsymbol{w}}_i\|_2^2}, \tag{36}$$

holds for some constant $C$. Plugging (36) into (34), we obtain

$$\|\nabla^2 f(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W}^\star)\|\leq\frac{C}{K^{\frac{3}{2}}}e^{\|\widetilde{\boldsymbol{W}}\|_{\mathrm{F}}^2}\cdot\|\boldsymbol{W}-\boldsymbol{W}^\star\|_{\mathrm{F}}\cdot\max_{\|\boldsymbol{a}\|=1}\sum_{j=1}^{K}\sum_{l=1}^{K}\|\boldsymbol{a}_j\|_2\|\boldsymbol{a}_l\|_2\leq\frac{C}{K^{\frac{3}{2}}}e^{\|\widetilde{\boldsymbol{W}}\|_{\mathrm{F}}^2}\cdot\|\boldsymbol{W}-\boldsymbol{W}^\star\|_{\mathrm{F}}. \tag{37}$$

Further since $e^{\max_{1\leq i\leq k}\|\widetilde{\boldsymbol{w}}_i\|_2^2}\leq C$ 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

$$\|\nabla^2 f(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W}^\star)\|\leq\frac{C}{K^{\frac{3}{2}}}\|\boldsymbol{W}-\boldsymbol{W}^\star\|_{\mathrm{F}} \tag{38}$$

holds for some constant $C$.

- **The CNN case:** according to (13), we can calculate the upper bound of $\|\nabla^2 f(\boldsymbol{w})-\nabla^2 f(\boldsymbol{w}^\star)\|$ by definition as

$$\|\nabla^2 f(\boldsymbol{w})-\nabla^2 f(\boldsymbol{w}^\star)\|\leq\max_{\|\boldsymbol{u}\|_2=1}\sum_{j=1}^{K}\sum_{l=1}^{K}\mathbb{E}\left[\left(g_{j,l}(\boldsymbol{w})-g_{j,l}(\boldsymbol{w}^\star)\right)\cdot\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\cdot\boldsymbol{x}^{(l)\top}\boldsymbol{u}\right]. \tag{39}$$

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

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

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

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

where $S_{j,l,k}$ is a scalar that depends on $\widetilde{\boldsymbol{w}}$ 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

$$\begin{aligned}K^2\cdot\frac{\partial g_{j,l}(\boldsymbol{w})}{\partial\boldsymbol{w}}=&\frac{(1-H(\boldsymbol{w}))\,\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}^{(j)}\right)\phi''\left(\boldsymbol{w}^\top\boldsymbol{x}^{(l)}\right)}{(1-H(\boldsymbol{w}))^3}\cdot\boldsymbol{x}^{(l)}+\frac{(1-H(\boldsymbol{w}))\,\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}^{(l)}\right)\phi''\left(\boldsymbol{w}^\top\boldsymbol{x}^{(j)}\right)}{(1-H(\boldsymbol{w}))^3}\cdot\boldsymbol{x}^{(j)}\\&-\frac{\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}^{(l)}\right)\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}^{(j)}\right)}{(1-H(\boldsymbol{w}))^3}\cdot\left(\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{x}^{(k)}\right),\end{aligned} \tag{40}$$

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

$$S_{j,l,l}=\frac{1}{K^2}\frac{(1-H(\boldsymbol{w}))\,\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}^{(j)}\right)\phi''\left(\boldsymbol{w}^\top\boldsymbol{x}^{(l)}\right)}{(1-H(\boldsymbol{w}))^3}-\frac{1}{K^3}\frac{\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}^{(l)}\right)\phi'\left(\boldsymbol{w}^\top\boldsymbol{x}^{(j)}\right)}{(1-H(\boldsymbol{w}))^3}. \tag{41}$$

and

$$|S_{j,l,l}|\leq\frac{1}{K^2}\frac{1}{(1-H(\widetilde{\boldsymbol{w}}))^3}, \tag{42}$$

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

$$|S_{j,l,k}| \leq \begin{cases} \max\left\{\frac{1}{K^2}\frac{1}{(1-H(\widetilde{\boldsymbol{w}}))^3}, \frac{1}{K^2}\frac{1}{(H(\widetilde{\boldsymbol{w}}))^3}\right\} & j \neq l \\ \max\left\{\frac{1}{K}\frac{1}{(1-H(\widetilde{\boldsymbol{w}}))^2}, \frac{1}{K}\frac{1}{(H(\widetilde{\boldsymbol{w}}))^2}\right\} & j = l \end{cases}. \tag{43}$$

