

How does promoting the minority fraction affect generalization? A theoretical study of one-hidden-layer neural network on group imbalance

Hongkang Li, *Student Member, IEEE*; Shuai Zhang, *Member, IEEE*; Yihua Zhang, Meng Wang, *Senior Member, IEEE*; Sijia Liu, *Member, IEEE*; and Pin-Yu Chen, *Member, IEEE*

Abstract—Group imbalance has been a known problem in empirical risk minimization (ERM), where the achieved high average accuracy is accompanied by low accuracy in a *minority group*. Despite algorithmic efforts to improve the minority group accuracy, a theoretical generalization analysis of ERM on individual groups remains elusive. By formulating the group imbalance problem with the Gaussian Mixture Model, this paper quantifies the impact of individual groups on the sample complexity, the convergence rate, and the average and group-level testing performance. Although our theoretical framework is centered on binary classification using a one-hidden-layer neural network, to the best of our knowledge, we provide the first theoretical analysis of the group-level generalization of ERM in addition to the commonly studied average generalization performance. Sample insights of our theoretical results include that when all group-level co-variance is in the medium regime and all mean are close to zero, the learning performance is most desirable in the sense of a small sample complexity, a fast training rate, and a high average and group-level testing accuracy. Moreover, we show that increasing the fraction of the minority group in the training data does not necessarily improve the generalization performance of the minority group. Our theoretical results are validated on both synthetic and empirical datasets, such as CelebA and CIFAR-10 in image classification.

Index Terms—Explainable machine learning, group imbalance, generalization analysis, Gaussian mixture model

I. INTRODUCTION

Training neural networks with empirical risk minimization (ERM) is a common practice to reduce the average loss of a machine learning task evaluated on a dataset. However, recent findings [1], [2], [3], [4], [5] have shown empirical evidence about a critical challenge of ERM, known as *group imbalance*, where a well-trained model that has high average accuracy may have significant errors on the minority group that infrequently appears in the data. Moreover, the group attributes that determine the majority and minority groups are usually hidden and unknown during the training. The training set can be augmented by data augmentation methods [6] with varying performance, such as cropping and rotation [7], noise injection

[8], and generative adversarial network (GAN)-based methods [9].

As ERM is a prominent method and enjoys great empirical success, it is important to characterize the impact of ERM on group imbalance theoretically. However, the technical difficulty of analyzing the nonconvex ERM problem of neural networks results from the concatenation of nonlinear functions across layers, and the existing generalization analyses of ERM often require strong assumptions and focus on the average performance of all data. For example, the neural tangent kernel type of analysis [10], [11], [12], [13], [14], [15], [16] linearizes the neural network around the random initialization. The generalization results are independent of the feature distribution and cannot be exploited to characterize the impact of individual groups. Ref. [14] provides the sample complexity analysis when the data comes from the mixtures of well-separated distributions but still cannot characterize the learning performance of individual groups. In another line of works [17], [18], [19], [20], [21], [22], [23], [24], [25], people make data assumptions that the labels are determined merely by some input features and are irrelevant to other features or model parameters. The generalization analysis characterizes how the neurons learn important features. Our work follows the line of works [26], [27], [28], [29], [30], where the label of each data is generated by both the input distribution and the ground-truth model so that group imbalance can be characterized.

Contribution: To the best of our knowledge, *this paper provides the first theoretical characterization of both the average and group-level generalization of a one-hidden-layer neural network trained by ERM on data generated from a mixture of distributions*. This paper considers the binary classification problem with the cross entropy loss function, with training data generated by a ground-truth neural network with known architecture and unknown weights. The optimization problem is challenging due to a high non-convexity from the multi-neuron architecture and the non-linear sigmoid activation.

Assuming the features follow a Gaussian Mixture Model (GMM), where samples of each group are generated from a Gaussian distribution with an arbitrary mean vector and co-variance matrix, this paper quantifies the impact of individual groups on the sample complexity, the training convergence rate, and the average and group-level test error. The training

The authors Hongkang Li and Dr. Meng Wang are with the Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute. Email: {lih35, wangm7}@rpi.edu. The author Dr. Shuai Zhang is with New Jersey Institute of Technology. Email: sz457@njit.edu. The authors Yihua Zhang and Dr. Sijia Liu are with the Dept. of Computer Science and Engineering, Michigan State University. Email: {zhan1908, liusijia5}@msu.edu. The author Dr. Pin-Yu Chen is with IBM Research. Email: pin-yu.chen@ibm.com.

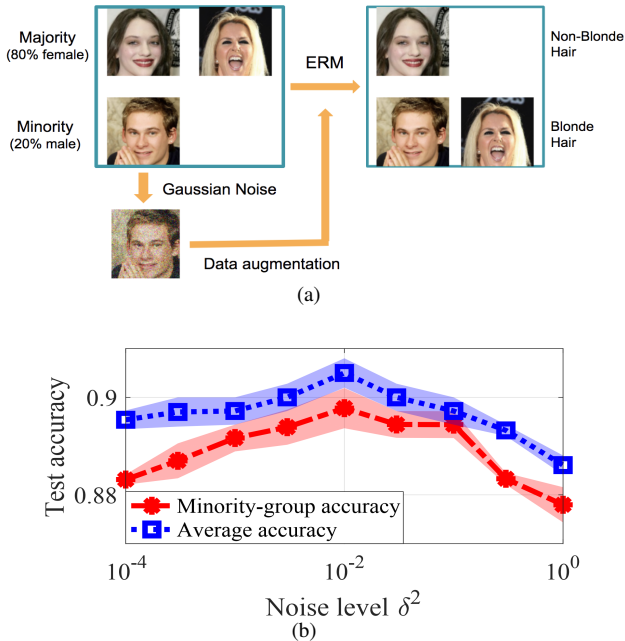


Fig. 1: Group imbalance experiment. (a) Binary classification on CelebA dataset using Gaussian augmentation to control the minority group co-variance. (b) Test accuracy against the augmented noise level.

algorithm is the gradient descent following a tensor initialization and converges linearly. Our key results include

(1) *Medium-range group-level co-variance enhances the learning performance.* When a group-level co-variance deviates from the medium regime, the learning performance degrades in terms of higher sample complexity, slower convergence in training, and worse average and group-level generalization performance. As shown in Figure 1(a), we introduce Gaussian augmentation to control the co-variance level of the minority group in the CelebA dataset [31]. The learned model achieves the highest test accuracy when the co-variance is at the medium level, see Figure 1(b). Another implication is that the diverse performance of different data augmentation methods might partially result from the different group-level co-variance introduced by these methods. Furthermore, although our setup does not directly model the batch normalization approach [32] that modifies the mean and variance in each layer to achieve fast and stable convergence, our result provides a theoretical insight that co-variance indeed affects the learning performance.

(2) *Group-level mean shifts from zero hurt the learning performance.* When a group-level mean deviates from zero, the sample complexity increases, the algorithm converges slower, and both the average and group-level test error increases. Thus, the learning performance is improved if each distribution is zero-mean. This paper provides a similar theoretical insight to practical tricks such as whitening [33], subgroup shift [34], [35], population shift [36], [37] and the pre-processing of making data zero-mean [38], that data mean affects the learning performance.

(3) *Increasing the fraction of the minority group in the*

training data does not always improve its generalization performance. The generalization performance is also affected by the mean and co-variance of individual groups. In fact, increasing the fraction of the minority group in the training data can have a completely opposite impact in different datasets.

II. BACKGROUND AND RELATED WORK

Improving the minority-group performance with known group attributes. With known group attributes, distributionally robust optimization (DRO) [4] minimizes the worst-group training loss instead of solving ERM. DRO is more computationally expensive than ERM and does not always outperform ERM in the minority-group test error. Spurious correlations [3] can be viewed as one reason of group imbalance, where strong associations between labels and irrelevant features exist in training samples. Different from the approaches that address spurious correlations, such as down-sampling the majority [39], [40], up-weight the minority group [41], and removing spurious features [42], [43], this paper does not require the special model of spurious correlations and any group attribute information.

Imbalance learning and long-tailed learning focus on learning from imbalanced data with a long-tailed distribution, which means that a few classes of the data make up the majority of the dataset, while the majority of classes have little data samples [44], [45], [46], [47], [48], [49], [50], [51], [52]. Some works [45], [52] claimed that naively increasing the number of the minority does not always improve the generalization. Therefore, some recent works develop novel oversampling and data augmentation methods [49], [48], [51] that can promote the minority fraction by generating diverse and context-rich minority data. However, there are very limited theoretical explanations of how these techniques affect the generalization.

Generalization performance with the standard Gaussian input for one-hidden-layer neural networks. [53], [54], [55], [56] consider infinite training samples. [26] characterize the sample complexity of fully connected neural networks with smooth activation functions. [57], [58], [30] extend to the non-smooth ReLU activation for fully-connected and convolutional neural networks, respectively. [28] analyzes the cross entropy loss function for binary classification problems. [27] analyzes the generalizability of graph neural networks for both regression and binary classification problems. One-hidden-layer case of neural network pruning and self-training are also studied in [59] and [29], respectively.

Theoretical characterization of learning performance from other input distributions for one-hidden-layer neural networks. [60] analyzes the training loss with a single Gaussian with an arbitrary co-variance. [61] quantifies the SGD evolution trained on the Gaussian mixture model. When the hidden layer only contains one neuron, [62] analyzes rotationally invariant distributions. With an infinite number of neurons and an infinite input dimension, [63] analyzes the generalization error based on the mean-field analysis for distributions like Gaussian Mixture with the same mean. [64]

considers inputs with low-dimensional structures. No sample complexity is provided in all these works.

Notations: \mathbf{Z} is a matrix with $Z_{i,j}$ as the (i, j) -th entry. \mathbf{z} is a vector with z_i as the i -th entry. $[K]$ denotes the set including integers from 1 to K . \mathbf{I}_d and \mathbf{e}_i represent the identity matrix in $\mathbb{R}^{d \times d}$ and the i -th standard basis vector, respectively. $\delta_i(\mathbf{Z})$ denotes the i -th largest singular value of \mathbf{Z} . The matrix norm $\|\mathbf{Z}\| = \delta_1(\mathbf{Z})$. $f(x) = O(g(x))$ (or $\Omega(g(x))$, $\Theta(g(x))$) means that $f(x)$ increases at most, at least, or in the order of $g(x)$, respectively.

III. PROBLEM FORMULATION AND ALGORITHM

We consider the classification problem with an unbalanced dataset using fully connected neural networks over n independent training examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from a data distribution. The learning algorithm is to minimize the empirical risk function via gradient descent (GD). In what follows, we will present the data model and neural network model considered in this paper.

Data Model. Let $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ denote the input feature and label, respectively. We consider an unbalanced dataset that consists of L ($L \geq 2$) groups of data, where the feature \mathbf{x} in the group l ($l \in [L]$) is drawn from a multi-variate Gaussian distribution with mean $\boldsymbol{\mu}_l \in \mathbb{R}^d$, and covariance $\boldsymbol{\Sigma}_l \in \mathbb{R}^{d \times d}$. Specifically, \mathbf{x} follows the Gaussian mixture model (GMM) [65], [66], [67], [68], denoted as $\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ ¹. $\lambda_l \in (0, 1)$ is the probability of sampling from distribution- l and represents the expected fraction of group- l data. $\sum_{l=1}^L \lambda_l = 1$. Group l is defined as a minority group if λ_l is less than $1/L$. We use $\Psi = \{\lambda_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \forall l\}$ to denote all parameters of the mixture model². We consider binary classification with label y generated by a ground-truth neural network with unknown weights $\mathbf{W}^* = [\mathbf{w}_1^*, \dots, \mathbf{w}_K^*] \in \mathbb{R}^{d \times K}$ and sigmoid activation³. function $\phi(x) = \frac{1}{1 + \exp(-x)}$, where⁴

$$\mathbb{P}(y = 1 | \mathbf{x}) = H(\mathbf{W}^*, \mathbf{x}) := \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*\top} \mathbf{x}). \quad (1)$$

Learning model. Learning is performed over a neural network that has the same architecture as in (1), which is a one-hidden-layer fully connected neural network⁵ with its weights denoted by $\mathbf{W} \in \mathbb{R}^{d \times K}$. Given n training samples $\{\mathbf{x}_i, y_i\}_{i=1}^n$ where \mathbf{x}_i follows the GMM model, and y_i is from

¹We consider this data model inspired by existing works on group imbalance and practical datasets. Details can be found in Appendix F.

²In practice, Ψ can be estimated by the EM algorithm [69] and the moment-based method [66]. The EM algorithm returns model parameters within Euclidean distance $O((\frac{d}{n})^{\frac{1}{2}})$ when the number of mixture components L is known. When L is unknown, one usually over-specifies an estimate $\bar{L} > L$, then the estimation error by the EM algorithm scales as $O((\frac{d}{n})^{\frac{1}{4}})$. Please refer to [70], [71], [72] for details.

³The results can be generalized to any activation function ϕ with bounded ϕ , ϕ' and ϕ'' , where ϕ' is even. Examples include tanh and erf.

⁴Our data model is reduced to logistic regression in the special case that $K = 1$. We mainly study the more challenging case when $K > 1$, because the learning problem becomes highly non-convex when there are multiple neurons in the network.

⁵All the weights in the second layer are assumed to be fixed to facilitate the analysis. This is a standard assumption in theoretical generalization analysis [57], [28], [27].

(1), we aim to find the model weights via solving the empirical risk minimization (ERM), where $f_n(\mathbf{W})$ is the empirical risk,

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times K}} f_n(\mathbf{W}) := \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{W}; \mathbf{x}_i, y_i), \quad (2)$$

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

$$\begin{aligned} \ell(\mathbf{W}; \mathbf{x}_i, y_i) &= -y_i \cdot \log(H(\mathbf{W}, \mathbf{x}_i)) \\ &\quad - (1 - y_i) \cdot \log(1 - H(\mathbf{W}, \mathbf{x}_i)). \end{aligned} \quad (3)$$

Note that for any permutation matrix \mathbf{P} , $\mathbf{W}\mathbf{P}$ corresponds permuting neurons of a network with weights \mathbf{W} . Therefore, $H(\mathbf{W}, \mathbf{x}) = H(\mathbf{W}\mathbf{P}, \mathbf{x})$, and $f_n(\mathbf{W}\mathbf{P}) = f_n(\mathbf{W})$. The estimation is considered successful if one finds any column permutation of \mathbf{W}^* .

The average generalization performance of a learned model \mathbf{W} is evaluated by the average risk

$$\bar{f}(\mathbf{W}) = \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \ell(\mathbf{W}; \mathbf{x}_i, y_i), \quad (4)$$

and the generalization performance on group l is evaluated by the group- l risk

$$\bar{f}_l(\mathbf{W}) = \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \ell(\mathbf{W}; \mathbf{x}_i, y_i). \quad (5)$$

Training Algorithm. Our algorithm starts from an initialization $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$ computed based on the tensor initialization method (Subroutine 1 in Appendix) and then updates the iterates \mathbf{W}_t using gradient descent with the step size⁶ η_0 . The computational complexity of tensor initialization is $O(Knd)$. The per-iteration complexity of the gradient step is $O(Knd)$. We defer the details of Algorithm 1 in Appendix.

Algorithm 1 Our ERM learning algorithm

- 1: **Input:** Training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the step size $\eta_0 = O\left(\left(\sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right)^{-1}\right)$, the total number of iterations T
- 2: **Initialization:** $\mathbf{W}_0 \leftarrow$ Tensor initialization method via Subroutine 1
- 3: **Gradient Descent:** for $t = 0, 1, \dots, T - 1$

$$\begin{aligned} \mathbf{W}_{t+1} &= \mathbf{W}_t - \eta_0 \cdot \frac{1}{n} \sum_{i=1}^n (\nabla \ell(\mathbf{W}, \mathbf{x}_i, y_i) + \nu_i) \\ &= \mathbf{W}_t - \eta_0 \left(\nabla f_n(\mathbf{W}) + \frac{1}{n} \sum_{i=1}^n \nu_i \right) \end{aligned} \quad (6)$$

- 4: **Output:** \mathbf{W}_T
-

IV. MAIN THEORETICAL RESULTS

We will formally present our main theory below, and the insights are summarized in Section IV-A. For the convenience of presentation, some quantities are defined here, and all of them can be viewed as constant. Define $\sigma_{\max} = \max_{l \in [L]} \{\|\boldsymbol{\Sigma}_l\|^{\frac{1}{2}}\}$, $\sigma_{\min} = \min_{l \in [L]} \{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\}$. Let $\tau = \sigma_{\max}/\sigma_{\min}$. We assume $\tau = \Theta(1)$, indicating that σ_{\max} and σ_{\min} are in

⁶Algorithm 1 employs a constant step size. One can potentially speed up the convergence, i.e., reduce η , by using a variable step size. We leave the corresponding theoretical analysis for future work.

the same order⁷. Let $\delta_i(\mathbf{W}^*)$ denote the i -th largest singular value of \mathbf{W}^* . Let $\kappa = \frac{\delta_1(\mathbf{W}^*)}{\delta_K(\mathbf{W}^*)}$, and define $\eta = \prod_{i=1}^K (\delta_i(\mathbf{W}^*)/\delta_K(\mathbf{W}^*))$.

Theorem 1. *There exist $\epsilon_0 \in (0, \frac{1}{4})$ and positive value functions $\mathcal{B}(\Psi)$ (sample complexity parameter), $q(\Psi)$ (convergence rate parameter), and $\mathcal{E}_w(\Psi)$, $\mathcal{E}(\Psi)$, $\mathcal{E}_l(\Psi)$ (generalization parameters) such that as long as the sample size n satisfies*

$$n \geq n_{sc} := \text{poly}(\epsilon_0^{-1}, \kappa, \eta, \tau, K, \delta_1(\mathbf{W}^*)) \mathcal{B}(\Psi) d \log^2 d, \quad (7)$$

we have that with probability at least $1 - d^{-10}$, the iterates $\{\mathbf{W}_t\}_{t=1}^T$ returned by Algorithm 1 with step size $\eta_0 = O\left(\left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|^{\frac{1}{2}})^2\right)^{-1}\right)$ converge linearly with a statistical error to a critical point $\widehat{\mathbf{W}}_n$ with the rate of convergence v , i.e.,

$$\begin{aligned} \|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F &\leq v(\Psi)^t \|\mathbf{W}_0 - \widehat{\mathbf{W}}_n\|_F \\ &\quad + \frac{\eta_0 \xi}{1 - v(\Psi)} \sqrt{dK \log n/n}, \end{aligned} \quad (8)$$

$$v(\Psi) = 1 - K^{-2} q(\Psi), \quad (9)$$

where $\xi \geq 0$ is the upper bound of the entry-wise additive noise in the gradient computation.

Moreover, there exists a permutation matrix \mathbf{P}^* such that

$$\begin{aligned} \|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}^*\|_F &\leq \mathcal{E}_w(\Psi) \cdot \text{poly}(\kappa, \eta, \tau, \delta_1(\mathbf{W}^*)) \\ &\quad \cdot \Theta\left(K^{\frac{5}{2}}(1 + \xi) \cdot \sqrt{d \log n/n}\right). \end{aligned} \quad (10)$$

The average population risk \bar{f} and the group- l risk \bar{f}_l satisfy

$$\begin{aligned} \bar{f} &\leq \mathcal{E}(\Psi) \cdot \text{poly}(\kappa, \eta, \tau, \delta_1(\mathbf{W}^*)) \\ &\quad \cdot \Theta\left(K^{\frac{5}{2}}(1 + \xi) \cdot \sqrt{d \log n/n}\right) \end{aligned} \quad (11)$$

$$\begin{aligned} \bar{f}_l &\leq \mathcal{E}_l(\Psi) \cdot \text{poly}(\kappa, \eta, \tau, \delta_1(\mathbf{W}^*)) \\ &\quad \cdot \Theta\left(K^{\frac{5}{2}}(1 + \xi) \cdot \sqrt{d \log n/n}\right) \end{aligned} \quad (12)$$

The closed-form expressions of \mathcal{B} , q , \mathcal{E}_w , \mathcal{E} , and \mathcal{E}_l are in Section D of the supplementary material and skipped here. The quantitative impact of the GMM model parameters Ψ on the learning performance varies in different regimes and can be derived from Theorem 1. The following corollary summarizes the impact of Ψ on the learning performance in some sample regimes.

Corollary 1. *When we vary one parameter of group l for any $l \in [L]$ of the GMM model Ψ and fix all the others, the learning performance degrades in the sense that the sample complexity n_{sc} , the convergence rate v , $\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}^*\|_F$, average risk \bar{f} and group- l risk \bar{f}_l all increase (details summarized in Table 1), as long as any of the following conditions happens,*

(i) $\|\boldsymbol{\Sigma}_l\|$ approaches 0; (ii) $\|\boldsymbol{\Sigma}_l\|$ increases from some constant; (iii) $\|\boldsymbol{\mu}_l\|$ increases from 0,

(iv) λ_l decreases, provided that $\|\boldsymbol{\Sigma}_l\| = \sigma_{\min}^2$, i.e., group l has the smallest group-level co-variance, where $\|\boldsymbol{\Sigma}_j\|$ are all constants, and $\|\boldsymbol{\mu}_i\| = \|\boldsymbol{\mu}_j\|$ for all $i, j \in [L]$.

(v) λ_l increases, provided that $\|\boldsymbol{\Sigma}_l\| = \sigma_{\max}^2$, i.e., group l has the largest group-level co-variance, where $\|\boldsymbol{\Sigma}_j\|$ are all constants, and $\|\boldsymbol{\mu}_i\| = \|\boldsymbol{\mu}_j\|$ for all $i, j \in [L]$.

To the best of our knowledge, Theorem 1 provides the first characterization of the sample complexity, learning rate, and generalization performance under the Gaussian mixture model. It also firstly characterizes the per-group generalization performance in addition to the average generalization.

A. Theoretical Insights

We summarize the crucial implications of Theorem 1 and Corollary 1 as follows.

(P1). Training convergence and generalization guarantee. The iterates \mathbf{W}_t converge to a critical point $\widehat{\mathbf{W}}_n$ linearly, and the distance between $\widehat{\mathbf{W}}_n$ and $\mathbf{W}^* \mathbf{P}^*$ is $O(\sqrt{d \log n/n})$ for a certain permutation matrix \mathbf{P}^* . When the computed gradients contain noise, there is an additional error term of $O(\xi \sqrt{d \log n/n})$, where ξ is the noise level ($\xi = 0$ for noiseless case). Moreover, the average risk of all groups and the risk of each individual group are both $O((1 + \xi) \sqrt{d \log n/n})$.

(P2). Sample complexity. For a given GMM, the sample complexity is $\Theta(d \log^2 d)$, where d is the feature dimension. This result is in the same order as the sample complexity for the standard Gaussian input in [28] and [26]. Our bound is almost order-wise optimal with respect to d because the degree of freedom is dK . The additional multiplier of $\log^2 d$ results from the concentration bound in the proof technique. We focus on the dependence on the feature dimension d and treat the network width K as constant. The sample complexity in [28] and [26] is also $d \cdot \text{poly}(K, \log d)$.

(P3). Learning performance is improved at a medium regime of group-level co-variance. On the one hand, when $\|\boldsymbol{\Sigma}_l\|$ is $\Omega(1)$, the learning performance degrades as $\|\boldsymbol{\Sigma}_l\|$ increases in the sense that the sample complexity n_{sc} , the convergence rate v , the estimation error of \mathbf{W}^* , the average risk \bar{f} , and the group- l risk \bar{f}_l all increase. This is due to the saturation of the loss and gradient when the samples have a large magnitude. On the other hand, when $\|\boldsymbol{\Sigma}_l\|$ is $o(1)$, the learning performance also degrades when $\|\boldsymbol{\Sigma}_l\|$ approaches zero. The intuition is that in this regime, the input data are concentrated on a few vectors, and the optimization problem does not have a benign landscape.

(P4). Increasing the fraction of the minority group data does not always improve the generalization, while the performance also depends on the mean and co-variance of individual groups. Take $\|\boldsymbol{\Sigma}_j\| = \Theta(1)$ for all group j , and $\|\boldsymbol{\mu}_j\|$ is the same for all j as an example (columns 5 and 6 of Table 1). When $\|\boldsymbol{\Sigma}_l\|$ is the smallest among all groups, increasing λ_l improves the learning performance. When $\|\boldsymbol{\Sigma}_l\|$ is the largest among all groups, increasing λ_l actually degrades the performance. The intuition is that from (P3), the learning performance is enhanced at a medium regime of group-level co-variance. Thus, increasing the fraction of a group with a medium level of co-variance improves the performance, while

⁷Note that it is a very mild assumption that σ_{\min} is not very close to zero, or equivalently, $\tau = \Theta(1)$. We verify this in Appendix G.

⁷ $\text{poly}(\|\boldsymbol{\mu}_l\|)$ is $\|\boldsymbol{\mu}_l\|^4$ for $\|\boldsymbol{\mu}_l\| \leq 1$; $\|\boldsymbol{\mu}_l\|^{12}$ for $\|\boldsymbol{\mu}_l\| > 1$.

TABLE I: Impact of GMM parameters on the learning performance in sample regimes

	Σ_l changes		μ_l changes	λ_l changes, constant $\ \Sigma_j\ $'s, equal $\ \mu_j\ $'s	
	$\ \Sigma_l\ = o(1)$	$\ \Sigma_l\ = \Omega(1)$		if $\ \Sigma_l\ = \sigma_{\min}^2$	if $\ \Sigma_l\ = \sigma_{\max}^2$
$\mathcal{B}(\Psi)$, sample complexity n_{sc}	$O(\ \Sigma_l\ ^{-3})$	$O\ \Sigma_l\ ^3$	$O(\text{poly}(\ \mu_l\))^8$	$O(\frac{1}{(1+\lambda_l)^2})$	$O(1) - \frac{\Theta(1)}{(1+\lambda_l)^2}$
convergence rate $v(\Psi) \propto -q(\Psi)$	$1 - \Theta(\ \Sigma_l\ ^3)$	$1 - \Theta(\frac{1}{1+\ \Sigma_l\ })$	$1 - \Theta(\frac{1}{\ \mu_l\ ^2+1})$	$\Theta(\frac{1}{1+\lambda_l})$	$1 - \Theta(\frac{1}{1+\lambda_l})$
$\mathcal{E}_w(\Psi)$, $\ \widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\ _F$	$O(1) - \Theta(\ \Sigma_l\ ^3)$	$O(\sqrt{\ \Sigma_l\ })$	$O(1 + \ \mu_l\)$	$O(\frac{1}{1+\sqrt{\lambda_l}})$	$O(1 + \sqrt{\lambda_l})$
$\mathcal{E}(\Psi)$, average risk \bar{f}	$O(1) - \Theta(\ \Sigma_l\ ^3)$	$O(\ \Sigma_l\)$	$O(1 + \ \mu_l\ ^2)$	$O(\frac{1}{1+\lambda_l})$	$O(1) - \frac{\Theta(1)}{1+\lambda_l}$
$\mathcal{E}_l(\Psi)$, group- l risk \bar{f}_l	$O(1) - \Theta(\ \Sigma_l\ ^3)$	$O(\ \Sigma_l\)$	$O(1 + \ \mu_l\ ^2)$	$O(\frac{1}{1+\sqrt{\lambda_l}})$	$O(1 + \sqrt{\lambda_l})$

increasing the fraction of a group with large co-variance degrades the learning performance. Similarly, when augmenting the training data, an argumentation method that introduces medium variance could improve the learning performance, while an argumentation method that introduces a significant level of variance could hurt the learning performance.

(P5). Group-level mean shifts from zero degrade the learning performance. The learning performance degrades as $\|\mu_l\|$ increases. An intuitive explanation of the degradation is that some training samples have a significant large magnitude such that the sigmoid function saturates.

B. Proof Idea and Technical Novelty

1) *Proof Idea:* Different from the analysis of logistic regression for generalized linear models, our paper deals with more technical challenges of nonconvex optimization due to the multi-neuron architecture, the GMM model, and a more complicated activation and loss. The establishment of Theorem 1 consists of three key lemmas.

Lemma 1. (informal version) *As long as the number of training samples is larger than $\Omega(dK^5 \log^2 d)$, the empirical risk function is strongly convex in the neighborhood of \mathbf{W}^* (or a permutation of \mathbf{W}^*). The size of the convex region is characterized by the Gaussian mixture distribution.*

The main proof idea of Lemma 1 is to show that the nonconvex empirical risk $f_n(\mathbf{W})$ in a small neighborhood around \mathbf{W}^* (or any permutation $\mathbf{W}^* \mathbf{P}$) is almost convex with a sufficiently large n . The difficulty is to find a positive lower bound of the smallest singular value of $\nabla^2 \bar{f}(\mathbf{W})$, which should also be a function of the GMM. Then, we can obtain $\nabla^2 f_n(\mathbf{W})$ from $\nabla^2 \bar{f}(\mathbf{W})$ by concentration inequalities.

Lemma 2. (informal version) *If initialized in the convex region, the gradient descent algorithm converges linearly to a critical point $\widehat{\mathbf{W}}_n$, which is close to \mathbf{W}^* (or any permutation of \mathbf{W}^*), and the distance is diminishing as the number of training samples increases.*

Given the locally strong convexity, Lemma 2 provides the linear convergence to a critical point. The convergence rate is determined by the GMM.

Lemma 3. (informal version) *Tensor Initialization Method initializes $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$ around \mathbf{W}^* (or a permutation of \mathbf{W}^*).*

The idea of tensor initialization is to first find quantities (see \mathbf{Q}_j in Definition 1) in the supplementary material) which are proven to be functions of tensors of \mathbf{w}_i^* . Then the method approximates these quantities numerically using training samples and then applies the tensor decomposition method on the estimated quantities to obtain \mathbf{W}_0 , which is an estimation of \mathbf{W}^* .

Combining the above three lemmas together, one can derive the required sample complexity and the upper bound of \bar{f} and \bar{f}_l in (7), (11), and (12), respectively. The idea is first to compute the sample complexity bound such that the tensor initialization method initializes \mathbf{W}_0 in the local convex region by Lemma 3. Then the final sample complexity is obtained by comparing two sample complexities from Lemma 1 and 3.

By further looking into the order of the terms $\mathcal{B}(\Psi)$, $v(\Psi)$, $\mathcal{E}(\Psi)$, $\mathcal{E}_w(\Psi)$, and $\mathcal{E}_l(\Psi)$ in several cases of Ψ , Theorem 1 leads to Corollary 1. To be more specific, we only vary parameters Σ_l , or μ_l , or λ_l following the cases in Table I, while fixing all other parameters of Ψ . We apply the Taylor expansion to approximate the terms and derive error bounds with the Lipschitz smoothness of the loss function.

2) *Technical Novelty:* Our algorithmic and analytical framework is built upon some recent works on the generalization analysis of one-hidden-layer neural networks, see, e.g., [26], [57], [28], [27], [59], which assume that \mathbf{x}_i follows the standard Gaussian distribution and cannot be directly extended to GMM. This paper makes new technical contributions from the following aspects.

First, we characterize the local convex region near \mathbf{W}^* for the GMM model. To be more specific, we explicitly characterize the positive lower bound of the smallest singular value of $\nabla^2 \bar{f}(\mathbf{W})$ with respect to Ψ , while existing results either only hold for standard Gaussian data [26], [28], [59], [29], or can only show $\nabla^2 \bar{f}(\mathbf{W})$ is positive definite regardless the impact of Ψ [10].

Second, new tools, including matrix concentration bounds are developed to explicitly quantify the impact of Ψ on the sample complexity.

Third, we investigate and provide the order of the bound

for sample complexity, convergence rate, generalization error, average risk, and group- l risk in terms of Ψ for the first time in the line of research of model estimation [26], [28], [27], [59], [29], which is also a novel result for the case of Gaussian inputs.

Fourth, we design and analyze new tensors for the mixture model to initialize properly, while the previous tensor methods in [26], [57], [28], [27] utilize the rotation invariant property that only holds for zero mean Gaussian.

V. NUMERICAL EXPERIMENTS

A. Experiments on Synthetic datasets

We first verify the theoretical bounds in Theorem 1 on synthetic data. Each entry of $\mathbf{W}^* \in \mathbb{R}^{d \times K}$ is generated from $\mathcal{N}(0, 1)$. The training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ is generated using the GMM model and (1). If not otherwise specified, $L = 2$, $d = 5$, and $K = 3^9$. To reduce the computational time, we randomly initialize near \mathbf{W}^* instead of computing the tensor initialization¹⁰.

Sample complexity. We first study the impact of d on the sample complexity. Let $\boldsymbol{\mu}_1 = \mathbf{1}$ in \mathbb{R}^d and let $\boldsymbol{\mu}_2 = \mathbf{0}$. Let $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$. $\lambda_1 = \lambda_2 = 0.5$. We randomly initialize M times and let $\widehat{\mathbf{W}}_n^{(m)}$ denote the output of Algorithm 1 in the m th trail. Let $\bar{\mathbf{W}}_n$ denote the mean values of all $\widehat{\mathbf{W}}_n^{(m)}$, and let $V_W = \sqrt{\sum_{m=1}^M \|\widehat{\mathbf{w}}_n^m - \bar{\mathbf{W}}_n\|^2 / M}$ denote the variance. An experiment is successful if $V_W \leq 10^{-3}$ and fails otherwise. M is set to 20. For each pair of d and n , 20 independent sets of \mathbf{W}^* and the corresponding training samples are generated. Figure 2 shows the success rate of these independent experiments. A black block means that all the experiments fail. A white block means that they all succeed. The sample complexity is indeed almost linear in d , as predicted by (7).

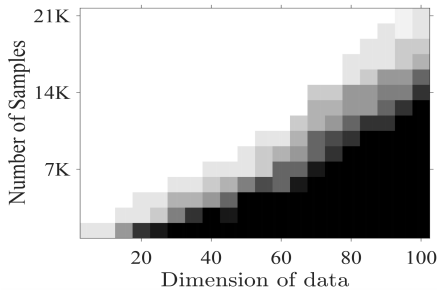


Fig. 2: The sample complexity when the feature dimension changes

We next study the impact on the sample complexity of the GMM model. In Figure 3 (a), $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$, and let $\boldsymbol{\mu}_1 = \mu \cdot \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$. $\|\boldsymbol{\mu}_1\|$ varies from 0 to 5. Figure 3(a)

⁹Like [26], [57], [28], we consider a small-sized network in synthetic experiments to reduce the computational time, especially for computing the sample complexity in Figure 3. Our results hold for large networks too.

¹⁰The existing methods based on tensor initialization all use random initialization in synthetic experiments to reduce the computational time. See [28], [57], [27], [29] as examples. We compare tensor initialization and local random initialization numerically in Section B of the supplementary material and show that they have the same performance.

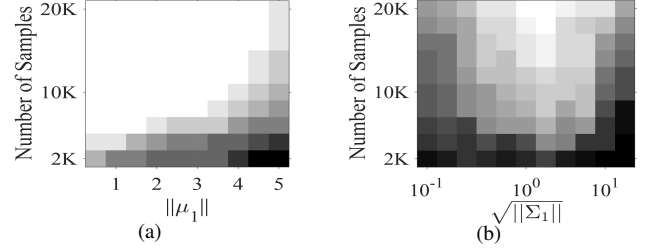


Fig. 3: The sample complexity (a) when one mean changes, (b) when one co-variance changes.

shows that when the mean increases, the sample complexity increases. In Figure 3 (b), we fix $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$, and let $\boldsymbol{\Sigma}_1 = \sigma^2 \mathbf{I}$ and $\boldsymbol{\Sigma}_2 = \mathbf{I}$. σ varies from 10^{-1} to 10^1 . The sample complexity increases both when $\|\boldsymbol{\Sigma}_1\|$ increases and when $\|\boldsymbol{\Sigma}_1\|$ approaches zero. All results match predictions in Corollary 1.

Convergence analysis. We next study the convergence rate of Algorithm 1. Figure 4(a) shows the impact of $\|\boldsymbol{\mu}_l\|$. $\lambda_1 = \lambda_2 = 0.5$, $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = C \cdot \mathbf{1}$ for a positive C , and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Lambda}^\top \mathbf{D} \boldsymbol{\Lambda}$. Here $\boldsymbol{\Lambda}$ is generated by computing the left-singular vectors of a $d \times d$ random matrix from the Gaussian distribution. $\mathbf{D} = \text{diag}(1, 1.1, 1.2, 1.3, 1.4)$. $n = 1 \times 10^4$. Algorithm 1 always converges linearly when $\|\boldsymbol{\mu}_l\|$ changes. Moreover, as $\|\boldsymbol{\mu}_1\|$ increases, Algorithm 1 converges slower. Figure 4 (b) shows the impact of the variance of the Gaussian mixture model. $\lambda_1 = \lambda_2 = 0.5$, $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = \sigma^2 \cdot \boldsymbol{\Lambda}^\top \mathbf{D} \boldsymbol{\Lambda}$. $n = 5 \times 10^4$. We change $\|\boldsymbol{\Sigma}\|$ by changing σ . Among the values we test, Algorithm 1 converges fastest when $\|\boldsymbol{\Sigma}\| = 1$. The convergence rate slows down when $\|\boldsymbol{\Sigma}\|$ increases or decreases from 1. All results are consistent with the predictions in Corollary 1. We then study the impact of K on the convergence rate. $\lambda_1 = \lambda_2 = 0.5$, $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$. Figure 5 (a) shows that, as predicted by (9), the convergence rate is linear in $-1/K^2$.