Plug these back to (39) we obtain

$$\begin{aligned} \|\nabla^2 f(\boldsymbol{w}) - \nabla^2 f(\boldsymbol{w}^\star)\| &\leq \max_{\|\boldsymbol{u}\|_2=1} \sum_{j=1}^K \sum_{l=1}^K \mathbb{E}\left[\sum_{k=1}^K \left\langle S_{j,l,k} \cdot \boldsymbol{x}^{(k)}, \boldsymbol{w} - \boldsymbol{w}^\star \right\rangle \cdot \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \cdot \boldsymbol{x}^{(l)\top} \boldsymbol{u}\right] \\ &= \max_{\|\boldsymbol{u}\|_2=1} \sum_{j=1}^K \sum_{l=1}^K \mathbb{E}\left[\sum_{k=1}^K S_{j,l,k} \cdot (\boldsymbol{w} - \boldsymbol{w}^\star)^\top \boldsymbol{x}^{(k)} \cdot \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \cdot \boldsymbol{x}^{(l)\top} \boldsymbol{u}\right] \\ &\leq \max_{\|\boldsymbol{u}\|_2=1} \sum_{j=1}^K \sum_{l=1}^K \sqrt{\mathbb{E}\left[\sum_{k=1}^K S_{j,l,k}^2\right] \cdot \mathbb{E}\left[\sum_{k=1}^K \left((\boldsymbol{w} - \boldsymbol{w}^\star)^\top \boldsymbol{x}^{(k)}\right)^2 \left(\boldsymbol{u}^\top \boldsymbol{x}^{(j)}\right)^2 \left(\boldsymbol{x}^{(l)\top} \boldsymbol{u}\right)^2\right]} \\ &\leq \max_{\|\boldsymbol{u}\|_2=1} \sum_{j=1}^K \sum_{l=1}^K \sqrt{\mathbb{E}\left[\sum_{k=1}^K S_{j,l,k}^2\right] \cdot \sum_{k=1}^K \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 \cdot \|\boldsymbol{u}\|_2^2 \cdot \|\boldsymbol{u}\|_2^2} \\ &\leq C \cdot K \cdot e^{\|\widetilde{\boldsymbol{w}}\|_2^2} \cdot \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2, \end{aligned} \tag{44}$$

where the second inequality follows from Cauchy-Schwarz inequality, and the last inequality follows from (43) and Lemma 5. Further since $e^{\|\widetilde{\boldsymbol{w}}\|_2^2} \leq C \cdot \left(1 + \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2\right)$ given that $\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \leq 0.7$, we conclude that

$$\|\nabla^2 f(\boldsymbol{w}) - \nabla^2 f(\boldsymbol{w}^\star)\| \leq C \cdot K \cdot \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \tag{45}$$

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

### *B. Proof of Lemma 2*

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 1 to obtain a uniform bound in the neighborhood of $\boldsymbol{W}^\star$.

- **The FCN case:** Recall

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

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

$$\nabla^2 f(\boldsymbol{W}^\star) \succeq \left(\min_{\|\boldsymbol{a}\|_2=1} \boldsymbol{a}^\top \nabla^2 f(\boldsymbol{W}^\star) \boldsymbol{a}\right) \cdot \boldsymbol{I} = \min_{\|\boldsymbol{a}\|_2=1} \frac{1}{K^2} \mathbb{E}\left[\frac{\left(\sum_{j=1}^K \phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)\right)^2}{H(\boldsymbol{W}^\star)(1 - H(\boldsymbol{W}^\star))}\right] \cdot \boldsymbol{I}. \tag{46}$$

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

$$\nabla^2 f(\boldsymbol{W}^\star) \succeq \min_{\|\boldsymbol{a}\|_2=1} \frac{4}{K^2} \mathbb{E}\left[\left(\sum_{j=1}^K \phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)\right)^2\right] \cdot \boldsymbol{I} \succeq \frac{4}{K^2} \cdot \frac{\rho_{\mathrm{FCN}}(\sigma_K)}{\kappa^2 \lambda} \cdot \boldsymbol{I}, \tag{47}$$

where the last inequality follows from [15, Lemmas D.4 and D.6]. To derive an upper bound of $\nabla^2 f(\boldsymbol{W}^\star)$, we have

$$\nabla^2 f(\boldsymbol{W}^\star) \preceq \left(\max_{\|\boldsymbol{a}\|_2=1} \boldsymbol{a}^\top \nabla^2 f(\boldsymbol{W}^\star) \boldsymbol{a}\right) \cdot \boldsymbol{I} = \max_{\|\boldsymbol{a}\|_2=1} \frac{1}{K^2} \mathbb{E}\left[\frac{\left(\sum_{j=1}^K \phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)\right)^2}{\frac{1}{K^2}\sum_{j,l} \phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1 - \phi\left(\boldsymbol{w}_l^{\star\top}\boldsymbol{x}\right)\right)}\right]. \tag{48}$$

Then by Cauchy-Schwarz inequality, we have

$$\frac{\left(\sum_{j=1}^K \phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)\right)^2}{\frac{1}{K^2}\sum_{j,l} \phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1 - \phi\left(\boldsymbol{w}_l^{\star\top}\boldsymbol{x}\right)\right)} \leq \frac{\left(\sum_{j=1}^K \phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)^2\right) \cdot \left(\sum_{j=1}^K \left(\boldsymbol{a}_j^\top \boldsymbol{x}\right)^2\right)}{\frac{1}{K^2}\sum_{j,l} \phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1 - \phi\left(\boldsymbol{w}_l^{\star\top}\boldsymbol{x}\right)\right)}. \tag{49}$$

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

$$\sum_{j,l}\phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1-\phi\left(\boldsymbol{w}_l^{\star\top}\boldsymbol{x}\right)\right) \ge \sum_{j=1}^{K}\phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\left(1-\phi\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)\right) = \sum_{j=1}^{K}\phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right) \ge 4\sum_{j=1}^{K}\phi'\left(\boldsymbol{w}_j^{\star\top}\boldsymbol{x}\right)^2, \tag{50}$$

we obtain

$$\boldsymbol{a}^\top\nabla^2 f\left(\boldsymbol{W}^\star\right)\boldsymbol{a} \preceq \frac{1}{K^2}\mathbb{E}\left[\frac{CK^2}{4}\sum_{j=1}^{K}\left(\boldsymbol{a}_j^\top\boldsymbol{x}\right)^2\right]. \tag{51}$$