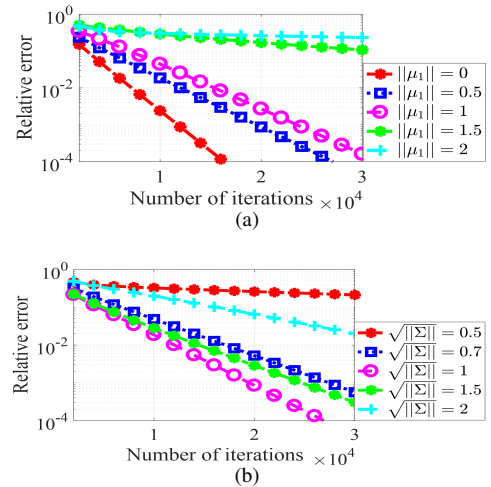


Fig. 4: (a) The convergence rate with different $\boldsymbol{\mu}_1$. (b) The convergence rate with different $\boldsymbol{\Sigma}$. (c) Convergence rate when the number of neurons K changes.

Average and group-level generalization performance.

The distance between $\widehat{\mathbf{W}}_n$ returned by Algorithm 1 and \mathbf{W}^* is measured by $\|\widehat{\mathbf{W}}_n - \mathbf{W}^*\|_F$. n ranges from 2×10^3 to 6×10^4 . $\Sigma_1 = \Sigma_2 = 9\mathbf{I}$, $\mu_1 = \mathbf{1}$, $\mu_2 = -\mathbf{1}$. Each point in Figure 5 (b) is averaged over 20 experiments of different \mathbf{W}^* and training set. The error is indeed linear in $\sqrt{\log(n)/n}$, as predicted by (8).

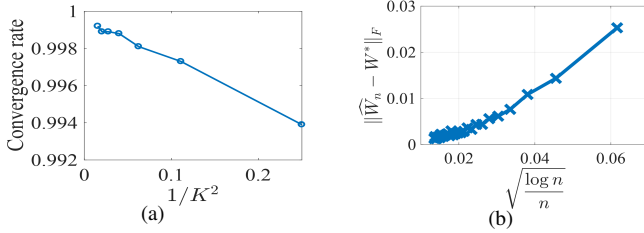


Fig. 5: (a) Convergence rate when the number of neurons K changes. (b) The relative error of the learned model when n changes.

We evaluate the impact of one mean/co-variance of the minority group on the generalization. $n = 2 \times 10^4$. Let $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, $\mu_1 = 2 \cdot \mathbf{1}$, $\Sigma_1 = \mathbf{I}$. First, we let $\mu_2 = (\mu_2 - 2) \cdot \mathbf{1}$ and $\Sigma_2 = \mathbf{I}$. Figure 6 (b) shows that both the average risk and the group-2 risk increase as μ_2 increases, consistent with (P5). Then we set $\mu_2 = -2 \cdot \mathbf{1}$, $\Sigma_2 = \sigma_2^2 \cdot \mathbf{I}$. Figure 6 (a) indicates that both the average and the group-2 risk will first decrease and then increase as $\|\Sigma_2\|_2$ increases, consistent with (P3).

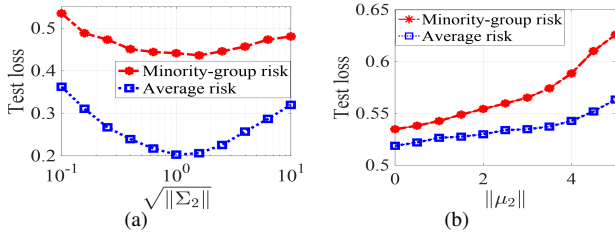


Fig. 6: (a) The cross-entropy test loss when the co-variance of the minority group changes. (b) The cross-entropy test loss when the mean of the minority group changes.

Next, we study the impact of increasing the fraction of the minority group. $\mu_1 = \mu_2 = 0$. Let group 2 be the minority group. In Figure 7 (a), $\Sigma_1 = 10 \cdot \mathbf{I}$ and $\Sigma_2 = \mathbf{I}$, the minority group has a smaller level of co-variance. Then when λ_2 increases from 0 to 0.5, both the average and group-2 risk decrease. In Figure 7 (b), $\Sigma_1 = \mathbf{I}$ and $\Sigma_2 = 10 \cdot \mathbf{I}$, and the minority group has a higher-level of co-variance. Then when λ_2 increases from 0 to 0.3, both the average and group-2 risk increase. As predicted by insight (P4), increasing λ_2 does not necessarily improve the generalization of group 2.

B. Image classification on dataset CelebA

We choose the attribute “blonde hair” as the binary classification label. ResNet 9 [73] is selected to be the learning model

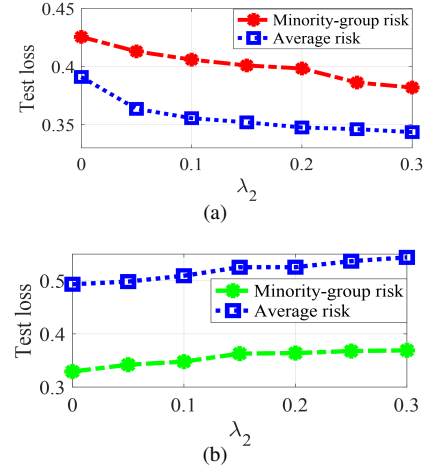


Fig. 7: The test loss (cross entropy loss) of synthetic data with different λ_2 values. (a) Group 2 has a smaller level of co-variance. (b) Group 2 has a larger level of co-variance.

here because it was applied in many simple computer vision tasks [74], [75]. To study the impact of co-variance, we pick 4000 female (majority) and 1000 male (minority) images and implement Gaussian data augmentation to create additional 300 images for the male group. Specifically, we select 300 out of 1000 male images and add i.i.d. noise drawn from $\mathcal{N}(0, \delta^2)$ to every entry. The test set includes 500 male and 500 female images. Figure 1 shows that when δ^2 increases, i.e., when the co-variance of the minority group increases, both the minority-group and average test accuracy increase first and then decrease, coinciding with our insight (P3).

Then we fix the total number of training data to be 5000 and vary the fractions of the two groups. From Figure 8(a)¹¹ and (b), we observe opposite trends if we increase the fraction of the minority group in the training data with the male being the minority and the female being the minority. The norm of covariance of the male and female group in the feature space is 5.1833 and 4.9716, respectively. This is consistent with Insight (P4). Due to space limit, our results on the CIFAR10 dataset are deferred to Section A in the supplementary material.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

This paper provides a novel theoretical framework for characterizing neural network generalization with group imbalance. The group imbalance is formulated using the Gaussian mixture model. This paper explicitly quantifies the impact of each group on the sample complexity, convergence rate, and the average and the group-level generalization. The learning performance is enhanced when the group-level covariance is at a medium regime, and the group-level mean is close to zero. Moreover, increasing the fraction of minority group does not guarantee improved group-level generalization.

¹¹In Figure 8(a), when the minority fraction is less than 0.01, the minority group distribution is almost removed from the Gaussian mixture model. Then the $O(1)$ constants in the last column of Table I have some minor changes, and the order-wise analyses do not reflect the minor fluctuations in this regime.

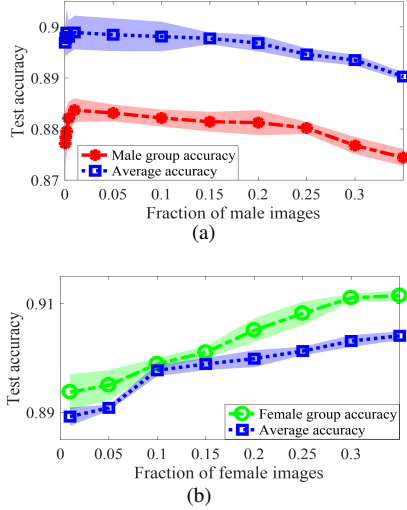


Fig. 8: The test accuracy on CelebA dataset has opposite trends when the minority group fraction increases. (a) Male group is the minority (b) Female group is the minority

One future direction is to extend the analysis to multiple-hidden-layer neural networks and multi-class classification. Because of the concatenation of nonlinear activation functions, the analysis of the landscape of the empirical risk and the design of a proper initialization is more challenging and requires the development of new tools. Another future direction is to analyze other robust training methods, such as DRO. We see no ethical or immediate negative societal consequence of our work.

APPENDIX

A. Definitions

Definition 1. (ρ -function). Let $\mathbf{z} \sim \mathcal{N}(\mathbf{u}, \mathbf{I}_d) \in \mathbb{R}^d$. Define $\alpha_q(i, \mathbf{u}, \sigma) = \mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi^q(\sigma \cdot z_i) z_i^q]$ and $\beta_q(i, \mathbf{u}, \sigma) = \mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'^2(\sigma \cdot z_i) z_i^q]$, $\forall q \in \{0, 1, 2\}$, where z_i and u_i is the i -th entry of \mathbf{z} and \mathbf{u} , respectively. Define $\rho(\mathbf{u}, \sigma)$ as

$$\rho(\mathbf{u}, \sigma) = \min_{i, j \in [d], j \neq i} \left\{ (u_j^2 + 1)(\beta_0(i, \mathbf{u}, \sigma) - \alpha_0(i, \mathbf{u}, \sigma)^2), \right. \\ \left. \beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2(i, \mathbf{u}, \sigma)^2}{u_i^2 + 1} \right\} \quad (13)$$

Definition 2. (D -function). Given the Gaussian Mixture Model and any positive integer m , define $D_m(\Psi)$ as

$$D_m(\Psi) = \sum_{l=1}^L \lambda_l \left(\frac{\|\boldsymbol{\mu}_l\|}{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}} + 1 \right)^m, \quad (14)$$

ρ -function is defined to compute the lower bound of the Hessian of the population risk with Gaussian input. D -function is a normalized parameter for the means and variances. It is lower bounded by 1. D -function is an increasing function of $\|\boldsymbol{\mu}_l\|$ and a decreasing function of σ_l .

B. Proof of Lemma 1

We first restate the formal version of Lemma 1 in the following.

Lemma 1. (Strongly local convexity) Consider the classification model with FCN (1) and the sigmoid activation function. There exists a constant C such that as long as the sample size

$$n \geq C_1 \epsilon_0^{-2} \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^2 \\ \cdot \left(\sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \right. \right. \\ \left. \left. \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \right)^{-2} d K^5 \log^2 d \right) \quad (15)$$

for some constant $C_1 > 0$, $\epsilon_0 \in (0, \frac{1}{4})$, and any fixed permutation matrix $\mathbf{P} \in \mathbb{R}^{K \times K}$ we have for all $\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$,

$$\Omega \left(\frac{1 - 2\epsilon_0}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \right. \right. \\ \left. \left. \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \right) \right) \cdot \mathbf{I}_{dK} \\ \leq \nabla^2 f_n(\mathbf{W}) \leq C_2 \sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot \mathbf{I}_{dK}$$

with probability at least $1 - d^{-10}$ for some constant $C_2 > 0$.

1) Useful lemmas: Lemmas 4, 5, 6, 7, and 8 are required for the proof.

Lemma 4.

$$\mathbb{E}_{\mathbf{x} \sim \frac{1}{2} \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2} \mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[\left(\sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i) \right)^2 \right] \\ \geq \rho(\boldsymbol{\mu}, \sigma) \|\mathbf{R}\|_F^2, \quad (17)$$

where $\rho(\boldsymbol{\mu}, \sigma)$ is defined in Definition 1 and $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_k) \in \mathbb{R}^{d \times k}$ is an arbitrary matrix.

Lemma 5. With the FCN model (1) and the Gaussian Mixture Model, for any permutation matrix \mathbf{P} , for some constant $C_{12} > 0$, we have we have

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla^2 \ell(\mathbf{W}, \mathbf{x})\| \right. \\ \left. - \nabla^2 \ell(\mathbf{W}', \mathbf{x})\| \right] / \|\mathbf{W} - \mathbf{W}'\|_F \\ \lesssim d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^4} \quad (18)$$

Lemma 6. (Hessian smoothness of population loss) In the FCN model (1), for any permutation matrix \mathbf{P} , we have

$$\|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \lesssim K^{\frac{3}{2}} \cdot \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \\ \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \quad (19)$$

Lemma 7. (Local strong convexity of population loss) In the FCN model (1), for any permutation matrix \mathbf{P} , if $\|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \leq r$ for an $\epsilon_0 \in (0, \frac{1}{4})$, then,

$$\begin{aligned} & \frac{4(1-\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\right) \\ & \cdot \|\Sigma_l^{-1}\|^{-\frac{1}{2}} \cdot \mathbf{I}_{dK} \preceq \nabla^2 \bar{f}(\mathbf{W}) \preceq \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \cdot \mathbf{I}_{dK} \end{aligned} \quad (20)$$

Lemma 8. In the FCN model (1), for any permutation matrix \mathbf{P} , as long as $n \geq C' \cdot dK \log dK$ for some constant $C' > 0$, we have

$$\begin{aligned} & \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \\ & \leq \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}} \end{aligned} \quad (21)$$

with probability at least $1 - d^{-10}$.

We next show the proof of Lemma 1.

2) *Proof:* From Lemma 7 and 8, with probability at least $1 - d^{-10}$,

$$\begin{aligned} \nabla^2 f_n(\mathbf{W}) & \succeq \nabla^2 \bar{f}(\mathbf{W}) - \|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 f_n(\mathbf{W})\| \cdot \mathbf{I} \\ & \succeq \Omega\left(\frac{(1-\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\right) \cdot \mathbf{I}\right. \\ & \left. - O\left(C_6 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}}\right) \cdot \mathbf{I}\right) \end{aligned} \quad (22)$$

As long as the sample complexity is set to satisfy

$$\begin{aligned} & \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \cdot \sqrt{\frac{dK \log n}{n}} \leq \frac{\epsilon_0}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K \kappa^2} \\ & \cdot \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\right) \cdot \mathbf{I} \end{aligned} \quad (23)$$

i.e.,

$$\begin{aligned} & n \gtrsim \epsilon_0^{-2} \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2\right)^2 \\ & \cdot \left(\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\right) \cdot \mathbf{I}\right)^{-2} dK^5 \log^2 d \end{aligned} \quad (24)$$

for some constant $C_1 > 0$, then we have the lower bound of the Hessian with probability at least $1 - d^{-10}$.

$$\begin{aligned} \nabla^2 f_n(\mathbf{W}) & \succeq \Omega\left(\frac{(1-2\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\right) \cdot \mathbf{I}\right) \end{aligned} \quad (25)$$

By (20) and (21), we can also derive the upper bound as follows,

$$\begin{aligned} \|\nabla^2 f_n(\mathbf{W})\| & \leq \|\nabla^2 \bar{f}(\mathbf{W})\| + \|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \\ & \lesssim \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \\ & \quad + \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}} \\ & \lesssim \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \end{aligned} \quad (26)$$

Combining (25) and (26), we have

$$\begin{aligned} & \Omega\left(\frac{(1-2\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\right) \cdot \mathbf{I}\right) \\ & \preceq \nabla^2 f_n(\mathbf{W}) \\ & \preceq \sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\|_\infty + \|\Sigma_l^{\frac{1}{2}}\|)^2 \cdot \mathbf{I} \end{aligned} \quad (27)$$

with probability at least $1 - d^{-10}$.

C. Proof of Lemma 2

We restate the formal version of Lemma 2 in the following.

Lemma 2. (Linear convergence of gradient descent) Assume the conditions in Lemma 1 hold. Given any fixed permutation matrix $\mathbf{P} \in \mathbb{R}^{K \times K}$, if the local convexity of $\mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ holds, there exists a critical point in $\mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ for some constant $C_3 > 0$, and $\epsilon_0 \in (0, \frac{1}{2})$, such that

$$\begin{aligned} & \|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \\ & \lesssim \frac{K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 (1+\xi)} \cdot \sqrt{d \log n/n}}{\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\right) \cdot \mathbf{I}} \end{aligned} \quad (28)$$

If the initial point $\mathbf{W}_0 \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$, the gradient descent linearly converges to $\widehat{\mathbf{W}}_n$, i.e.,

$$\begin{aligned} \|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F & \leq \|\mathbf{W}_0 - \widehat{\mathbf{W}}_n\|_F \cdot \left(1 - \Omega\left(\frac{\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\right) \cdot \mathbf{I}}{K^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2}\right)\right)^t \end{aligned} \quad (29)$$

with probability at least $1 - d^{-10}$.

1) A useful lemma:

Lemma 9. If r is defined in (139) for $\epsilon_0 \in (0, \frac{1}{4})$, then with probability at least $1 - d^{-10}$, we have¹²

$$\begin{aligned} & \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla f_n(\mathbf{W}) - \nabla \bar{f}(\mathbf{W})\| \\ & \lesssim \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l\|)^2 \sqrt{\frac{d \log n}{n}}} (1+\xi) \end{aligned} \quad (30)$$

¹² $\nabla f_n(\mathbf{W})$ is defined as $\frac{1}{n} \sum_{i=1}^n (\nabla l(\mathbf{W}, \mathbf{x}_i, y_i) + \nu_i)$ in algorithm 1

, where \mathbf{P} is a permutation matrix.

We next show the proof of Lemma 2.

2) *Proof:* Following the proof of Theorem 2 in [28], first, we have Taylor's expansion of $f_n(\widehat{\mathbf{W}}_n)$

$$\begin{aligned} f_n(\widehat{\mathbf{W}}_n) &= f_n(\mathbf{W}^* \mathbf{P}) + \left\langle \nabla \tilde{f}_n(\mathbf{W}^* \mathbf{P}), \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \right\rangle \\ &\quad + \frac{1}{2} \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \nabla^2 f_n(\mathbf{W}') \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \end{aligned} \quad (31)$$

Here \mathbf{W}' is on the straight line connecting $\mathbf{W}^* \mathbf{P}$ and $\widehat{\mathbf{W}}_n$. By the fact that $f_n(\widehat{\mathbf{W}}_n) \leq f_n(\mathbf{W}^* \mathbf{P})$, we have

$$\begin{aligned} &\frac{1}{2} \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \nabla^2 f_n(\mathbf{W}') \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \\ &\leq \left| \nabla f_n(\mathbf{W}^* \mathbf{P})^\top \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \right| \end{aligned} \quad (32)$$

From Lemma 7 and Lemma 9, we have

$$\begin{aligned} &\frac{4}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}\right), \\ &\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}} \|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F^2 \\ &\leq \frac{1}{2} \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \nabla^2 f_n(\mathbf{W}') \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \end{aligned} \quad (33)$$

and

$$\begin{aligned} &\left| \nabla \tilde{f}_n(\mathbf{W}^* \mathbf{P})^\top \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \right| \\ &\leq \|\nabla \tilde{f}_n(\mathbf{W}^* \mathbf{P})\| \cdot \|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \\ &\leq (\|\nabla \tilde{f}_n(\mathbf{W}^* \mathbf{P}) - \nabla \tilde{f}(\mathbf{W}^* \mathbf{P})\| + \|\nabla \tilde{f}(\mathbf{W}^* \mathbf{P})\|) \\ &\quad \cdot \|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \\ &\leq O\left(\sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{d \log n}{n}} (1 + \xi)\right) \\ &\quad \|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \end{aligned} \quad (34)$$

The second to last step of (34) comes from the triangle inequality, and the last step follows from the fact $\nabla f(\mathbf{W}^* \mathbf{P}) = 0$. Combining (32), (33) and (34), we have

$$\begin{aligned} &\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \\ &\lesssim \frac{K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2} (1 + \xi) \cdot \sqrt{d \log n / n}}{\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}\right), \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}} \end{aligned} \quad (35)$$

Therefore, we have concluded that there indeed exists a critical point $\widehat{\mathbf{W}}$ in $\mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$. Then we show the linear convergence of Algorithm 1 as below. By the update rule, we have

$$\begin{aligned} &\mathbf{W}_{t+1} - \widehat{\mathbf{W}}_n \\ &= \mathbf{W}_t - \eta_0 (\nabla f_n(\mathbf{W}_t) + \frac{1}{n} \sum_{i=1}^n \nu_i) - (\widehat{\mathbf{W}}_n - \eta_0 \nabla f_n(\widehat{\mathbf{W}}_n)) \\ &= \left(\mathbf{I} - \eta_0 \int_0^1 \nabla^2 f_n(\mathbf{W}(\gamma)) \right) (\mathbf{W}_t - \widehat{\mathbf{W}}_n) - \frac{\eta_0}{n} \sum_{i=1}^n \nu_i \end{aligned} \quad (36)$$

where $\mathbf{W}(\gamma) = \gamma \widehat{\mathbf{W}}_n + (1 - \gamma) \mathbf{W}_t$ for $\gamma \in (0, 1)$. Since $\mathbf{W}(\gamma) \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$, by Lemma 1, we have

$$H_{\min} \cdot \mathbf{I} \preceq \nabla^2 f_n(\mathbf{W}(\gamma)) \preceq H_{\max} \cdot \mathbf{I} \quad (37)$$

where $H_{\min} = \Omega\left(\frac{1}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}\right), \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)$, $H_{\max} = \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l\|)^2$. Therefore,

$$\begin{aligned} &\|\mathbf{W}_{t+1} - \widehat{\mathbf{W}}_n\|_F \\ &= \|\mathbf{I} - \eta_0 \int_0^1 \nabla^2 f_n(\mathbf{W}(\gamma))\| \cdot \|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F + \left\| \frac{\eta_0}{n} \sum_{i=1}^n \nu_i \right\|_F \\ &\leq (1 - \eta_0 H_{\min}) \|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F + \left\| \frac{\eta_0}{n} \sum_{i=1}^n \nu_i \right\|_F \end{aligned} \quad (38)$$

$$\begin{aligned} &\text{By setting } \eta_0 = \frac{1}{H_{\max}} = O\left(\frac{1}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l\|)^2}\right), \text{ we obtain} \\ &\|\widehat{\mathbf{W}}_{t+1} - \widehat{\mathbf{W}}_n\|_F \leq \left(1 - \frac{H_{\min}}{H_{\max}}\right) \|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F + \frac{\eta_0}{n} \sum_{i=1}^n \|\nu_i\|_F \end{aligned} \quad (39)$$

Therefore, Algorithm 1 converges linearly to the local minimizer with an extra statistical error.

By Hoeffding's inequality in [76] and Property 2, we have

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \|\nu_i\|_F \geq \sqrt{\frac{dK \log n}{n}} \xi\right) \lesssim \exp\left(-\frac{\xi^2 dK \log n}{dK \xi^2}\right) \\ &\lesssim d^{-10} \end{aligned} \quad (40)$$

Therefore, with probability $1 - d^{-10}$ we can derive

$$\begin{aligned} &\|\widehat{\mathbf{W}}_t - \widehat{\mathbf{W}}_n\|_F \\ &\leq \left(1 - \frac{H_{\min}}{H_{\max}}\right)^t \|\mathbf{W}_0 - \widehat{\mathbf{W}}_n\|_F + \frac{H_{\max} \eta_0}{H_{\min}} \sqrt{\frac{dK \log n}{n}} \xi \end{aligned} \quad (41)$$

D. Proof of Lemma 3

We first restate the formal version of Lemma 3 in the following.

Lemma 3. (Tensor initialization) For classification model, with $D_6(\Psi)$ defined in Definition 2, we have that if the sample size

$$n \geq \kappa^8 K^4 \tau^{12} D_6(\Psi) \cdot d \log^2 d, \quad (42)$$

then the output $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$ satisfies

$$\|\mathbf{W}_0 - \mathbf{W}^* \mathbf{P}^*\| \lesssim \kappa^6 K^3 \cdot \tau^6 \sqrt{D_6(\Psi)} \sqrt{\frac{d \log n}{n}} \|\mathbf{W}^*\| \quad (43)$$

with probability at least $1 - n^{-\Omega(\delta_1^4)}$ for a specific permutation matrix $\mathbf{P}^* \in \mathbb{R}^{K \times K}$.

1) *Useful lemmas:* Lemmas 10, 11, 12, 13, and 14 are needed to prove Lemma 3.

Lemma 10. Let \mathbf{Q}_2 and \mathbf{Q}_3 follow Definition 3. Let S be a set of i.i.d. samples generated from the mixed Gaussian distribution $\sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$. Let $\hat{\mathbf{Q}}_2, \hat{\mathbf{Q}}_3$ be the empirical version of $\mathbf{Q}_2, \mathbf{Q}_3$ using data set S , respectively. Then with a probability at least $1 - 2n^{-\Omega(\delta_1(\mathbf{W}^*)^{4d})}$, we have

$$\|\mathbf{Q}_2 - \hat{\mathbf{Q}}_2\| \lesssim \sqrt{\frac{d \log n}{n}} \cdot \delta_1(\mathbf{W}^*)^2 \cdot \tau^6 \sqrt{D_2(\Psi)D_4(\Psi)} \quad (44)$$

if the mixed Gaussian distribution is not symmetric. We also have

$$\begin{aligned} & \|\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \hat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\| \\ & \lesssim \sqrt{\frac{d \log n}{n}} \cdot \delta_1(\mathbf{W}^*)^2 \cdot \tau^6 \sqrt{D_2(\Psi)D_4(\Psi)} \end{aligned} \quad (45)$$

for any arbitrary vector $\boldsymbol{\alpha} \in \mathbb{R}^d$, if the mixed Gaussian distribution is symmetric.

Lemma 11. Let $\mathbf{U} \in \mathbb{E}^{d \times K}$ be the orthogonal column span of \mathbf{W}^* . Let $\boldsymbol{\alpha}$ be a fixed unit vector and $\hat{\mathbf{U}} \in \mathbb{R}^{d \times K}$ denote an orthogonal matrix satisfying $\|\mathbf{U}\mathbf{U}^\top - \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\| \leq \frac{1}{4}$. Define $\mathbf{R}_3 = \mathbf{Q}_3(\hat{\mathbf{U}}, \hat{\mathbf{U}}, \hat{\mathbf{U}})$, where \mathbf{Q}_3 is defined in Definition 3. Let $\hat{\mathbf{R}}_3$ be the empirical version of \mathbf{R}_3 using data set S , where each sample of S is i.i.d. sampled from the mixed Gaussian distribution $\sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$. Then with a probability at least $1 - n^{-\Omega(\delta_1^4(\mathbf{W}^*))}$, we have

$$\|\hat{\mathbf{R}}_3 - \mathbf{R}_3\| \lesssim \delta_1(\mathbf{W}^*)^2 \cdot (\tau^6 \sqrt{D_6(\Psi)}) \cdot \sqrt{\frac{\log n}{n}} \quad (46)$$

Lemma 12. Let $\hat{\mathbf{Q}}_1$ be the empirical version of \mathbf{Q}_1 using dataset S . Then with a probability at least $1 - 2n^{-\Omega(d)}$, we have

$$\|\hat{\mathbf{Q}}_1 - \mathbf{Q}_1\| \lesssim (\tau^2 \sqrt{D_2(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}} \quad (47)$$

Lemma 13. ([26], Lemma E.6) Let $\mathbf{Q}_2, \mathbf{Q}_3$ be defined in Definition 3 and $\hat{\mathbf{Q}}_2, \hat{\mathbf{Q}}_3$ be their empirical version, respectively. Let $\mathbf{U} \in \mathbb{R}^{d \times K}$ be the column span of \mathbf{W}^* . Assume $\|\mathbf{Q}_2 - \hat{\mathbf{Q}}_2\| \leq \frac{\delta_K(\mathbf{Q}_2)}{10}$ for non-symmetric distribution cases and $\|\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \hat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\| \leq \frac{\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}))}{10}$ for symmetric distribution cases and any arbitrary vector $\boldsymbol{\alpha} \in \mathbb{R}^d$. Then after $T = O(\log(\frac{1}{\epsilon}))$ iterations, the output of the Tensor Initialization Method 1, $\hat{\mathbf{U}}$ will satisfy

$$\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\| \lesssim \frac{\|\hat{\mathbf{Q}}_2 - \mathbf{Q}_2\|}{\delta_K(\mathbf{Q}_2)} + \epsilon, \quad (48)$$

which implies

$$\|(\mathbf{I} - \hat{\mathbf{U}}\hat{\mathbf{U}}^\top)\mathbf{w}_i^*\| \lesssim \left(\frac{\|\mathbf{Q}_2 - \hat{\mathbf{Q}}_2\|}{\delta_K(\mathbf{Q}_2)} + \epsilon\right)\|\mathbf{w}_i^*\| \quad (49)$$

if the mixed Gaussian distribution is not symmetric. Similarly, we have

$$\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\| \lesssim \frac{\|\hat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\|}{\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}))} + \epsilon, \quad (50)$$

which implies

$$\begin{aligned} & \|(\mathbf{I} - \hat{\mathbf{U}}\hat{\mathbf{U}}^\top)\mathbf{w}_i^*\| \\ & \lesssim \left(\frac{\|\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \hat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\|}{\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}))} + \epsilon\right)\|\mathbf{w}_i^*\| \end{aligned} \quad (51)$$

if the mixed Gaussian distribution is symmetric.

Lemma 14. ([26], Lemma E.13) Let $\mathbf{U} \in \mathbb{R}^{d \times K}$ be the orthogonal column span of \mathbf{W}^* . Let $\hat{\mathbf{U}} \in \mathbb{R}^{d \times K}$ be an orthogonal matrix such that $\|\mathbf{U}\mathbf{U}^\top - \hat{\mathbf{U}}\hat{\mathbf{U}}^\top\| \lesssim \gamma_1 \lesssim \frac{1}{\kappa^2 \sqrt{K}}$. For each $i \in [K]$, let $\hat{\mathbf{v}}_i$ denote the vector satisfying $\|\hat{\mathbf{v}}_i - \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_i^*\| \leq \gamma_2 \lesssim \frac{1}{\kappa^2 \sqrt{K}}$. Let \mathbf{Q}_1 be defined in Lemma 12 and $\hat{\mathbf{Q}}_1$ be its empirical version. If $\|\mathbf{Q}_1 - \hat{\mathbf{Q}}_1\| \leq \gamma_3 \|\mathbf{Q}_1\| \lesssim \frac{1}{4} \|\mathbf{Q}_1\|$, then we have

$$\|\|\mathbf{w}_i^*\| - \hat{\alpha}_i\| \leq (\kappa^4 K^{\frac{3}{2}}(\gamma_1 + \gamma_2) + \kappa^2 K^{\frac{1}{2}} \gamma_3) \|\mathbf{w}_i^*\| \quad (52)$$

We next show the proof of Lemma 3.

2) *Proof:* By the triangle inequality, we have

$$\begin{aligned} & \|\mathbf{w}_j^* - \hat{\alpha}_j \hat{\mathbf{U}} \hat{\mathbf{v}}_j\| \\ & = \left\| \mathbf{w}_j^* - \|\mathbf{w}_j^*\| \hat{\mathbf{U}} \hat{\mathbf{v}}_j + \|\mathbf{w}_j^*\| \hat{\mathbf{U}} \hat{\mathbf{v}}_j - \hat{\alpha}_j \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| \\ & \leq \left\| \mathbf{w}_j^* - \|\mathbf{w}_j^*\| \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| + \left\| \|\mathbf{w}_j^*\| \hat{\mathbf{U}} \hat{\mathbf{v}}_j - \hat{\alpha}_j \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| \\ & \leq \|\mathbf{w}_j^*\| \left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| + \left\| \|\mathbf{w}_j^*\| - \hat{\alpha}_j \right\| \|\hat{\mathbf{U}} \hat{\mathbf{v}}_j\| \\ & \leq \|\mathbf{w}_j^*\| \left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* + \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| \\ & \quad + \left\| \|\mathbf{w}_j^*\| - \hat{\alpha}_j \right\| \|\hat{\mathbf{U}} \hat{\mathbf{v}}_j\| \\ & \leq \delta_1(\mathbf{W}^*) \left(\left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* \right\| + \left\| \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* - \hat{\mathbf{v}}_j \right\| \right) \\ & \quad + \left\| \|\mathbf{w}_j^*\| - \hat{\alpha}_j \right\| \end{aligned} \quad (53)$$

From Lemma 10, Lemma 13, $\delta_K(\mathbf{Q}_2) \lesssim \delta_K^2(\mathbf{W}^*)$ and $\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})) \lesssim \delta_K^2(\mathbf{W}^*)$ for any arbitrary vector $\boldsymbol{\alpha} \in \mathbb{R}^d$, we have

$$\begin{aligned} & \left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* \right\| \\ & \lesssim \frac{\|\mathbf{Q}_2 - \hat{\mathbf{Q}}_2\|}{\delta_K(\mathbf{Q}_2)} \lesssim \sqrt{\frac{d \log n}{n}} \cdot \frac{\delta_1(\mathbf{W}^*)^2}{\delta_K(\mathbf{W}^*)^2} \cdot \tau^6 \sqrt{D_2(\Psi)D_4(\Psi)} \\ & = \sqrt{\frac{d \log n}{n}} \cdot \kappa^2 \cdot \tau^6 \sqrt{D_2(\Psi)D_4(\Psi)} \end{aligned} \quad (54)$$

if the mixed Gaussian distribution is not symmetric, and

$$\begin{aligned} & \left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* \right\| \lesssim \frac{\|\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \hat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\|}{\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}))} \\ & = \sqrt{\frac{d \log n}{n}} \cdot \kappa^2 \cdot \tau^6 \sqrt{D_2(\Psi)D_4(\Psi)} \end{aligned} \quad (55)$$

if the mixed Gaussian distribution is symmetric. Moreover, we have

$$\begin{aligned} & \left\| \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* - \hat{\mathbf{v}}_j \right\| \\ & \leq \frac{K^{\frac{3}{2}}}{\delta_K^2(\mathbf{W}^*)} \|\mathbf{R}_3 - \hat{\mathbf{R}}_3\| \lesssim \kappa^2 \cdot (\tau^6 \sqrt{D_6(\Psi)}) \cdot \sqrt{\frac{K^3 \log n}{n}} \end{aligned} \quad (56)$$

in which the first step is by Theorem 3 in [77], and the second step is by Lemma 11. By Lemma 14, we have

$$\left\| \|\mathbf{w}_j^*\| - \hat{\alpha}_j \right\| \leq (\kappa^4 K^{\frac{3}{2}} (\gamma_1 + \gamma_2) + \kappa^2 K^{\frac{1}{2}} \gamma_3) \|\mathbf{W}^*\| \quad (57)$$

Therefore, taking the union bound of failure probabilities in Lemmas 10, 11, and 12 and by $D_2(\Psi)D_4(\Psi) \leq D_6(\Psi)$ from Property 10, we have that if the sample size $n \geq \kappa^8 K^4 \tau^{12} D_6(\Psi) \cdot d \log^2 d$, then the output $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$ satisfies

$$\|\mathbf{W}_0 - \mathbf{W}^*\| \lesssim \kappa^6 K^3 \cdot \tau^6 \sqrt{D_6(\Psi)} \sqrt{\frac{d \log n}{n}} \|\mathbf{W}^*\| \quad (58)$$

with probability at least $1 - n^{-\Omega(\delta_1^4(\mathbf{W}^*))}$

E. Extension to Multi-Classification

We only show the analysis of binary classification in the main body of the paper due to the simplicity of presentation and highlight our major conclusions on the group imbalance. We briefly introduce how to extend our analysis on binary classification to multi-classification in this section. The main idea is to define the label as a multi-dimensional vector and apply the analysis for the binary classification case multiple times. Specifically, let C be the number of classes, where $C = 2^c$ for a positive integer c . The label \mathbf{y}_i is a c -dimensional vector, and its j th entry $y_{i,j} \in \{0, 1\}$ for $j \in [c]$ and $i \in [n]$. Such a formulation for the multi-classification problem can be found in [45], [78]. Then, following the binary setup, data \mathbf{x}_i, y_i satisfies

$$\mathbb{P}(y_{i,j} = 1 | \mathbf{x}_i) = H_j(\mathbf{W}^*, \mathbf{x}_i), \quad (59)$$

for some unknown ground-truth neural network with unknown weights \mathbf{W}^* , where $H_j(\mathbf{W}^*, \mathbf{x}_i)$ is the j -th entry of $\mathbf{H}(\mathbf{W}^*, \mathbf{x}_i) \in \mathbb{R}^c$ with the parameter $\mathbf{W}_j^* \in \mathbb{R}^{d \times K}$.

The training process is to minimize the empirical risk function with a cross-entropy loss

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c -y_{i,j} \log(H_j(\mathbf{W}, \mathbf{x}_i)) \\ & - (1 - y_{i,j}) \log(1 - H_j(\mathbf{W}, \mathbf{x}_i)) \\ & := \sum_{j=1}^c f_n^{(j)}(\mathbf{W}). \end{aligned} \quad (60)$$

Note that $f_n^{(j)}(\mathbf{W})$ has exactly the form as (2) in our paper for the binary case. Therefore, we can apply the existing theoretical results for $f_n^{(j)}(\mathbf{W})$ with all $j \in [c]$, and summing up all the bounds yields the theoretical results for the multi-class case.

We implement experiments on the CelebA dataset for 4-classification. The only change is that we use the combinations of two attributes, “blonde hair” and “pale skin” to generate four classes of data. All other settings are the same. The results are the following.