Plugging (51) back to (48), we obtain

$$\nabla^2 f\left(\boldsymbol{W}^\star\right) \preceq C\cdot\boldsymbol{I}. \tag{52}$$

Thus together with the lower bound (47), we conclude that

$$\frac{4}{K^2}\cdot\frac{\rho_{\mathrm{FCN}}\left(\sigma_K\right)}{\kappa^2\lambda}\cdot\boldsymbol{I} \preceq \nabla^2 f\left(\boldsymbol{W}^\star\right) \preceq C\cdot\boldsymbol{I}. \tag{53}$$

From Lemma 1, we have

$$\|\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^\star)\| \lesssim \frac{C}{K^{\frac{3}{2}}}\|\boldsymbol{W}-\boldsymbol{W}^\star\|_F. \tag{54}$$

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

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

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 (13), we have

$$\nabla^2 f\left(\boldsymbol{w}^\star\right) = \mathbb{E}\left[\frac{\frac{1}{K^2}\sum_{j,l}\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(l)}\right)\boldsymbol{x}^{(j)}\boldsymbol{x}^{(l)\top}}{H\left(\boldsymbol{w}^\star\right)\left(1-H\left(\boldsymbol{w}^\star\right)\right)}\right]. \tag{55}$$

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

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

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

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

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,

$$\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right]\cdot\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(l)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(l)}\right]=\left(\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\right)^2\left(\mathbb{E}\left[\phi'\left(z\right)z\right]\right)^2=0, \tag{56}$$

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 $\phi'\left(z\right)z=-\left(\phi'\left(-z\right)\cdot(-z)\right)$. Similarly,

$$\begin{aligned}
&\mathbb{E}\left[\left(\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right)^2\right]\\
&=\mathbb{E}\Bigg[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2\cdot\left(\left(\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2+\left(\left(\boldsymbol{u}-\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\cdot\boldsymbol{w}^\star\right)^\top\boldsymbol{x}^{(j)}\right)^2\right)\Bigg]\\
&=\left(\frac{\boldsymbol{u}^\top\boldsymbol{w}^\star}{\|\boldsymbol{w}^\star\|_2^2}\right)^2\cdot\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2\right]+\left(\|\boldsymbol{u}\|_2^2-\frac{\left(\boldsymbol{u}^\top\boldsymbol{w}^\star\right)^2}{\|\boldsymbol{w}^\star\|_2^2}\right)\cdot\mathbb{E}\left[\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2\right].
\end{aligned} \tag{57}$$

Together with Definition 2, we have

$$\mathbb{E}\left[\left(\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right)^2\right]\geq\rho_{\mathrm{CNN}}\left(\|\boldsymbol{w}^\star\|_2\right). \tag{58}$$

Hence,

$$\nabla^2 f\left(\boldsymbol{w}^\star\right)\succeq\frac{4}{K}\cdot\rho_{\mathrm{CNN}}\left(\|\boldsymbol{w}^\star\|_2\right)\cdot\boldsymbol{I}. \tag{59}$$

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

$$\nabla^2 f\left(\boldsymbol{w}^\star\right)\leq\left(\max_{\|\boldsymbol{u}\|_2=1}\boldsymbol{u}^\top\nabla^2 f\left(\boldsymbol{w}^\star\right)\boldsymbol{u}\right)\cdot\boldsymbol{I}\leq\max_{\|\boldsymbol{u}\|_2=1}\mathbb{E}\left[\frac{\sum_{j=1}^K\left(\frac{1}{K}\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)\right)^2\cdot\sum_{j=1}^K\left(\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right)^2}{H\left(\boldsymbol{w}^\star\right)\left(1-H\left(\boldsymbol{w}^\star\right)\right)}\right]\cdot\boldsymbol{I}. \tag{60}$$

Using (50), i.e.,

$$\frac{\frac{1}{K^2}\sum_{j=1}^K\phi'\left(\boldsymbol{w}^{\star\top}\boldsymbol{x}^{(j)}\right)^2}{H\left(\boldsymbol{w}^\star\right)\left(1-H\left(\boldsymbol{w}^\star\right)\right)}\leq\frac{1}{4}, \tag{61}$$

we upper bound the right-hand side of (60) as

$$\nabla^2 f\left(\boldsymbol{w}^\star\right)\preceq\max_{\|\boldsymbol{u}\|_2=1}\mathbb{E}\left[\frac{1}{4}\sum_{j=1}^K\left(\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\right)^2\right]\cdot\boldsymbol{I}=\frac{K}{4}\cdot\boldsymbol{I}. \tag{62}$$

Together with the lower bound, we now conclude that

$$\frac{4}{K}\cdot\rho_{\mathrm{CNN}}\left(\|\boldsymbol{w}^\star\|_2\right)\cdot\boldsymbol{I}\preceq\nabla^2 f\left(\boldsymbol{w}^\star\right)\preceq\frac{K}{4}\cdot\boldsymbol{I}. \tag{63}$$

And following from (15) in Lemma 1, we have

$$\|\nabla^2 f\left(\boldsymbol{w}\right)-\nabla^2 f\left(\boldsymbol{w}^\star\right)\|\leq C\cdot K\cdot\|\boldsymbol{w}-\boldsymbol{w}^\star\|_2. \tag{64}$$

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

$$C\cdot\frac{\rho_{\mathrm{CNN}}\left(\|\boldsymbol{w}^\star\|_2\right)}{K}\cdot\boldsymbol{I}\preceq\nabla^2 f\left(\boldsymbol{w}\right)\preceq C\cdot K\cdot\boldsymbol{I}. \tag{65}$$

### *C. Proof of Lemma 3*

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 [17] to our setting. Let $N_\epsilon$ be the $\epsilon$-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 $\log N_\epsilon\leq dK\log\left(3r/\epsilon\right)$ [43]. Let $\mathcal{W}_\epsilon=\{\boldsymbol{W}_1,\cdots,\boldsymbol{W}_{N_\epsilon}\}$ be the $\epsilon$-cover set with $N_\epsilon$ elements. For any $\boldsymbol{W}\in\mathbb{B}\left(\boldsymbol{W}^\star,r\right)$, let $j\left(\boldsymbol{W}\right)=\mathrm{argmin}_{j\in[N_\epsilon]}\|\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}\|_{\mathrm{F}}\leq\epsilon$ 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

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

where the events $A_t$, $B_t$ and $C_t$ are defined as

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

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

1) **Upper bound on $\mathbb{P}\left(B_t\right)$.** Before continuing, we state a useful technical lemma, whose proof can be found in [17].

**Lemma 6.** *Let $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ be a symmetric $d \times d$ matrix and $V_\epsilon$ be an $\epsilon$-cover of unit-Euclidean-norm ball $\mathbb{B}\left(\mathbf{0}, 1\right)$, then*

$$\|\boldsymbol{M}\| \leq \frac{1}{1-2 \epsilon} \sup_{\boldsymbol{v} \in V_\epsilon}|\langle\boldsymbol{v}, \boldsymbol{M} \boldsymbol{v}\rangle|. \tag{66}$$

Let $V_{\frac{1}{4}}$ be a $\left(\frac{1}{4}\right)$-cover of the ball $\mathbb{B}(\mathbf{0}, 1)=\{\boldsymbol{W} \in \mathbb{R}^{d \times K}:\|\boldsymbol{W}\|_{\mathrm{F}}=1\}$, where $\log \left|V_{\frac{1}{4}}\right| \leq dK \log 12$. Following from Lemma 6, we have

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

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

$$\mathbb{P}\left(B_t\right) \leq \mathbb{P}\left(\sup_{\boldsymbol{W} \in \mathcal{W}_\epsilon, \boldsymbol{v} \in V_{\frac{1}{4}}}\left|\frac{1}{n} \sum_{i=1}^n G_i\right| \geq \frac{t}{6}\right) \leq e^{dK\left(\log \frac{3 r}{\epsilon}+\log 12\right)} \sup_{\boldsymbol{W} \in \mathcal{W}_\epsilon, \boldsymbol{v} \in V_{\frac{1}{4}}} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n G_i\right| \geq \frac{t}{6}\right), \tag{67}$$

where $G_i=\left\langle\boldsymbol{v},\left(\nabla^2 \ell\left(\boldsymbol{W} ; \boldsymbol{x}_i\right)-\mathbb{E}\left[\nabla^2 \ell\left(\boldsymbol{W} ; \boldsymbol{x}\right)\right]\right) \boldsymbol{v}\right\rangle$ 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 D-F.

**Lemma 7.** *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 [17, Theorem 9] to (67), we have that for fixed $\boldsymbol{W} \in \mathcal{W}_\epsilon, \boldsymbol{v} \in V_{\frac{1}{4}}$,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n\left\langle\boldsymbol{v},\left(\nabla^2 \ell\left(\boldsymbol{W} ; \boldsymbol{x}_i\right)-\mathbb{E}\left[\nabla^2 \ell\left(\boldsymbol{W} ; \boldsymbol{x}\right)\right]\right) \boldsymbol{v}\right\rangle\right| \geq \frac{t}{6}\right) \leq 2 \exp\left(-c \cdot n \cdot \min\left(\frac{t^2}{\tau^4}, \frac{t}{\tau^2}\right)\right), \tag{68}$$

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

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

Thus as long as

$$t>C \cdot \max\left\{\sqrt{\frac{\tau^4\left(dK \log \frac{36 r}{\epsilon}+\log \frac{4}{\delta}\right)}{n}}, \frac{\tau^2\left(dK \log \frac{36 r}{\epsilon}+\log \frac{4}{\delta}\right)}{n}\right\} \tag{69}$$

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

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 D-H

**Lemma 8.** *Suppose the loss is associated with FCN. There exists some constant $C$ such that*

$$\mathbb{E}\left[\sup_{\boldsymbol{W}\neq\boldsymbol{W}'\in\mathbb{B}(\boldsymbol{W}^\star,r)}\frac{\|\nabla^2\ell(\boldsymbol{W},\boldsymbol{x})-\nabla^2\ell(\boldsymbol{W}',\boldsymbol{x})\|}{\|\boldsymbol{W}-\boldsymbol{W}'\|_{\mathrm{F}}}\right]\leq C\cdot d\sqrt{K}. \tag{70}$$

Consider the event $C_t$ first. We derive

$$\begin{aligned}
&\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\|\mathbb{E}\left[\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}\right)\right]-\mathbb{E}\left[\nabla^2\ell(\boldsymbol{W};\boldsymbol{x})\right]\|\\
&\leq\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\frac{\|\mathbb{E}\left[\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}\right)\right]-\mathbb{E}\left[\nabla^2\ell(\boldsymbol{W};\boldsymbol{x})\right]\|}{\|\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}\|_{\mathrm{F}}}\cdot\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\|\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}\|_{\mathrm{F}}\\
&\leq C\cdot d\sqrt{K}\cdot\epsilon.
\end{aligned} \tag{71}$$

Therefore, $C_t$ holds as long as

$$t\geq C\cdot d\sqrt{K}\cdot\epsilon. \tag{72}$$

We can bound the event $A_t$ as below.