One can observe from Figure 9 that when the noise level δ^2 increases, i.e., when the co-variance of the minority group increases, both the minority-group and average test accuracy increase first and then decrease, coinciding with our insight

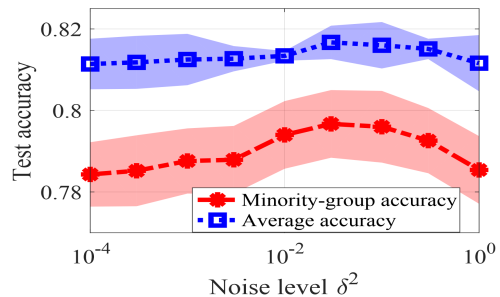


Fig. 9: Test accuracy against the augmented noise level for 4-classification.

(P3). In Figure 10 (a) and (b), we can see opposite trends if we increase the fraction of the minority group in the training data, with the male being the minority or the female being the minority. Figures 9 and 10 are consistent with our findings in Figures 1 and 8, respectively.

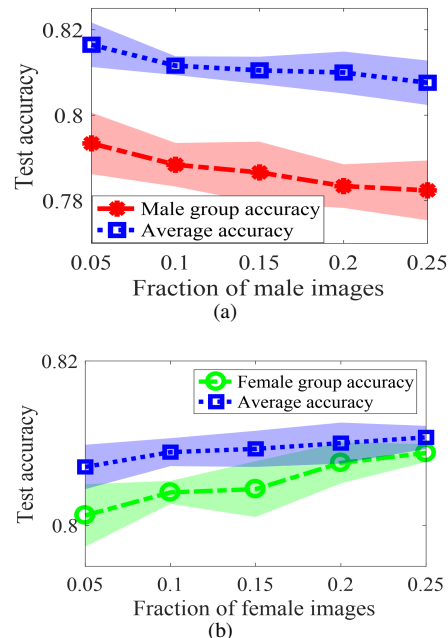


Fig. 10: The test accuracy on CelebA dataset has opposite trends when the minority group fraction increases for 4-classification. (a) Male group is the minority (b) Female group is the minority

F. Discussion about Gaussian Mixture Model (GMM)

The GMM distribution intuitively means that each data comes from a certain group, which is represented by a certain Gaussian component with mean μ_l and co-variance Σ_l , $l \in [L]$. The fraction λ_l stands for the fraction of group $l \in [L]$. This formulation is motivated by existing works [3], [25], which are related to group imbalance in the case of convolutional neural networks. One can see that each data feature follows GMM by Eqn (4) of [3]. In our setup, we define the data following the GMM for fully connected neural networks, where labels are determined by the mixture of Gaussian input and the ground-truth model.

We also conduct an experiment on CelebA to show some practical datasets satisfy the GMM model. We select data with two attributes, male and female. We extract features before the fully connected layer of the ResNet 9 model and fit the features to a two-component GMM using the EM algorithm [69]. The goodness of fit is measured by the average log-likelihood score as in [79]. We compute the average log-likelihood score of the CelebA dataset as 1.63 bits/dimension. To see that 1.63 bits/dimension reflects a good fitting, we generate synthetic data following the estimated GMM by CelebA and then compute the log-likelihood score of fitting the synthetic data to a two-component GMM. The resulting score is 1.80 bits/dimension for the synthetic two-component GMM data. Therefore, we can see that the quality of fitting CelebA is almost as good as fitting synthetic data generated by a GMM, which indicates that a two-component GMM is a good fitting for the studied practical dataset generated by CelebA.

Since many existing theoretical works [26], [28], [27], [59], [29] consider the data as standard Gaussian, we also compute the score if we use a single Gaussian to fit the data. The resulting average log-likelihood score is 1.08 bits/dimension, which is evidently smaller than the two-component GMM considered in our manuscript. This shows our GMM can better describe the real data.

Moreover, our GMM assumption goes beyond the state-of-the-art assumption of the standard Gaussian for loss landscape analysis for one-hidden-layer neural networks with convergence guarantees [26], [80], [57], [28], [29]. When generalizing from the standard Gaussian to GMM, we make new technical contributions to analyzing the more complicated and challenging landscape of the risk function because of a mixture of non-zero mean and non-unit standard deviation Gaussians. We characterize the impact of the parameters of the GMM model on the learning convergence and generalization performance. In contrast, other existing theoretical works [10], [11], [12], [81] that consider other input distributions that are more general than the standard Gaussian model do not explicitly quantify the impact of the distribution parameters on the loss landscape and generalization performance.

G. Discussion about σ_{\min} and τ

In this section, we show that the assumption that σ_{\min} is not very close to zero, or equivalently, $\tau = \Theta(1)$, is mild. Even when the real data have singular values very close to zero, they can be approximated by low-rank data without hurting the performance by only keeping a few significant singular values and setting the small ones to zero. Thus, every practical dataset can be approximated by a dataset with $\tau = \Theta(1)$ while maintaining the same performance. We verify this by an experiment of binary classification on CelebA [31]. After training with a ResNet-9, the output feature of each testing image is 256-dimensional. One can find that the singular value of the covariance matrix of features is close to 0 except for the top singular values. The feature matrix reconstructed with top singular values can achieve comparable testing accuracy as using all singular values, as shown in Table II. One can observe

that the feature matrix reconstruction with top-5 singular values, which is 2% of all the singular vectors, leads to a test accuracy already close to that using all singular vectors, and the performance gap is smaller than 4.5%. We can compute that $\tau = 4.6155 = \Theta(1)$ for the feature matrix reconstructed by top 5 singular values.

TABLE II: Testing accuracy with a reconstructed feature matrix using different amounts of singular values (s.v.)

Reconstruct with	top 5 s.v.	top 25 s.v.	all 256 s.v.
Accuracy	84.00%	85.00%	88.50%

REFERENCES

- [1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.
- [2] T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3428–3448.
- [3] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, "An investigation of why overparameterization exacerbates spurious correlations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8346–8356.
- [4] S. Sagawa*, P. W. Koh*, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=ryxGuJrFvS>
- [5] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, "Improving out-of-distribution robustness via selective augmentation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25407–25437.
- [6] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [8] F. J. Moreno-Barea, F. Strazzera, J. M. Jerez, D. Urda, and L. Franco, "Forward noise adjustment scheme for data augmentation," in *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2018, pp. 728–734.
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [10] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *36th International Conference on Machine Learning, ICML 2019*. International Machine Learning Society (IMLS), 2019, pp. 477–502.
- [11] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 242–252.
- [12] Z. Allen-Zhu, Y. Li, and Y. Liang, "Learning and generalization in overparameterized neural networks, going beyond two layers," in *Advances in neural information processing systems*, 2019, pp. 6158–6169.
- [13] Y. Cao and Q. Gu, "Generalization bounds of stochastic gradient descent for wide and deep neural networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 10836–10846.
- [14] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in *Advances in Neural Information Processing Systems*, 2018, pp. 8157–8166.
- [15] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in neural information processing systems*, 2018, pp. 8571–8580.
- [16] H. Li, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Generalization guarantee of training graph convolutional networks with graph topology sampling," in *International Conference on Machine Learning*. PMLR, 2022, pp. 13014–13051.

- [17] Z. Shi, J. Wei, and Y. Liang, "A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features," in *International Conference on Learning Representations*, 2021.
- [18] S. Karp, E. Winston, Y. Li, and A. Singh, "Local signal adaptivity: Provable feature learning in neural networks beyond kernels," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 883–24 897, 2021.
- [19] Z. Allen-Zhu and Y. Li, "Feature purification: How adversarial training performs robust deep learning," in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2022, pp. 977–988.
- [20] H. Li, M. Wang, S. Liu, and P.-Y. Chen, "A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity," in *The Eleventh International Conference on Learning Representations*, 2023.
- [21] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: <https://openreview.net/forum?id=Uuf2q9TFXGA>
- [22] H. Li, M. Wang, T. Ma, S. Liu, Z. ZHANG, and P.-Y. Chen, "What improves the generalization of graph transformer? a theoretical dive into self-attention and positional encoding," in *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*, 2023. [Online]. Available: <https://openreview.net/forum?id=BaxFC3z9R6>
- [23] Y. Chen, W. Huang, K. Zhou, Y. Bian, B. Han, and J. Cheng, "Understanding and improving feature learning for out-of-distribution generalization," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: <https://openreview.net/forum?id=cozEoAtjG8>
- [24] H. Li, M. Wang, S. Lu, H. Wan, X. Cui, and P.-Y. Chen, "Transformers as multi-task feature selectors: Generalization analysis of in-context learning," in *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. [Online]. Available: <https://openreview.net/forum?id=BMQ4i2RVbE>
- [25] Y. Deng, Y. Yang, B. Mirzasoleiman, and Q. Gu, "Robust learning with progressive data expansion against spurious correlation," *arXiv e-prints*, pp. arXiv-2306, 2023.
- [26] 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-Volume 70*, 2017, pp. 4140–4149. [Online]. Available: <https://arxiv.org/pdf/1706.03175.pdf>
- [27] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Fast learning of graph neural networks with guaranteed generalizability: One-hidden-layer case," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 268–11 277.
- [28] H. Fu, Y. Chi, and Y. Liang, "Guaranteed recovery of one-hidden-layer neural networks via cross entropy," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3225–3235, 2020.
- [29] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis," in *International Conference on Learning Representations*, 2021.
- [30] S. Zhang, H. Li, M. Wang, M. Liu, P.-Y. Chen, S. Lu, S. Liu, K. Murugesan, and S. Chaudhury, "On the convergence and sample complexity analysis of deep q-networks with ϵ -greedy exploration," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: <https://openreview.net/forum?id=HWGWeaN76q>
- [31] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.
- [32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.
- [33] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*. Springer, 1998, pp. 9–50.
- [34] L. M. Koch, C. M. Schürch, A. Gretton, and P. Berens, "Hidden in plain sight: Subgroup shifts escape OOD detection," in *Medical Imaging with Deep Learning*, 2022. [Online]. Available: <https://openreview.net/forum?id=aZgiUNye2Cz>
- [35] J. Ma, J. Deng, and Q. Mei, "Subgroup generalization and fairness of graph neural networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [36] A. Biswas and S. Mukherjee, "Ensuring fairness under prior probability shifts," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 414–424.
- [37] S. Giguere, B. Metevier, Y. Brun, P. S. Thomas, S. Niekum, and B. C. da Silva, "Fairness guarantees under demographic shift," in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=wbPobLm6ueA>
- [38] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [39] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert systems with applications*, vol. 73, pp. 220–239, 2017.
- [40] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.
- [41] J. Byrd and Z. Lipton, "What is the effect of importance weighting in deep learning?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 872–881.
- [42] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 219–226.
- [43] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.
- [44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [45] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 680–697.
- [46] S. S. Mullick, S. Datta, and S. Das, "Generative adversarial minority oversampling," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1695–1704.
- [47] J. Kim, J. Jeong, and J. Shin, "M2m: Imbalanced classification via major-to-minor translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 896–13 905.
- [48] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *European Conf. on Computer Vision (ECCV)*, 2020.
- [49] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5212–5221.
- [50] C. Fang, H. He, Q. Long, and W. J. Su, "Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training," *Proceedings of the National Academy of Sciences*, vol. 118, no. 43, p. e2103091118, 2021.
- [51] L. Yang, H. Jiang, Q. Song, and J. Guo, "A survey on long-tailed visual recognition," *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1837–1872, 2022.
- [52] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi, "The majority can help the minority: Context-rich minority oversampling for long-tailed classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6887–6896.
- [53] S. S. Du, J. D. Lee, Y. Tian, A. Singh, and B. Póczos, "Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima," in *International Conference on Machine Learning*, 2018, pp. 1338–1347.
- [54] R. Ge, J. D. Lee, and T. Ma, "Learning one-hidden-layer neural networks with landscape design," in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=BkwHObbRZ>
- [55] 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.
- [56] I. Safran and O. Shamir, "Spurious local minima are common in two-layer relu neural networks," in *International Conference on Machine Learning*, 2018, pp. 4430–4438.
- [57] X. Zhang, Y. Yu, L. Wang, and Q. Gu, "Learning one-hidden-layer relu networks via gradient descent," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1524–1534.
- [58] S. Zhang, M. Wang, J. Xiong, S. Liu, and P.-Y. Chen, "Improved linear convergence of training cnns with generalizability guarantees: A one-hidden-layer case," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [59] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Why lottery ticket wins? a theoretical perspective of sample complexity on sparse neural networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

- [60] Y. Yoshida and M. Okada, “Data-dependence of plateau phenomenon in learning with neural network — statistical mechanical analysis,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019, pp. 1722–1730.
- [61] F. Mignacco, F. Krzakala, P. Urbani, and L. Zdeborová, “Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9540–9550, 2020.
- [62] S. S. Du, J. D. Lee, and Y. Tian, “When is a convolutional filter easy to learn?” in *International Conference on Learning Representations*, 2018.
- [63] S. Mei, A. Montanari, and P.-M. Nguyen, “A mean field view of the landscape of two-layer neural networks,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 33, pp. E7665–E7671, 2018.
- [64] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, “When do neural networks outperform kernel methods?” *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 820–14 830, 2020.
- [65] K. Pearson, “Contributions to the mathematical theory of evolution,” *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.
- [66] D. Hsu and S. M. Kakade, “Learning mixtures of spherical gaussians: moment methods and spectral decompositions,” in *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, 2013, pp. 11–20.
- [67] A. Moitra and G. Valiant, “Settling the polynomial learnability of mixtures of gaussians,” in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 93–102.
- [68] O. Regev and A. Vijayaraghavan, “On learning mixtures of well-separated gaussians,” in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2017, pp. 85–96.
- [69] R. A. Redner and H. F. Walker, “Mixture densities, maximum likelihood and the em algorithm,” *SIAM review*, vol. 26, no. 2, pp. 195–239, 1984.
- [70] N. Ho and X. Nguyen, “Convergence rates of parameter estimation for some weakly identifiable finite mixtures,” *Ann. Statist.*, vol. 44, no. 6, pp. 2726–2755, 12 2016. [Online]. Available: <https://doi.org/10.1214/16-AOS1444>
- [71] R. Dwivedi, N. Ho, K. Khamaru, M. I. Jordan, M. J. Wainwright, and B. Yu, “Singularity, misspecification, and the convergence rate of em,” *To appear, Annals of Statistics*, 2020.
- [72] R. Dwivedi, N. Ho, K. Khamaru, M. Wainwright, M. Jordan, and B. Yu, “Sharp analysis of expectation-maximization for weakly identifiable models,” in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. Online: PMLR, 26–28 Aug 2020, pp. 1866–1876.
- [73] 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.
- [74] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer, “Shift: A zero flop, zero parameter alternative to spatial convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9127–9135.
- [75] A. Dutta, E. H. Bergou, A. M. Abdelmoniem, C.-Y. Ho, A. N. Sahu, M. Canini, and P. Kalnis, “On the discrepancy between the theoretical analysis and practical implementations of compressed communication for distributed deep learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3817–3824.
- [76] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *arXiv preprint arXiv:1011.3027*, 2010.
- [77] V. Kuleshov, A. Chaganty, and P. Liang, “Tensor factorization via matrix factorization,” in *Artificial Intelligence and Statistics*, 2015, pp. 507–516.
- [78] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, “Learning spatial regularization with image-level supervisions for multi-label image classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5513–5522.
- [79] D. Zoran and Y. Weiss, “Natural images, gaussian mixtures and dead leaves,” *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [80] K. Zhong, Z. Song, and I. S. Dhillon, “Learning non-overlapping convolutional neural networks with multiple kernels,” *arXiv preprint arXiv:1711.03440*, 2017.
- [81] D. Zou, Y. Cao, D. Zhou, and Q. Gu, “Gradient descent optimizes over-parameterized deep relu networks,” *Machine Learning*, vol. 109, no. 3, pp. 467–492, 2020.
- [82] M. Janzamin, H. Sedghi, and A. Anandkumar, “Score function features for discriminative learning: Matrix and tensor framework,” *arXiv preprint arXiv:1412.2863*, 2014.
- [83] S. Mei, Y. Bai, and A. Montanari, “The landscape of empirical risk for non-convex losses,” *arXiv preprint arXiv:1607.06534*, 2016.

Supplementary Material

We begin our Supplementary Material here.

Section H provides more experiment results as a supplement of Section V.

Section I introduces the algorithm, especially the tensor initialization in detail.

Section J includes some definitions and properties as a preliminary to our proof.

Section K shows the proof of Theorem 1 and Corollary 1, followed by **Section L, M, and N** as the proof of three key Lemmas about local convexity, linear convergence and tensor initialization, respectively.

H. More Experiment Results

We present our experiment results on empirical datasets CelebA [31] and CIFAR-10¹³ in this section. To be more specific, we evaluate the impact of the variance levels introduced by different data augmentation methods on the learning performance. We also evaluate the impact of the minority group fraction in the training data on the learning performance. All the experiments are reported in a format of “mean \pm 2 \times standard deviation” with a random seed equal to 10. We implement our experiments on an NVIDIA GeForce RTX 2070 super GPU and a work station with 8 cores of 3.40GHz Intel i7 CPU.

1) *Tests on CelebA*: In addition to the Gaussian augmentation method in Figure 1 (b), we also evaluate the performance of data augmentation by cropping in Figure 11. The setup is exactly the same as that for Gaussian augmentation, except that we augment the data by cropping instead of adding Gaussian noise. Specifically, to generate an augmented image, we randomly crop an image with a size $w \times w \times 3$ and then resize back to $224 \times 224 \times 3$. One can observe that the minority-group and average test accuracy first increase and then decrease as w increases, which is in accordance with Insight (P3).

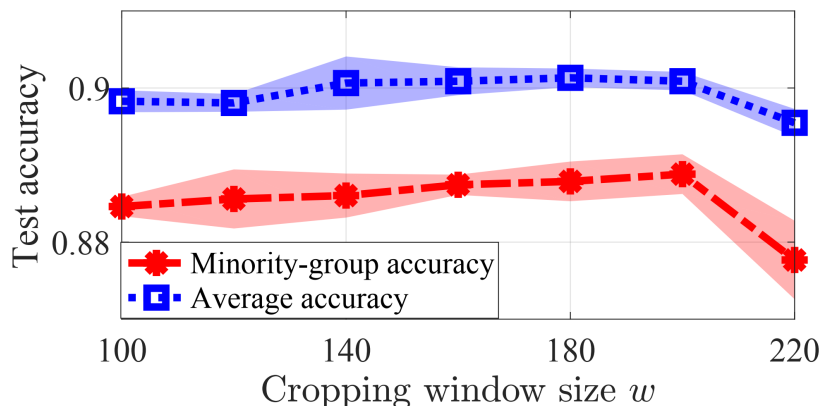


Fig. 11: The test accuracy of CelebA dataset with the data augmentation method of cropping.

2) *Tests on CIFAR-10*: Group 1 contains images with attributes “bird”, “cat”, “deer”, “dog”, “frog” and “horse.” Group 2 contains “airplane” images. In this setting, Group 1 has a larger variance. Because each image in CIFAR-10 only has one attribute, we consider the binary classification setting where all images in Group 1 are labeled as “animal” and all images are labeled as “airplane.” This is a special scenario that the group label is also the classification label. Note that our results hold for general setups where group labels and classification labels are irrelevant, like our previous results on CelebA. LeNet 5 [38] is selected to be the learning model.

We first pick 8000 animal images (majority) and 2000 airplane images (minority). We select 1000 out of 2000 airplane images to implement data augmentation, including both Gaussian augmentation and random cropping. For Gaussian augmentation, we add i.i.d. Gaussian noise drawn from $\mathcal{N}(0, \delta^2)$ to each entry¹⁴. For random cropping, we randomly crop the image with a certain size $w \times w \times 3$ and then resizing back to $32 \times 32 \times 3$. Figure 12 shows that when δ or w increase, i.e., the variance introduced by either augmentation method increases, both the minority-group and average test accuracy increase first and then decrease, which is consistent with our Insight (P3).

Then we fix the total number of training data to be 5000 and vary the fractions of the two groups. One can see opposite trends in Figure 13 if we increase the fraction of the minority group with the airplane being the minority and the animal being the minority, which reflects our Insight (P4).

¹³Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The CIFAR-10 dataset. www.cs.toronto.edu/~ekriz/cifar.html

¹⁴In this experiment, the noise is added to the raw image where the pixel value ranges from 0 to 255, while in the experiment of CelebA (Figure 1 (b)), the noise is added to the image after normalization where the pixel value ranges from 0 to 1.

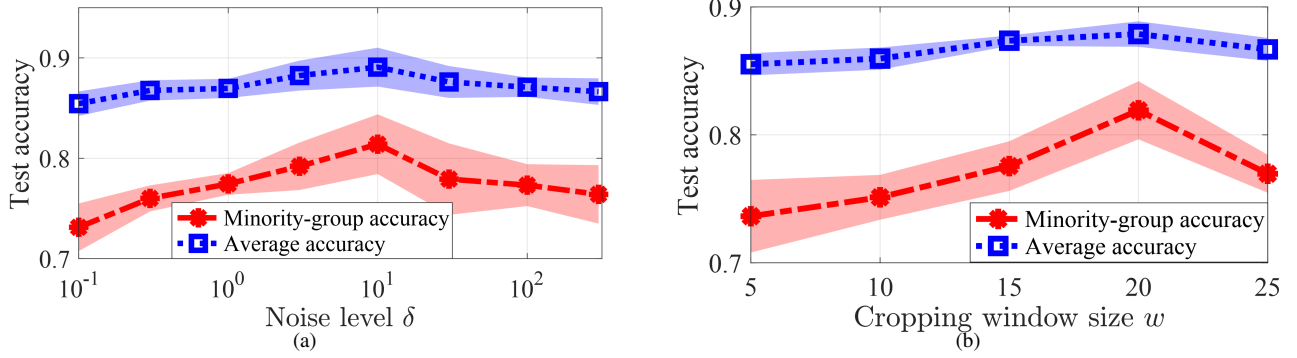


Fig. 12: The test accuracy of CIFAR-10 dataset with different data augmentation methods (a) Gaussian noise (b) cropping.

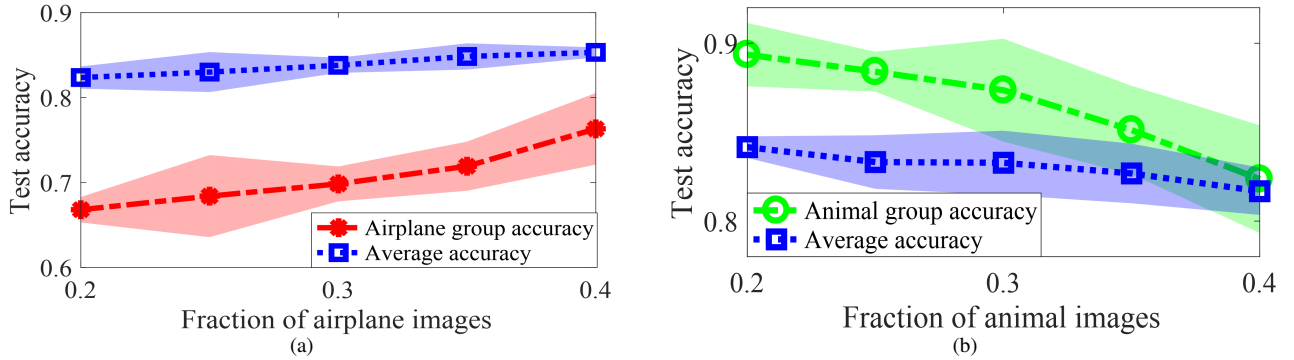


Fig. 13: The test accuracy of CIFAR-10 dataset has opposite trends when the minority group fraction increases (a) Airplane group is the minority. (b) Animal group is the minority.

I. Algorithm

We first introduce new notations to be used in this part and summarize key notions in Table III. We write $f(x) \lesssim (\gtrsim)g(x)$ if $f(x) \leq (\geq)\Theta(g(x))$. The gradient and the Hessian of a function $f(\mathbf{W})$ are denoted by $\nabla f(\mathbf{W})$ and $\nabla^2 f(\mathbf{W})$, respectively. $\mathbf{A} \succeq 0$ means \mathbf{A} is a positive semi-definite (PSD) matrix. $\mathbf{A}^{\frac{1}{2}}$ means that $\mathbf{A} = (\mathbf{A}^{\frac{1}{2}})^2$. The outer product of vectors $\mathbf{z}_i \in \mathbb{R}^{n_i}$, $i \in [l]$, is defined as $\mathbf{T} = \mathbf{z}_1 \otimes \dots \otimes \mathbf{z}_l \in \mathbb{R}^{n_1 \times \dots \times n_l}$ with $\mathbf{T}_{j_1 \dots j_l} = (\mathbf{z}_1)_{j_1} \dots (\mathbf{z}_l)_{j_l}$. Given a tensor $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times d_1}$, $\mathbf{B} \in \mathbb{R}^{n_2 \times d_2}$, $\mathbf{C} \in \mathbb{R}^{n_3 \times d_3}$, the (i_1, i_2, i_3) -th entry of the tensor $\mathbf{T}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is given by

$$\sum_{i'_1}^{n_1} \sum_{i'_2}^{n_2} \sum_{i'_3}^{n_3} \mathbf{T}_{i'_1, i'_2, i'_3} \mathbf{A}_{i'_1, i_1} \mathbf{B}_{i'_2, i_2} \mathbf{C}_{i'_3, i_3}. \quad (61)$$

TABLE III: Summary of notations

$\lambda_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, l \in [L]$	The fraction, mean, and covariance of the l -th component in the Gaussian mixture distribution, respectively.
d, n, K	The feature dimension, the number of training samples, and the number of neurons, respectively.
$\mathbf{W}^*, \mathbf{W}_t$	\mathbf{W}^* is the ground truth weight. \mathbf{W}_t is the updated weight in the t -th iteration.
f_n, \bar{f}, ℓ	f_n is the empirical risk function. \bar{f} is the average risk or the population risk function. ℓ is the cross-entropy loss function.
$\Psi, \sigma_{\max}, \sigma_{\min}, \tau$	Ψ denotes our Gaussian mixture model $(\lambda_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \forall l)$. $\sigma_{\max} = \max_{l \in [L]} \{\ \boldsymbol{\Sigma}_l\ ^{\frac{1}{2}}\}$. $\sigma_{\min} = \min_{l \in [L]} \{\ \boldsymbol{\Sigma}_l^{-1}\ ^{-\frac{1}{2}}\}$. $\tau = \sigma_{\max}/\sigma_{\min}$.
$\delta_i(\mathbf{W}^*), \eta, \kappa, i \in [K]$	$\delta_i(\mathbf{W}^*)$ is the i -th largest singular value of \mathbf{W}^* . η and κ are two functions of \mathbf{W}^* .
$\rho(\mathbf{u}, \sigma), \Gamma(\Psi), D_m(\Psi)$	These items are functions of the Gaussian mixture distribution Ψ used to develop our Theorem 1.
$\boldsymbol{\nu}_i, \xi$	$\boldsymbol{\nu}_i$ is the gradient noise. ξ is the upper bound of the noise level.
$\mathbf{Q}_j, j = 1, 2, 3$	\mathbf{Q}_j 's are tensors used in the initialization.
$\mathcal{B}(\Psi)$	A parameter appeared in the sample complexity bound (7).
$v(\Psi), q(\Psi)$	$v(\Psi)$ is the convergence rate (8). $q(\Psi)$ is a parameter in the definition of $v(\Psi)$ (9).
$\mathcal{E}_w(\Psi), \mathcal{E}, \mathcal{E}_l$	Generalization parameters. $\mathcal{E}_w(\Psi)$ appears in the error bound of the model (10). $\mathcal{E}(\Psi)$ and $\mathcal{E}_l(\Psi)$ are to characterize the average risk (11) and the group- l risk (12), respectively.

The method starts from an initialization $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$ computed based on the tensor initialization method (Subroutine 1) and then updates the iterates \mathbf{W}_t using gradient descent with the step size η_0 . To model the inaccuracy in computing the gradient, an i.i.d. zero-mean noise $\{\nu_i\}_{i=1}^n \in \mathbb{R}^{d \times K}$ with bounded magnitude $|(\nu_i)_{jk}| \leq \xi$ ($j \in [d], k \in [K]$) for some $\xi \geq 0$ are added in (6) when computing the gradient of the loss in (3).

Our tensor initialization method in Subroutine 1 is extended from [82] and [26]. The idea is to compute quantities (\mathbf{Q}_j in (62)) that are tensors of \mathbf{w}_i^* and then apply the tensor decomposition method to estimate \mathbf{w}_i^* . Because \mathbf{Q}_j can only be estimated from training samples, tensor decomposition does not return \mathbf{w}_i^* exactly but provides a close approximation, and this approximation is used as the initialization for Algorithm 1. Because the existing method on tensor construction only applies to the standard Gaussian distribution, we exploit the relationship between probability density functions and tensor expressions developed in [82] to design tensors suitable for the Gaussian mixture model. Formally,

Definition 3. For $j = 1, 2, 3$, we define

$$\mathbf{Q}_j := \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [y \cdot (-1)^j p^{-1}(\mathbf{x}) \nabla^{(j)} p(\mathbf{x})], \quad (62)$$

where $p(\mathbf{x})$, the probability density function of GMM is defined as

$$p(\mathbf{x}) = \sum_{l=1}^L \lambda_l (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_l|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l)\right) \quad (63)$$

If the Gaussian mixture model is symmetric, the symmetric distribution can be written as

$$\mathbf{x} \sim \begin{cases} \sum_{l=1}^{\frac{L}{2}} \lambda_l (\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) + \mathcal{N}(-\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)) & L \text{ is even} \\ \lambda_1 \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1) + \sum_{l=2}^{\frac{L-1}{2}} \lambda_l (\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) + \mathcal{N}(-\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)) & L \text{ is odd} \end{cases} \quad (64)$$

\mathbf{Q}_j is a j -th-order tensor of \mathbf{w}_i^* , e.g., $\mathbf{Q}_3 = \frac{1}{K} \sum_{i=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\phi'''(\mathbf{w}_i^{*\top} \mathbf{x})] \mathbf{w}_i^{*\otimes 3}$. These quantities cannot be directly computed from (62) but can be estimated by sample means, denoted by $\widehat{\mathbf{Q}}_j$ ($j = 1, 2, 3$), from samples $\{\mathbf{x}_i, y_i\}_{i=1}^n$. The following assumption guarantees that these tensors are nonzero and can thus be leveraged to estimate \mathbf{W}^* .

Assumption 1. The Gaussian Mixture Model in (64) satisfies the following conditions:

- 1) \mathbf{Q}_1 and \mathbf{Q}_3 are nonzero.
- 2) If the distribution is not symmetric, then \mathbf{Q}_2 is nonzero.

Assumption 1 is a very mild assumption¹⁵. Moreover, as indicated in [82], in the rare case that some quantities \mathbf{Q}_i ($i = 1, 2, 3$) are zero, one can construct higher-order tensors in a similar way as in Definition 3 and then estimate \mathbf{W}^* from higher-order tensors.

Subroutine 1 describes the tensor initialization method, which estimates the direction and magnitude of \mathbf{w}_j^* , $j \in [K]$, separately. The direction vectors are denoted as $\bar{\mathbf{w}}_j^* = \mathbf{w}_j^* / \|\mathbf{w}_j^*\|$ and the magnitude $\|\mathbf{w}_j^*\|$ is denoted as z_j . Lines 2-6 estimate the subspace \widehat{U} spanned by $\{\mathbf{w}_1^*, \dots, \mathbf{w}_K^*\}$ using $\widehat{\mathbf{Q}}_2$ or, in the case that $\mathbf{Q}_2 = 0$, a second-order tensor projected by $\widehat{\mathbf{Q}}_3$. Lines 7-8 estimate $\bar{\mathbf{w}}_j^*$ by employing the KCL algorithm [77]. Lines 9-10 estimate the magnitude z_j . Finally, the returned estimation of \mathbf{W}^* is used as an initialization \mathbf{W}_0 for Algorithm 1. The computational complexity of Subroutine 1 is $O(Knd)$ based on similar calculations as those in [26].

1) *Numerical Evaluation of Tensor Initialization:* Figure 14 shows the accuracy of the returned model by Algorithm 1. Here $n = 2 \times 10^5$, $d = 50$, $K = 2$, $\lambda_1 = \lambda_2 = 0.5$, $\boldsymbol{\mu}_1 = -0.3 \cdot \mathbf{1}$ and $\boldsymbol{\mu}_2 = \mathbf{0}$. We compare the tensor initialization with a random initialization in a local region $\{\mathbf{W} \in \mathbb{R}^{d \times K} : \|\mathbf{W} - \mathbf{W}^*\|_F \leq \epsilon\}$. Each entry of \mathbf{W}^* is selected from $[-0.1, 0.1]$ uniformly. Tensor initialization in Subroutine 1 returns an initial point close to one permutation of \mathbf{W}^* , with a relative error of 0.65. If the random initialization is also close to \mathbf{W}^* , e.g., $\epsilon = 0.1$, then the gradient descent algorithm converges to a critical point from both initializations, and the linear convergence rate is the same. We also test a random initialization with each entry drawn from $\mathcal{N}(0, 25)$. The initialization is sufficiently far from \mathbf{W}^* , and the algorithm does not converge. On a MacBook Pro with Intel(R) Core(TM) i5-7360U CPU at 2.30GHz and MATLAB 2017a, it takes 5.52 seconds to compute the tensor initialization. Thus, to reduce the computational time, we consider a random initialization with $\epsilon = 0.1$ in the experiments instead of computing tensor initialization.

¹⁵By mild, we mean given L , if Assumption 1 is not met for some Ψ_0 , there exists an infinite number of Ψ' in any neighborhood of Ψ_0 such that Assumption 1 holds for Ψ' .

Subroutine 1 Tensor Initialization Method

- 1: **Input:** Partition n pairs of data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ into three disjoint subsets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$
- 2: **if** the Gaussian Mixture distribution is not symmetric **then**
- 3: Compute $\widehat{\mathbf{Q}}_2$ using \mathcal{D}_1 . Estimate the subspace $\widehat{\mathbf{U}}$ by orthogonalizing the eigenvectors with respect to the K largest eigenvalues of $\widehat{\mathbf{Q}}_2$
- 4: **else**
- 5: Pick an arbitrary vector $\boldsymbol{\alpha} \in \mathbb{R}^d$, and use \mathcal{D}_1 to compute $\widehat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})$. Estimate $\widehat{\mathbf{U}}$ by orthogonalizing the eigenvectors with respect to the K largest eigenvalues of $\widehat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})$.
- 6: **end if**
- 7: Compute $\widehat{\mathbf{R}}_3 = \widehat{\mathbf{Q}}_3(\widehat{\mathbf{U}}, \widehat{\mathbf{U}}, \widehat{\mathbf{U}})$ from data set \mathcal{D}_2
- 8: Employ the KCL algorithm to compute vectors $\{\widehat{\mathbf{v}}_i\}_{i \in [K]}$, which are the estimates of $\{\widehat{\mathbf{U}}^\top \bar{\mathbf{w}}_i^*\}_{i=1}^K$. Then the direction vectors $\{\bar{\mathbf{w}}_i^*\}_{i=1}^K$ can be approximated by $\{\widehat{\mathbf{U}} \widehat{\mathbf{v}}_i\}_{i=1}^K$.
- 9: Compute $\widehat{\mathbf{Q}}_1$ from data set \mathcal{D}_3 .
- 10: Estimate the magnitude \widehat{z} by solving the optimization problem

$$\widehat{z} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \frac{1}{2} \left\| \widehat{\mathbf{Q}}_1 - \sum_{j=1}^K \alpha_j \bar{\mathbf{w}}_j^* \right\|^2 \quad (65)$$

- 11: **Return:** Use $\widehat{z}_j \widehat{\mathbf{U}} \widehat{\mathbf{v}}_j$ as the j th column of \mathbf{W}_0 , $j \in [K]$.
-

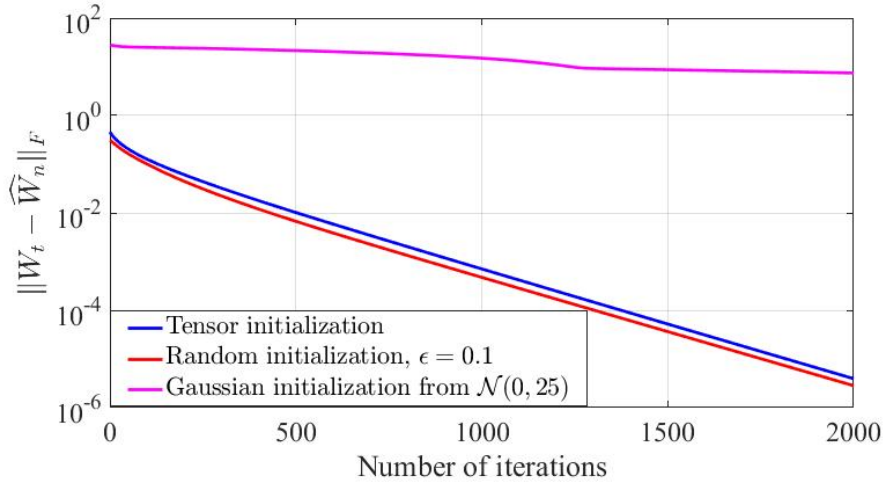


Fig. 14: Comparison between tensor initialization, a random initialization near \mathbf{W}^* , and an arbitrary random initialization

J. Preliminaries of the Main Proof

In this section, we introduce some **definitions** and **properties** that will be used to prove the main results.