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\frac{1}{n}\left\|\sum_{i=1}^n\left[\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)-\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i\right)\right]\right\|\geq\frac{t}{3}\right)\\
&\leq\frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\left\|\frac{1}{n}\sum_{i=1}^n\left[\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)-\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i\right)\right]\right\|\right] &(73)\\
&\leq\frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\left\|\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)-\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i\right)\right\|\right]\\
&\leq\frac{3}{t}\mathbb{E}\left[\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\frac{\|\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)-\nabla^2\ell\left(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i\right)\|}{\|\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}\|_{\mathrm{F}}}\right]\cdot\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\|\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}\|_{\mathrm{F}}\\
&\leq\frac{C\cdot d\sqrt{K}\cdot\epsilon}{t} &(74)
\end{aligned}$$

where (73) follows from the Markov inequality. Thus, taking

$$t\geq\frac{6\epsilon\cdot C\cdot d\sqrt{K}}{\delta} \tag{75}$$

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

3) **Final step.** Let $\epsilon=\frac{\delta\tau^2}{C\cdot d\sqrt{K}\cdot ndK}$ and $\delta=d^{-10}$. Plugging $\epsilon$ and $\delta$ into (69) we need

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

The middle term can be bounded as

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

If $n\geq C\cdot dK\log(dK)$ 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 $n\geq 3$. Thus we can set

$$t\geq\tau^2\sqrt{\frac{\left(dK\log(36rnd^{11}K)+\log\frac{4}{\delta}\right)}{n}} \tag{76}$$

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

$$\mathbb{P}\left(\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r)}\|\nabla^2 f_n(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W})\|\geq C\tau^2\sqrt{\frac{dK\log n}{n}}\right)\leq d^{-10}. \tag{77}$$

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

**Lemma 9.** *Suppose the loss is associated CNN. There exists some constant $C$ such that*

$$\|G_i\|_{\psi_1} \leq C \cdot K^2 := \tau^2. \tag{78}$$

**Lemma 10.** *Suppose the loss is associated with CNN. There exists some constant $C$ such that*

$$\mathbb{E}\left[\sup_{\boldsymbol{W}\neq\boldsymbol{W}'\in\mathbb{B}(\boldsymbol{W}^\star,r)} \frac{\|\nabla^2\ell\left(\boldsymbol{W},\boldsymbol{x}\right)-\nabla^2\ell\left(\boldsymbol{W}',\boldsymbol{x}\right)\|}{\|\boldsymbol{W}-\boldsymbol{W}'\|_{\mathrm{F}}}\right] \leq C\cdot d\sqrt{K}. \tag{79}$$

Following argument similar to the proof of Lemma 3, we can obtain the following concentration inequality:

$$\sup_{\boldsymbol{w}\in\mathbb{B}(\boldsymbol{w}^\star,r)} \|\nabla^2 f_n\left(\boldsymbol{w}\right)-\nabla^2 f\left(\boldsymbol{w}\right)\| \leq C\cdot K^2\sqrt{\frac{\frac{d}{K}\cdot\log n}{n}}, \tag{80}$$

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

### *D. Proof of Lemma 4*

In order to proceed we need the following Lemma 11 whose proof is given in Appendix D-J.

**Lemma 11.** *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 3, and applies Lemma 11, for the loss associated with FCN, we can get the following concentration inequality

$$\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^\star,r_{\mathrm{FCN}})} \|\nabla f_n\left(\boldsymbol{W}\right)-\nabla f\left(\boldsymbol{W}\right)\|_2 \leq C\cdot\sqrt{\frac{d\sqrt{K}\log n}{n}} \tag{81}$$

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

$$\sup_{\boldsymbol{w}\in\mathbb{B}(\boldsymbol{w}^\star,r_{\mathrm{CNN}})} \|\nabla f_n\left(\boldsymbol{w}\right)-\nabla f\left(\boldsymbol{w}\right)\| \leq C\cdot\sqrt{K}\sqrt{\frac{\frac{d}{K}\log n}{n}} = C\cdot\sqrt{\frac{d\log n}{n}}, \tag{82}$$

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

### *E. Proof of Lemma 5*

We take the first term in (35) as an example, since the second term follows exactly in the same way. We first derive

$$\mathbb{E}\left[\left(\frac{1}{K}\sum_{i=1}^K\phi\left(\boldsymbol{w}_i^\top\boldsymbol{x}\right)\right)^{-z}\right] \leq \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^K\left(\phi\left(\boldsymbol{w}_i^\top\boldsymbol{x}\right)\right)^{-z}\right], \tag{83}$$

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

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

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 $p\geq 1$,

$$\mathbb{E}\left[\left(\frac{1}{\frac{1}{K}\sum_{i=1}^K\phi\left(\boldsymbol{w}_i^\top\boldsymbol{x}\right)}\right)^z\right] \leq C\cdot e^{t^2}, \tag{84}$$

holds for some constant $C$ depending on $z$.

### F. Proof of Lemma 7

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 2. 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

$$\| \left\langle \boldsymbol{u}, \nabla^2 \ell\left(\boldsymbol{W}; z\right) \boldsymbol{u} \right\rangle \|_{\psi_1} \leq \sum_{j=1}^{K}\sum_{l=1}^{K} \|\xi_{j,l} \cdot \boldsymbol{u}_j^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u}_l\|_{\psi_1} \leq \sum_{j=1}^{K}\sum_{l=1}^{K} \sup_{t\geq 1} \quad t^{-1}\left(\mathbb{E}\left|\xi_{j,l} \cdot \boldsymbol{u}_j^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{u}_l\right|^t\right)^{\frac{1}{t}}. \tag{85}$$

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

- For $j \neq l$,

$$|\xi_{j,l}| = \left| \frac{1}{K^2} \frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_l^\top \boldsymbol{x}\right) \cdot \left(H\left(\boldsymbol{W}\right)^2 + y - 2y \cdot H\left(\boldsymbol{W}\right)\right)}{\left(H\left(\boldsymbol{W}\right)\left(1 - H\left(\boldsymbol{W}\right)\right)\right)^2} \right| = \begin{cases} \frac{1}{K^2}\frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_l^\top \boldsymbol{x}\right)}{\left(1-H(\boldsymbol{W})\right)^2} & y = 0 \\ \frac{1}{K^2}\frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_l^\top \boldsymbol{x}\right)}{H(\boldsymbol{W})^2} & y = 1 \end{cases}. \tag{86}$$

  Moreover,

$$\frac{1}{K^2}\frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_l^\top \boldsymbol{x}\right)}{\left(1 - H\left(\boldsymbol{W}\right)\right)^2} \leq \frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_l^\top \boldsymbol{x}\right)}{\left(1 - \phi\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\right)\left(1 - \phi\left(\boldsymbol{w}_l^\top \boldsymbol{x}\right)\right)} \leq \phi\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi\left(\boldsymbol{w}_l^\top \boldsymbol{x}\right) \leq 1, \tag{87}$$

  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

$$\frac{1}{K^2}\frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_l^\top \boldsymbol{x}\right)}{H\left(\boldsymbol{W}\right)^2} \leq 1. \tag{88}$$

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

- For $j = l$,

$$|\xi_{j,j}| \leq \left| \frac{1}{K^2} \frac{\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\phi'\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right) \cdot \left(H\left(\boldsymbol{W}\right)^2 + y - 2y \cdot H\left(\boldsymbol{W}\right)\right)}{\left(H\left(\boldsymbol{W}\right)\left(1 - H\left(\boldsymbol{W}\right)\right)\right)^2} \right| + \left| \frac{1}{K} \frac{\phi''\left(\boldsymbol{w}_j^\top \boldsymbol{x}\right)\left(y - H\left(\boldsymbol{W}\right)\right)}{H\left(\boldsymbol{W}\right)\left(1 - H\left(\boldsymbol{W}\right)\right)} \right|. \tag{89}$$

  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,

$$\begin{aligned} \left\|\left\langle \boldsymbol{u}, \nabla^2 \ell\left(\boldsymbol{W}; z\right) \boldsymbol{u} \right\rangle\right\|_{\psi_1} &\leq C \cdot \sum_{j=1}^{K}\sum_{l=1}^{K} \sup_{t\geq 1} \quad t^{-1}\left(\sqrt{\mathbb{E}\left[\left(\boldsymbol{u}_j^\top \boldsymbol{x}\right)^{2t}\right]} \cdot \sqrt{\mathbb{E}\left[\left(\boldsymbol{u}_l^\top \boldsymbol{x}\right)^{2t}\right]}\right)^{\frac{1}{t}} \\ &\leq C \cdot \sum_{j=1}^{K}\sum_{l=1}^{K} \|\boldsymbol{u}_j\|_2 \|\boldsymbol{u}_l\|_2 \cdot \sup_{t\geq 1} \quad t^{-1}\left((2t-1)!!\right)^{\frac{1}{t}} \\ &\leq C :\equiv \tau^2, \end{aligned} \tag{90}$$

where the last inequality holds because

$$\begin{aligned} &\sup_{t\geq 1} \quad t^{-1}\left((2t-1)!!\right)^{\frac{1}{t}} \leq \sup_{t\geq 1} \quad t^{-1}\left((2t)^t\right)^{\frac{1}{t}} \leq 2, \\ &\sum_{j=1}^{K}\sum_{l=1}^{K} \|\boldsymbol{u}_j\|_2 \|\boldsymbol{u}_l\|_2 \leq \sum_{j=1}^{K}\sum_{l=1}^{K} \frac{\|\boldsymbol{u}_j\|_2^2 + \|\boldsymbol{u}_l\|_2^2}{2} = \frac{1}{2}. \end{aligned} \tag{91}$$

Thus, we conclude

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

### G. Proof of Lemma 9

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 2. 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

$$\| \left\langle \boldsymbol{u}, \nabla^2 \ell\left(\boldsymbol{w}\right) \boldsymbol{u} \right\rangle \|_{\psi_1} \leq \sum_{j \neq l} \|g_{j,l}\left(\boldsymbol{w}\right) \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \boldsymbol{u}^\top \boldsymbol{x}^{(l)}\|_{\psi_1} + \sum_{j=l} \|g_{j,l}\left(\boldsymbol{w}\right) \boldsymbol{u}^\top \boldsymbol{x}^{(j)} \boldsymbol{u}^\top \boldsymbol{x}^{(l)}\|_{\psi_1}. \tag{92}$$