First, we define the sub-Gaussian random variable and sub-Gaussian norm.

Definition 4. We say X is a sub-Gaussian random variable with sub-Gaussian norm $K > 0$, if $(\mathbb{E}|X|^p)^{\frac{1}{p}} \leq K\sqrt{p}$ for all $p \geq 1$. In addition, the sub-Gaussian norm of X , denoted $\|X\|_{\psi_2}$, is defined as $\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-\frac{1}{2}} (\mathbb{E}|X|^p)^{\frac{1}{p}}$.

Then we define the following three quantities. $\rho(\boldsymbol{\mu}, \sigma)$ is motivated by the ρ parameter for the standard Gaussian distribution in [26], and we generalize it to a Gaussian with an arbitrary mean and variance. We define the new quantities $\Gamma(\Psi)$ and $D_m(\Psi)$ for the Gaussian mixture model.

Definition 5. (Γ -function). With (13) and κ, η defined in Section III, we define

$$\Gamma(\Psi) = \sum_{l=1}^L \frac{\lambda_l}{\tau^K \kappa^2 \eta} \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\sigma_{\max}^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right) \quad (66)$$

Γ function is the weighted sum of ρ -function under mixture Gaussian distribution. This function is positive and upper bounded by a small value. Γ goes to zero if all $\|\boldsymbol{\mu}_l\|$ or all σ_l goes to infinity.

Property 1. Given $\mathbf{W}^* = \mathbf{U}\mathbf{V} \in \mathbb{R}^{d \times k}$, where $\mathbf{U} \in \mathbb{R}^{d \times k}$ is the orthogonal basis of \mathbf{W}^* . For any $\boldsymbol{\mu} \in \mathbb{R}^d$, we can find an orthogonal decomposition of $\boldsymbol{\mu}$ based on the column space of \mathbf{W}^* , i.e. $\boldsymbol{\mu} = \boldsymbol{\mu}_U + \boldsymbol{\mu}_{U^\perp}$. If we consider the recovery problem

of FCN with a dataset of Gaussian Mixture Model, in which $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ for some $h \in [L]$, the problem is equivalent to the problem of FCN with $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{U_h}, \boldsymbol{\Sigma}_h)$. Hence, we can assume without loss of generality that $\boldsymbol{\mu}_l$ belongs to the column space of \mathbf{W}^* for all $l \in [L]$.

Proof:

From (1) and (3), the recovery problem can be formulated as

$$\min_{\mathbf{W}^*} g(\mathbf{W}^{*\top} \mathbf{x}_i, y_i)$$

For any $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$, \mathbf{x}_i can be written as

$$\mathbf{x}_i = \mathbf{z} + \boldsymbol{\mu}_h$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_h)$. Therefore,

$$\mathbf{W}^{*\top} \mathbf{x}_i = \mathbf{W}^{*\top} (\mathbf{z} + \boldsymbol{\mu}_h) = \mathbf{W}^{*\top} (\mathbf{z} + \boldsymbol{\mu}_{U_h} + \boldsymbol{\mu}_{U_\perp h}) = \mathbf{W}^{*\top} (\mathbf{z} + \boldsymbol{\mu}_{U_h})$$

The final step is because $\mathbf{W}^{*\top} \boldsymbol{\mu}_{U_\perp} = \mathbf{0}$. So the problem is equivalent to the recovery problem of FCN with $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{U_h}, \boldsymbol{\Sigma}_h)$.

Recall that the gradient noise $\boldsymbol{\nu}_i \in \mathbb{R}^{d \times K}$ is zero-mean, and each of its entry is upper bounded by $\xi > 0$.

Property 2. We have that $\|\boldsymbol{\nu}_i\|_F$ is a sub-Gaussian random variable with its sub-Gaussian norm bounded by $\xi\sqrt{dK}$.

Proof:

$$(\mathbb{E}\|\boldsymbol{\nu}_i\|_F^p)^{\frac{1}{p}} \leq (\mathbb{E}|\sqrt{dK}\xi|^p)^{\frac{1}{p}} \leq \xi\sqrt{dK} \quad (67)$$

We state some general properties of the ρ function defined in Definition 1 in the following.

Property 3. $\rho(\mathbf{u}, \sigma)$ in Definition 1 satisfies the following properties,

- 1) **(Positive)** $\rho(\mathbf{u}, \sigma) > 0$ for any $\mathbf{u} \in \mathbb{R}^d$ and $\sigma \neq 0$.
- 2) **(Finite limit point for zero mean)** $\rho(\mathbf{u}, \sigma)$ converges to a positive value function of σ as u_i goes to 0, i.e. $\lim_{u_i \rightarrow 0} \rho(\mathbf{u}, \sigma) := \mathcal{C}_m(\sigma)$.
- 3) **(Finite limit point for zero variance)** When all $u_i \neq 0$ ($i \in [d]$), $\rho(\frac{\mathbf{u}}{\sigma}, \sigma)$ converges to a strictly positive real function of \mathbf{u} as σ goes to 0, i.e. $\lim_{\sigma \rightarrow 0} \rho(\frac{\mathbf{u}}{\sigma}, \sigma) := \mathcal{C}_s(\mathbf{u})$. When $u_i = 0$ for some $i \in [d]$, $\lim_{\sigma \rightarrow 0} \rho(\frac{\mathbf{u}}{\sigma}, \sigma) = 0$.
- 4) **(Lower bound function of the mean)** When everything else except $|u_i|$ is fixed, $\rho(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*))$ is lower bounded by a strictly positive real function, $\mathcal{L}_m(\frac{(\Delta \mathbf{W}^*)^\top \Delta \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*))$, which is monotonically decreasing as $|u_i|$ increases.
- 5) **(Lower bound function of the variance)** When everything else except σ is fixed, $\rho(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*))$ is lower bounded by a strictly positive real function, $\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*))$, which satisfies the following conditions: (a) there exists $\zeta_{s'} > 0$, such that $\sigma^{-1} \mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*))$ is an increasing function of σ when $\sigma \in (0, \zeta_{s'})$; (b) there exists $\zeta_s > 0$ such that $\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*))$ is a decreasing function of σ when $\sigma \in (\zeta_s, +\infty)$.

Proof:

(1) From Cauchy Schwarz's inequality, we have

$$\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'(\sigma \cdot z_i)] \leq \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'^2(\sigma \cdot z_i)]} \quad (68)$$

$$\begin{aligned} \mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'(\sigma \cdot z_i) z_i \cdot z_i] &\leq \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'^2(\sigma \cdot z_i) z_i^2]} \cdot \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[z_i^2]} \\ &= \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'^2(\sigma \cdot z_i) z_i^2]} \cdot \sqrt{u_i^2 + 1} \end{aligned} \quad (69)$$

The equalities of the (68) and (69) hold if and only if ϕ' is a constant function. Since that ϕ is the sigmoid function, the equalities of (68) and (69) cannot hold.

By the definition of $\rho(\mathbf{u}, \sigma)$ in Definition 1, we have

$$\beta_0(i, \mathbf{u}, \sigma) - \alpha_0^2(i, \mathbf{u}, \sigma) > 0, \quad (70)$$

$$\beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2^2(i, \mathbf{u}, \sigma)}{u_i^2 + 1} > 0. \quad (71)$$

Therefore,

$$\rho(\mathbf{u}, \sigma) > 0 \quad (72)$$

(2) We can derive that

$$\begin{aligned}
& \lim_{u_i \rightarrow 0} \left(\frac{u_j^2}{\sigma^2} + 1 \right) (\beta_0(i, \mathbf{u}, \sigma) - \alpha_0^2(i, \mathbf{u}, \sigma)) \\
&= \lim_{u_i \rightarrow 0} \left(\frac{u_j^2}{\sigma^2} + 1 \right) \left(\int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - u_i\|^2}{2}\right) dz_i \right. \\
&\quad \left. - \left(\int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - u_i\|^2}{2}\right) dz_i \right)^2 \right) \\
&= \left(\frac{u_j^2}{\sigma^2} + 1 \right) \left(\int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i\|^2}{2}\right) dz_i - \left(\int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i\|^2}{2}\right) dz_i \right)^2 \right),
\end{aligned} \tag{73}$$

where the first step is by Definition 1, and the second step comes from the limit laws. Similarly, we also have

$$\begin{aligned}
& \lim_{u_i \rightarrow 0} \left(\beta_2(i, \mathbf{u}, \sigma) - \frac{1}{u_i^2 + 1} \alpha_2^2(i, \mathbf{u}, \sigma) \right) \\
&= \lim_{u_i \rightarrow 0} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - u_i\|^2}{2}\right) dz_i \\
&\quad - \left(\frac{1}{u_i^2 + 1} \int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - u_i\|^2}{2}\right) dz_i \right)^2 \\
&= \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i\|^2}{2}\right) dz_i - \left(\int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i\|^2}{2}\right) dz_i \right)^2
\end{aligned} \tag{74}$$

Since that (73) and (74) are positive due to Jensen's inequality, we can derive that $\rho(\mathbf{u}, \sigma)$ converges to a positive value function of σ as u_i goes to 0, i.e.

$$\lim_{u \rightarrow 0} \rho(\mathbf{u}, \sigma) := \mathcal{C}_m(\sigma) \tag{75}$$

(3) When all $u_i \neq 0$ ($i \in [d]$),

$$\begin{aligned}
& \lim_{\sigma \rightarrow 0} \left(\beta_2(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \frac{1}{\frac{u_i^2}{\sigma^2} + 1} \alpha_2^2(i, \frac{\mathbf{u}}{\sigma}, \sigma) \right) \\
&= \lim_{\sigma \rightarrow 0} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - \frac{u_i}{\sigma}\|^2}{2}\right) dz_i \\
&\quad - \frac{1}{\frac{u_i^2}{\sigma^2} + 1} \left(\int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - \frac{u_i}{\sigma}\|^2}{2}\right) dz_i \right)^2 \\
&= \lim_{\sigma \rightarrow 0} \int_{-\infty}^{\infty} \phi'^2(u_i \cdot x_i) \frac{u_i^2}{\sigma^2} x_i^2 (2\pi \frac{\sigma^2}{u_i^2})^{-\frac{1}{2}} \exp\left(-\frac{\|x_i - 1\|^2}{2 \frac{\sigma^2}{u_i^2}}\right) dx_i \\
&\quad - \frac{1}{\frac{u_i^2}{\sigma^2} + 1} \left(\int_{-\infty}^{\infty} \phi'(u_i \cdot x_i) \frac{u_i^2}{\sigma^2} x_i^2 (2\pi \frac{\sigma^2}{u_i^2})^{-\frac{1}{2}} \exp\left(-\frac{\|x_i - 1\|^2}{2 \frac{\sigma^2}{u_i^2}}\right) dx_i \right)^2 \quad z_i = \frac{u_i}{\sigma} x_i \\
&= \lim_{\sigma \rightarrow 0} \phi'^2(u_i) \frac{u_i^2}{\sigma^2} - \frac{1}{\frac{u_i^2}{\sigma^2} + 1} \left(\phi'(u_i) \frac{u_i^2}{\sigma^2} \right)^2 \\
&= \lim_{\sigma \rightarrow 0} \phi'^2(u_i) \frac{u_i^2}{\sigma^2} \left(1 - \frac{\frac{u_i^2}{\sigma^2}}{1 + \frac{u_i^2}{\sigma^2}} \right) \\
&= \lim_{\sigma \rightarrow 0} \phi'^2(u_i) \frac{1}{1 + \frac{\sigma^2}{u_i^2}} \\
&= \phi'^2(u_i)
\end{aligned} \tag{76}$$

The first step of (76) comes from Definition 1. The second step and the last three steps are derived from some basic mathematical computation and the limit laws. The third step of (76) is by the fact that the Gaussian distribution goes to a

Dirac delta function when σ goes to 0. Then the integral will take the value when $x_i = 1$. Similarly, we can obtain the following

$$\begin{aligned}
& \lim_{\sigma \rightarrow 0} (\beta_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha_0^2(i, \frac{\mathbf{u}}{\sigma}, \sigma)) \\
&= \lim_{\sigma \rightarrow 0} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp(-\frac{\|z_i - \frac{u_i}{\sigma}\|^2}{2}) dz_i \\
&\quad - \left(\int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp(-\frac{\|z_i - \frac{u_i}{\sigma}\|^2}{2}) dz_i \right)^2 \\
&= \phi'^2(u_i) - \phi'^2(u_i) = 0
\end{aligned} \tag{77}$$

$$\begin{aligned}
& \lim_{\sigma \rightarrow 0} \left(\frac{\partial}{\partial \sigma} (\beta_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha_0^2(i, \frac{\mathbf{u}}{\sigma}, \sigma)) \right) \\
&= \lim_{\sigma \rightarrow 0} \left(\frac{\partial}{\partial \sigma} \left(\int_{-\infty}^{\infty} \phi'^2(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{\|x_i - u_i\|^2}{2\sigma^2}) dx_i \right. \right. \\
&\quad \left. \left. - \left(\int_{-\infty}^{\infty} \phi'(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{\|x_i - u_i\|^2}{2\sigma^2}) dx_i \right)^2 \right) \quad x_i = \sigma \cdot z_i \right) \\
&= \lim_{\sigma \rightarrow 0} \left(\int_{-\infty}^{\infty} \phi'^2(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{\|x_i - u_i\|^2}{2\sigma^2}) (-\sigma^{-1} + \|x_i - u_i\|^2 \sigma^{-2}) dx_i \right. \\
&\quad \left. - 2 \left(\int_{-\infty}^{\infty} \phi'(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{\|x_i - u_i\|^2}{2\sigma^2}) dx_i \right) \right. \\
&\quad \left. \cdot \int_{-\infty}^{\infty} \phi'(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\frac{\|x_i - u_i\|^2}{2\sigma^2}) (-\sigma^{-1} + \|x_i - u_i\|^2 \sigma^{-2}) dx_i \right) \\
&= \lim_{\sigma \rightarrow 0} \left(\frac{\phi'^2(u_i)}{-\sigma} - 2\phi'(u_i) \frac{\phi'(u_i)}{-\sigma} \right) \\
&= \lim_{\sigma \rightarrow 0} \frac{\phi'^2(u_i)}{\sigma} = +\infty
\end{aligned} \tag{78}$$

Therefore, by L'Hopital's rule and (77), (78), we have

$$\begin{aligned}
& \lim_{\sigma \rightarrow 0} \left(\frac{u_i^2}{\sigma^2} + 1 \right) (\beta_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha_0(i, \frac{\mathbf{u}}{\sigma}, \sigma)) \\
&= \lim_{\sigma \rightarrow 0} \frac{u_i^2}{2\sigma} \frac{\partial}{\partial \sigma} (\beta_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha_0(i, \frac{\mathbf{u}}{\sigma}, \sigma)) \\
&= +\infty
\end{aligned} \tag{79}$$

Combining (79) and (76), we can derive that $\rho(\frac{\mathbf{u}}{\sigma}, \sigma)$ converges to a positive value function of \mathbf{u} as σ goes to 0, i.e.

$$\lim_{\sigma \rightarrow 0} \rho\left(\frac{\mathbf{u}}{\sigma}, \sigma\right) := \mathcal{C}_s(\mathbf{u}). \tag{80}$$

When $u_i = 0$ for some $i \in [d]$, $\lim_{\sigma \rightarrow 0} \left(\frac{u_i^2}{\sigma^2} + 1 \right) (\beta_0(j, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha^2(j, \frac{\mathbf{u}}{\sigma}, \sigma)) = 0$ by (77). Then from the Definition 1, we have

$$\lim_{\sigma \rightarrow 0} \rho\left(\frac{\mathbf{u}}{\sigma}, \sigma\right) = 0 \tag{81}$$

(4) We can define $\mathcal{L}_m\left(\frac{(\Delta \mathbf{W}^*)^\top \Delta \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)$ as

$$\mathcal{L}_m\left(\frac{(\Delta \mathbf{W}^*)^\top \Delta \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) = \min_{v_i \in [0, u_i]} \left\{ \rho\left(\frac{(\Delta \mathbf{W}^*)^\top \Delta \mathbf{v}}{\sigma_l \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) : v_j = u_j \text{ for all } j \neq i \right\} \tag{82}$$

Then by this definition, we have

$$0 < \mathcal{L}_m\left(\frac{(\Delta \mathbf{W}^*)^\top \Delta \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) \leq \rho\left(\frac{(\Delta \mathbf{W}^*)^\top \Delta \mathbf{u}}{\sigma_l \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) \tag{83}$$

Meanwhile, for any $0 \leq u'_i \leq u_i^*$, since that $[0, u'_i] \subset [0, u_i^*]$, we can obtain

$$\mathcal{L}_m\left(\frac{(\Delta \mathbf{W}^*)^\top \Delta \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) \Big|_{u_i=u'_i} \geq \mathcal{L}_m\left(\frac{(\Delta \mathbf{W}^*)^\top \Delta \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) \Big|_{u_i=u_i^*} \tag{84}$$

Hence, $\mathcal{L}_m(\frac{(\mathbf{\Lambda}\mathbf{W}^*)^\top \mathbf{\Lambda}\mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$ is a strictly positive real function which is monotonically decreasing.

(5) Therefore, we only need to show the condition (a).

When $(\mathbf{W}^{*\top} \mathbf{u})_i \neq 0$ for all $i \in [K]$,

$$\lim_{\sigma \rightarrow 0} \rho(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*)) = \mathcal{C}_s(\mathbf{u}) > 0. \quad (85)$$

Therefore, there exists $\zeta_s > 0$, such that when $0 < \sigma < \zeta_s$,

$$\rho(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*)) > \frac{\mathcal{C}_s(\mathbf{W}^{*\top} \mathbf{u})}{2}. \quad (86)$$

Then we can define

$$\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*)) := \frac{\mathcal{C}_s(\mathbf{W}^{*\top} \mathbf{u})}{2\zeta_s} \sigma^2 \quad (87)$$

such that $\sigma^{-1}\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$ is an increasing function of σ below $\rho(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$.