Hence, we have

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

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

Plugging it back to (92), we obtain

$$\| \left\langle \boldsymbol{u}, \nabla^2 \ell\left(\boldsymbol{w}\right) \boldsymbol{u} \right\rangle \|_{\psi_1} \leq \sum_{j \neq l} \| \left(\boldsymbol{u}^\top \boldsymbol{x}^{(j)}\right) \left(\boldsymbol{u}^\top \boldsymbol{x}^{(l)}\right) \|_{\psi_1} + \sum_{j=1}^{K} \| \left(\boldsymbol{u}^\top \boldsymbol{x}^{(j)}\right)^2 \|_{\psi_1} \leq C \cdot K^2. \tag{93}$$

### H. Proof of Lemma 8

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,

$$\|\nabla^2 \ell\left(\boldsymbol{W}; z\right) - \nabla^2 \ell\left(\boldsymbol{W}'; z\right)\| \leq \sum_{j=1}^{K} \sum_{l=1}^{K} |\xi_{j,l}\left(\boldsymbol{W}\right) - \xi_{j,l}\left(\boldsymbol{W}'\right)| \cdot \|\boldsymbol{x}\boldsymbol{x}^\top\|. \tag{94}$$

Using the same method as shown in the proof of Lemma 1, we can upper bound $|\xi_{j,l}\left(\boldsymbol{W}\right) - \xi_{j,l}\left(\boldsymbol{W}'\right)|$ as

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

where following from (32),

$$|T_{j,l,k}| \leq \max\left\{ \frac{2}{K^3} \frac{1}{H\left(\boldsymbol{W}\right)^3}, \frac{1}{K^2} \frac{1}{H\left(\boldsymbol{W}\right)^2}, \frac{2}{K^3} \frac{1}{\left(1 - H\left(\boldsymbol{W}\right)\right)^3}, \frac{1}{K^2} \frac{1}{\left(1 - H\left(\boldsymbol{W}\right)\right)^2} \right\}. \tag{95}$$

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

$$\mathbb{E}\left[ \sup_{\boldsymbol{W} \neq \boldsymbol{W}'} \frac{\|\nabla^2 \ell\left(\boldsymbol{W}\right) - \nabla^2 \ell\left(\boldsymbol{W}'\right)\|}{\|\boldsymbol{W} - \boldsymbol{W}'\|_F} \right] \leq \sqrt{K} \cdot K^2 \cdot \mathbb{E}\left[ \left(\max_{j,l,k} |T_{j,l,k}|\right) \cdot \|\boldsymbol{x}\|_2 \cdot \|\boldsymbol{x}\boldsymbol{x}^\top\| \right] \leq C \cdot d\sqrt{K}. \tag{96}$$

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

### I. Proof of Lemma 10

Following from (13) we can write

$$\|\nabla^2 \ell\left(\boldsymbol{w}\right) - \nabla^2 \ell\left(\boldsymbol{w}'\right)\| \leq \sum_{j=1}^{K} \sum_{l=1}^{K} |g_{j,l}\left(\boldsymbol{w}\right) - g_{j,l}\left(\boldsymbol{w}'\right)| \cdot \|\boldsymbol{x}^{(j)} \boldsymbol{x}^{(l)\top}\|. \tag{97}$$

Similarly, the analysis in the proof of Lemma 1 implies that

$$|g_{j,l}\left(\boldsymbol{w}\right) - g_{j,l}\left(\boldsymbol{w}'\right)| \leq \left(\max_k |S_{j,l,k}|\right) \cdot \sqrt{K} \|\boldsymbol{x}\|_2 \cdot \|\boldsymbol{w} - \boldsymbol{w}'\|_2, \tag{98}$$

where we upper-bound $S_{j,l,k}$ in (43) as

$$|S_{j,l,k}| \leq \begin{cases} \max\left\{ \frac{1}{K^2} \frac{1}{(1 - H(\boldsymbol{w}))^3}, \frac{1}{K^2} \frac{1}{(H(\boldsymbol{w}))^3} \right\} & j \neq l \\ \max\left\{ \frac{1}{K} \frac{1}{(1 - H(\boldsymbol{w}))^2}, \frac{1}{K} \frac{1}{(H(\boldsymbol{w}))^2} \right\} & j = l \end{cases}. \tag{99}$$

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

$$\mathbb{E}\left[\sup_{\boldsymbol{w}\neq\boldsymbol{w}'}\frac{\|\nabla^2\ell(\boldsymbol{w})-\nabla^2\ell(\boldsymbol{w}')\|}{\|\boldsymbol{w}-\boldsymbol{w}'\|_F}\right] \leq \sqrt{K}\cdot\sum_{j=1}^{K}\sum_{l=1}^{K}\mathbb{E}\left[\left(\max_k|S_{j,l,k}|\right)\cdot\|\boldsymbol{x}\|_2\cdot\|\boldsymbol{x}^{(j)}\boldsymbol{x}^{(l)\top}\|\right] \leq C\cdot d\sqrt{K}. \tag{100}$$

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

### *J. Proof of Lemma 11*

- **The FCN case:** Following from (10), we have

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

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

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

  Thus we conclude that

$$\|\langle\nabla\ell(\boldsymbol{W}),\boldsymbol{u}\rangle\|_{\psi_2} \leq \sum_{j=1}^{K}\|\boldsymbol{u}_j\|_2 \leq \sqrt{K}, \tag{101}$$

  and the directional gradient is $\sqrt{K}$-sub-Gaussian.
- **The CNN case:** Following from (12), we have

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

  where

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

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

$$\|\langle\nabla\ell(\boldsymbol{w}),\boldsymbol{u}\rangle\|_{\psi_2} \leq K\cdot\frac{1}{K}\sum_{j=1}^{K}\|\boldsymbol{u}^\top\boldsymbol{x}^{(j)}\|_{\psi_2} \leq C\cdot K. \tag{102}$$

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

## APPENDIX E
## PROOF OF THEOREM 3

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 [15] 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 $i=1,\ldots,K$, [15] showed that for the regression problem, if the sample size $n \geq d\mathrm{poly}(K,\kappa,\zeta,\log d)$, where $\zeta>1$ is any constant, then

$$\|\overline{\boldsymbol{w}_i}^\star - s_i\boldsymbol{V}\widehat{\boldsymbol{u}}_i\| \leq \epsilon\mathrm{poly}(\mathrm{K},\kappa) \tag{103}$$

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 [15]. 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 [15] 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 [15], 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:

$$Q_1 = \boldsymbol{M}_{l_1}(\boldsymbol{I}, \underbrace{\boldsymbol{\alpha}, \cdots, \boldsymbol{\alpha}}_{(l_1-1)}), \tag{104}$$

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

$$Q_1 = \boldsymbol{M}_3(\boldsymbol{I}, \boldsymbol{\alpha}, \boldsymbol{\alpha}) = \frac{1}{K}\sum_{i=1}^{K} m_{3,i}(\|\boldsymbol{w}_i^\star\|)\left(\boldsymbol{\alpha}^\top \overline{\boldsymbol{w}}_i^\star\right)^2 \overline{\boldsymbol{w}}_i^\star, \tag{105}$$

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

$$m_{3,i}(\|\boldsymbol{w}_i^\star\|) = \mathbb{E}\left[\phi\left(\|\boldsymbol{w}_i^\star\| \cdot z\right) z^3\right] - 3\mathbb{E}\left[\phi\left(\|\boldsymbol{w}_i^\star\| \cdot z\right) z\right]. \tag{106}$$

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

$$\beta^\star = \mathrm{argmin}_{\beta \in \mathbb{R}^K} \left\| \frac{1}{K}\sum_{i=1}^{K} \beta_i s_i \overline{\boldsymbol{w}_i}^\star - Q_1 \right\|, \tag{107}$$

where each entry of the solution takes the form

$$\beta_i^\star = s_i^3 m_{3,i}(\|\boldsymbol{w}_i^\star\|)\left(\boldsymbol{\alpha}^T s_i \overline{\boldsymbol{w}_i}^\star\right)^2. \tag{108}$$

In the initialization, we substitute $\widehat{Q}_1$ (estimated from training data) for $Q_1$, $\boldsymbol{V}\widehat{u}_i$ (estimated in part (a)) for $s_i\overline{\boldsymbol{w}_i}^\star$ into (107), and obtain an estimate $\widehat{\beta}$ of $\beta^\star$. We then substitute $\widehat{\beta}$ for $\beta^\star$ and $\boldsymbol{V}\widehat{u}_i$ for $s_i\overline{\boldsymbol{w}_i}^\star$ into (108) to obtain an estimate $\widehat{a}_i$ of $\|\boldsymbol{w}_i^\star\|$ via the following equation

$$\widehat{\beta}_i = s_i^3 m_{3,i}(\widehat{a}_i)\left(\boldsymbol{\alpha}^\top \boldsymbol{V}\widehat{u}_i\right)^2. \tag{109}$$

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 $\widehat{\beta}_i$ for $i = 1, \ldots, K$.

For notational simplicity, let $\beta_{1,i}^\star := \frac{\beta_i^\star}{s_i^3(\boldsymbol{\alpha}^\top s_i \overline{\boldsymbol{w}_i}^\star)^2}$ and $\widehat{\beta}_{1,i} := \frac{\widehat{\beta}_i}{s_i^3(\boldsymbol{\alpha}^\top \boldsymbol{V}\widehat{u}_i)^2}$, and then (108) and (109) become

$$\widehat{\beta}_{1,i} = m_{3,i}(\widehat{a}_i), \quad \beta_{1,i}^\star = m_{3,i}(\|\boldsymbol{w}_i^\star\|). \tag{110}$$

By Assumption 2 and (108), 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 $\Gamma$. By employing the result in [15], if the sample size $n \geq d\mathrm{poly}\left(K, \kappa, t, \log d\right)$, then $\widehat{Q}_1$ and $Q_1$, $\boldsymbol{V}\widehat{u}_i$ and $s_i\overline{\boldsymbol{w}_i}^\star$ can be arbitrarily close so that $|\beta_{1,i}^\star - \widehat{\beta}_{1,i}| < \min\{\delta', \frac{r}{\sqrt{K}\Gamma}\}$.

Thus, by (110) and the mean value theorem, we obtain

$$|\widehat{a}_i - \|\boldsymbol{w}_i^\star\|| = |g'(\xi)||\beta_{1,i}^\star - \widehat{\beta}_{1,i}|, \tag{111}$$

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