When $(\mathbf{W}^{*\top} \mathbf{u})_i = 0$ for some $i \in [K]$, then

$$\lim_{\sigma \rightarrow 0} \rho(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*)) = 0. \quad (88)$$

We can define

$$\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*)) = \sigma \cdot \min_{v_i \in [u_i, \zeta_{s'}]} \left\{ \rho(\frac{\mathbf{W}^{*\top} \mathbf{v}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*)) : v_j \neq u_j \text{ for all } j \neq i \right\} \quad (89)$$

Then,

$$\sigma^{-1}\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*)) = \min_{v_i \in [u_i, \zeta_{s'}]} \left\{ \rho(\frac{\mathbf{W}^{*\top} \mathbf{v}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*)) : v_j = u_j \text{ for all } j \neq i \right\} \quad (90)$$

For any $0 \leq u'_i \leq u_i^* < \zeta_{s'}$, since that $[u_i^*, \zeta_{s'}] \subset [u'_i, \zeta_{s'}]$, we can obtain

$$\sigma^{-1}\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))|_{u_i=u'_i} \leq \sigma^{-1}\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))|_{u_i=u_i^*} \quad (91)$$

Therefore, we can derive that $\sigma^{-1}\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$ is monotonically increasing. Following the steps in (4), we can have that $\sigma^{-1}\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$ is a strictly positive real function which is upper bounded by $\rho(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$. In conclusion, condition (a) is proved.

For condition (b), since that $\zeta_s > 0$, $\rho(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$ is continuous and positive, we can obtain

$$\rho(\frac{\mathbf{W}^{*\top} \mathbf{v}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*)) \Big|_{\sigma=\zeta_s} > 0 \quad (92)$$

Then condition (b) can be easily proved as in (4).

We then characterize the order of the ρ function in different cases as follows.

Property 4. To specify the order with regard to the distribution parameters, $\rho(\mathbf{u}, \sigma)$ in Definition 1 satisfies the following properties,

- 1) (**Small variance**) $\lim_{\sigma \rightarrow 0^+} \rho(\mathbf{u}, \sigma) = \Theta(\sigma^4)$.
- 2) (**Large variance**) For any $\epsilon > 0$, $\lim_{\sigma \rightarrow \infty} \rho(\mathbf{u}, \sigma) \geq \Theta(\frac{1}{\sigma^{3+\epsilon}})$.
- 3) (**Large mean**) For any $\epsilon > 0$, $\lim_{\mu \rightarrow \infty} \rho(\mathbf{u}, \sigma) \geq \Theta(e^{-\frac{\|\mathbf{u}\|^2}{2}}) \frac{1}{\|\mathbf{u}\|^{3+\epsilon}}$.

Proof:

(1)

$$\begin{aligned}
& \beta_0(i, \mathbf{u}, \sigma) - \alpha_0(i, \mathbf{u}, \sigma)^2 \\
&= \mathbb{E}_{z \sim \mathcal{N}(\mu, 1)}[\phi'^2(\sigma \cdot z)] - (\mathbb{E}_{z \sim \mathcal{N}(\mu, 1)}[\phi'(\sigma \cdot z)])^2 \\
&= \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz - \left(\int_{-\infty}^{\infty} \phi'(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \right)^2 \\
&= \int_{-\infty}^{\infty} \left(\frac{1}{4} - \frac{t^2}{16} + \frac{t^4}{96} \cdots \right)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu\sigma)^2}{2\sigma^2}} dt \\
&\quad - \left(\int_{-\infty}^{\infty} \left(\frac{1}{4} - \frac{t^2}{16} + \frac{t^4}{96} + \cdots \right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu\sigma)^2}{2\sigma^2}} dt \right)^2 \\
&= \left(\frac{1}{16} - \frac{1}{32}(\mu^2\sigma^2 + \sigma^2) + \frac{7}{768}(3\sigma^4 + 6\mu^2\sigma^4 + \mu^4\sigma^4) + \cdots \right) \\
&\quad - \left(\frac{1}{4} - \frac{\mu^2\sigma^2 + \sigma^2}{16} + \frac{3\sigma^4 + 6\mu^2\sigma^4 + \mu^4\sigma^4}{192} + \cdots \right)^2 \\
&= \frac{1}{128}\sigma^4 + \frac{\mu^2\sigma^4}{64} + o(\sigma^4), \quad \text{as } \sigma \rightarrow 0^+.
\end{aligned} \tag{93}$$

The first step of (93) is by Definition 1. The second step and the last steps come from some basic mathematical computation. The third step is from Taylor expansion. Hence,

$$\lim_{\sigma \rightarrow 0^+} (\beta_0(i, \mathbf{u}, \sigma) - \alpha_0(i, \mathbf{u}, \sigma)^2) = \frac{1}{128}\sigma^4 + \frac{\mu^2\sigma^4}{64} + o(\sigma^4) \tag{94}$$

Similarly, we can obtain

$$\begin{aligned}
& \beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2(i, \mathbf{u}, \sigma)^2}{\mu^2 + 1} \\
&= \mathbb{E}_{z \sim \mathcal{N}(0, 1)}[\phi'^2(\sigma \cdot z)z^2] - \frac{(\mathbb{E}_{z \sim \mathcal{N}(0, 1)}[\phi'(\sigma \cdot z)z^2])^2}{\mu^2 + 1} \\
&= \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z)z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz - \frac{1}{\mu^2 + 1} \left(\int_{-\infty}^{\infty} \phi'(\sigma \cdot z)z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \right)^2 \\
&= \int_{-\infty}^{\infty} \left(\frac{t}{4\sigma} - \frac{t^3}{16\sigma} + \frac{t^5}{96\sigma} \cdots \right)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu\sigma)^2}{2\sigma^2}} dt \\
&\quad - \frac{1}{\mu^2 + 1} \left(\int_{-\infty}^{\infty} \left(\frac{t^2}{4\sigma^2} - \frac{t^4}{16\sigma^2} + \frac{t^6}{96\sigma^2} + \cdots \right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu\sigma)^2}{2\sigma^2}} dt \right)^2 \\
&= \left(\frac{1 + \mu^2}{16} - \frac{3\sigma^2 + 6\mu^2\sigma^2 + \mu^4\sigma^2}{32} + \cdots \right) \\
&\quad - \frac{1}{\mu^2 + 1} \left(\frac{1 + \mu^2}{4} - \frac{15\sigma^2 + 45\mu^2\sigma^2 + 15\mu^4\sigma^2 + \mu^6\sigma^2}{32} + \cdots \right)^2 \\
&= \frac{9}{64}\sigma^2 + \frac{33}{64}\mu^2\sigma^2 + \frac{13}{64}\mu^4\sigma^2 + \frac{1}{64}\mu^6\sigma^2 + o(\sigma^2), \quad \text{as } \sigma \rightarrow 0^+
\end{aligned} \tag{95}$$

Hence,

$$\lim_{\sigma \rightarrow 0^+} (\beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2(i, \mathbf{u}, \sigma)^2}{\mu^2 + 1}) = \frac{9}{64}\sigma^2 + o(\sigma^2) \tag{96}$$

Therefore,

$$\lim_{\sigma \rightarrow 0^+} \rho(\mathbf{u}, \sigma) = \min_{j \in [d], u_j \neq \mu} \{(u_j^2 + 1)\} \frac{1}{128}\sigma^4 \tag{97}$$

(2) Note that by some basic mathematical derivation,

$$\begin{aligned}
\int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{1}{(e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\
&\geq 2 \int_0^{\infty} \frac{1}{16e^{2\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z+|\mu|)^2}{2}} dz \\
&= \frac{1}{8} e^{2|\mu|\sigma + 2\sigma^2} \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z+2\sigma)^2}{2}} dz \\
&= \frac{1}{8\sqrt{2\pi}} e^{2|\mu|\sigma + 2\sigma^2} \int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt
\end{aligned} \tag{98}$$

We then provide the following Claim with its proof to give a lower bound for (98).

Claim: $\int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt > e^{-2|\mu|\sigma-2\sigma^2-k_1 \log \sigma}$ for $k_1 > 1$.

Proof: Let

$$f(\sigma) = \int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt - e^{-2|\mu|\sigma-2\sigma^2-k_1 \log \sigma}. \quad (99)$$

Then,

$$f'(\sigma) = e^{-2\sigma^2} \left((2|\mu| + 4\sigma + \frac{k_1}{\sigma}) \sigma^{-k_1} - 2e^{-\frac{1}{2}\mu^2} \right). \quad (100)$$

It can be easily verified that for a given $|\mu| \geq 0$, $f'(\sigma) < 0$ when σ is large enough if $k_1 > 1$. Combining that $\lim_{\sigma \rightarrow \infty} f(\sigma) = 0$, we have $f(\sigma) > 0$ when σ is large enough by showing the contradiction in the following:

Suppose there is a strictly increasing function $f(x) > 0$ with $\lim_{x \rightarrow \infty} f(x) = 0$ when x is large enough. Then there exists $x_0 > 0$ such that for any $\epsilon > 0$, $f(x) < \epsilon$ for $x > x_0$. Pick $\epsilon = f(x_0) > 0$, then for $x_1 > x_0$, $f(x_1) > f(x_0) = \epsilon$. Contradiction!

Similarly, we also have

$$\begin{aligned} \int_{-\infty}^{\infty} \phi'(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{1}{e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &\leq 2 \int_0^{\infty} \frac{1}{e^{\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &= e^{|\mu|\sigma + \frac{1}{2}\sigma^2} \int_0^{\infty} \frac{2}{\sqrt{2\pi}} e^{-\frac{(z+|\mu|+\sigma)^2}{2}} dz \\ &= \frac{2}{\sqrt{2\pi}} e^{|\mu|\sigma + \frac{1}{2}\sigma^2} \int_{|\mu|+\sigma}^{\infty} e^{-\frac{t^2}{2}} dt, \end{aligned} \quad (101)$$

and the **Claim:** $\int_{|\mu|+\sigma}^{\infty} e^{-\frac{t^2}{2}} dt < e^{-|\mu|\sigma - \frac{1}{2}\sigma^2 - k_2 \log \sigma}$ for $k_2 \leq 1$ to give an upper bound for (101).

Therefore, combining (98, 101) and two claims, we have that for any $\epsilon > 0$,

$$\beta_0(i, \mathbf{u}, \sigma) - \alpha_0(i, \mathbf{u}, \sigma)^2 \geq \frac{1}{8\sqrt{2\pi}} \frac{1}{\sigma^{k_1}} - \frac{1}{2\pi} \frac{1}{\sigma^{2k_2}} \gtrsim \frac{1}{\sigma^{1+\epsilon}} \quad (102)$$

(The above inequality holds for any $2k_2 > k_1$ where $k_1 > 1$ and $k_2 \leq 1$.)

Similarly,

$$\begin{aligned} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{z^2}{(e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &\geq 2 \int_0^{\infty} \frac{z^2}{16e^{2\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z+|\mu|)^2}{2}} dz \\ &= \frac{1}{8\sqrt{2\pi}} e^{|\mu|\sigma + 2\sigma^2} \int_{2|\mu|+2\sigma}^{\infty} (t-2\sigma)^2 e^{-\frac{t^2}{2}} dt \end{aligned} \quad (103)$$

Claim: $\int_{|\mu|+2\sigma}^{\infty} (t-2\sigma)^2 e^{-\frac{t^2}{2}} dt \geq e^{-2|\mu|\sigma-2\sigma^2-k_1 \log \sigma}$ if $k_1 > 3$.

Proof: Let

$$f(\sigma) = \int_{|\mu|+2\sigma}^{\infty} (t-2\sigma)^2 e^{-\frac{t^2}{2}} dt - e^{-2|\mu|\sigma-2\sigma^2-k_1 \log \sigma}. \quad (104)$$

$$f'(\sigma) = 8\sigma \int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt + e^{-2|\mu|\sigma-2\sigma^2} (4\sigma^{1-k_1} + k_1\sigma^{-1-k_1} + 2|\mu|\sigma^{-k_1} - 4e^{-\frac{1}{2}\mu^2}). \quad (105)$$

We need $f'(\sigma) < 0$ when σ is large enough. Since that $f'(\sigma) \rightarrow 0$, $f''(\sigma) \rightarrow 0$ when σ is large, we need $f''(\sigma) > 0$ and $f'''(\sigma) < 0$ recursively. Hence,

$$\begin{aligned} f'''(\sigma) &= e^{-2|\mu|\sigma-2\sigma^2} (64\sigma^{3-k_1} + 96\mu\sigma^{2-k_1} + 16(3k_1-3+\mu^2)\sigma^{1-k_1} + 8\mu(-\mu^2-3+6k_1)\sigma^{-k_1} \\ &\quad + 4k_1(3k_1+\mu^2)\sigma^{-1-k_1} + 2k_1(1+k_1)(\mu+2)\sigma^{-2-k_1} \\ &\quad + k_1(1+k_1)(2+k_1)\sigma^{-3-k_1} - 16e^{-\frac{1}{2}\mu^2}) < 0 \end{aligned} \quad (106)$$

requires $k_1 > 3$.

Similarly, we have

$$\int_{-\infty}^{\infty} \phi'(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \leq 2 \int_0^{\infty} \frac{1}{e^{\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} z^2 e^{-\frac{z^2}{2}} dz = \frac{2}{\sqrt{2\pi}} e^{\frac{1}{2}\sigma^2} \int_{\sigma}^{\infty} (t-\sigma)^2 e^{-\frac{t^2}{2}} dt \quad (107)$$

and the **Claim**: $\int_{\sigma}^{\infty} (t - \sigma)^2 e^{-\frac{t^2}{2}} dt < e^{-\frac{\sigma^2}{2} - k_2 \log \sigma}$. Hence,

$$\beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2(i, \mathbf{u}, \sigma)^2}{\mu^2 + 1} \geq \frac{1}{8\sqrt{2\pi}} \frac{1}{\sigma^{k_1}} - \frac{2}{\pi(\mu^2 + 1)} \frac{1}{\sigma^{2k_2}} \gtrsim \frac{1}{\sigma^{3.1}} \quad (108)$$

(The above inequality holds for any $2k_2 > k_1$ where $k_1 > 3$ and $k_2 < 3$.)

Therefore, by combining (102) and (108), for any $\epsilon > 0$

$$\lim_{\sigma \rightarrow \infty} \rho(\mathbf{u}, \sigma) \geq \Theta\left(\frac{1}{\sigma^{3+\epsilon}}\right). \quad (109)$$

(3) Let σ be fixed. For any $\epsilon > 0$, following the steps in (2), we can obtain

$$\begin{aligned} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{1}{(e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &\geq 2 \int_0^{\infty} \frac{1}{16e^{2\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z+|\mu|)^2}{2}} dz \\ &= \frac{1}{8\sqrt{2\pi}} e^{2|\mu|\sigma + 2\sigma^2} \int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt \\ &\geq \frac{1}{8\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \frac{1}{\mu^{1+\epsilon}} \end{aligned} \quad (110)$$

$$\begin{aligned} \int_{-\infty}^{\infty} \phi'(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{1}{e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &\leq 2 \int_0^{\infty} \frac{1}{e^{\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &= \frac{2}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \frac{1}{\mu^{1-\epsilon}} \end{aligned} \quad (111)$$

Similarly,

$$\int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \geq \frac{1}{8\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \frac{1}{\mu^{3+\epsilon}} \quad (112)$$

$$\int_{-\infty}^{\infty} \phi'(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \leq \frac{2}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \frac{1}{\mu^{3-\epsilon}} \quad (113)$$

We can conclude that $\lim_{\mu \rightarrow \infty} \rho(\mathbf{u}, \sigma) \geq \Theta\left(e^{-\frac{\|\mathbf{u}\|^2}{2}}\right) \frac{1}{\|\mathbf{u}\|^{3+\epsilon}}$.

Property 5. If a function $f(\mathbf{x})$ is an even function, then

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] \quad (114)$$

Proof:

Denote

$$g(\mathbf{x}) = f(\mathbf{x})(2\pi|\boldsymbol{\Sigma}|^2)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \quad (115)$$

By some basic mathematical computation,

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] &= \int_{\mathbf{x} \in \mathbb{R}^d} g(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_d) dx_1 \cdots dx_d \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2, \dots, x_d) d(-x_1) dx_2 \cdots dx_d \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(-x_1, x_2, \dots, x_d) dx_1 dx_2 \cdots dx_d \\ &= \int_{\mathbf{x} \in \mathbb{R}^d} g(-\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})(2\pi|\boldsymbol{\Sigma}|^2)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} + \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} + \boldsymbol{\mu})\right) \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] \end{aligned} \quad (116)$$

Therefore, we have

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] \quad (117)$$

Property 6. Under Gaussian Mixture Model $\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ where $\boldsymbol{\Sigma}_l = \text{diag}(\sigma_{l1}^2, \dots, \sigma_{ld}^2)$, we have the following upper bound.

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{u}^\top \mathbf{x})^{2t}] \leq (2t-1)!! \|\mathbf{u}\|^{2t} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \quad (118)$$

Proof:

Note that

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{u}^\top \mathbf{x})^{2t}] = \sum_{l=1}^L \lambda_l \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{u}^\top \mathbf{x})^{2t}] = \sum_{l=1}^L \lambda_l \mathbb{E}_{y \sim \mathcal{N}(\mathbf{u}^\top \boldsymbol{\mu}_l, \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u})} [y^{2t}], \quad (119)$$

where the last step is by that $\mathbf{u}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{u}^\top \boldsymbol{\mu}_l, \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u})$ for $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$. By some basic mathematical computation, we know

$$\begin{aligned} & \mathbb{E}_{y \sim \mathcal{N}(\mathbf{u}^\top \boldsymbol{\mu}_l, \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u})} [y^{2t}] \\ &= \int_{-\infty}^{\infty} (y - \mathbf{u}^\top \boldsymbol{\mu}_l + \mathbf{u}^\top \boldsymbol{\mu}_l)^{2t} \frac{1}{\sqrt{2\pi \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}}} e^{-\frac{(y - \mathbf{u}^\top \boldsymbol{\mu}_l)^2}{2\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}}} dy \\ &= \int_{-\infty}^{\infty} \sum_{p=0}^{2t} \binom{2t}{p} (\mathbf{u}^\top \boldsymbol{\mu}_l)^{2t-p} (y - \mathbf{u}^\top \boldsymbol{\mu}_l)^p \frac{1}{\sqrt{2\pi \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}}} e^{-\frac{(y - \mathbf{u}^\top \boldsymbol{\mu}_l)^2}{2\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}}} dy \\ &= \sum_{p=0}^{2t} \binom{2t}{p} (\mathbf{u}^\top \boldsymbol{\mu}_l)^{2t-p} \cdot \begin{cases} 0, & p \text{ is odd} \\ (p-1)!! (\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u})^{\frac{p}{2}}, & p \text{ is even} \end{cases} \quad (120) \\ &\leq \sum_{p=0}^{2t} \binom{2t}{p} |\mathbf{u}^\top \boldsymbol{\mu}_l|^{2t-p} (p-1)!! |\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}|^{\frac{p}{2}} \\ &\leq (2t-1)!! (|\mathbf{u}^\top \boldsymbol{\mu}_l| + |\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}|^{\frac{1}{2}})^{2t} \\ &\leq (2t-1)!! \|\mathbf{u}\|^{2t} (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t}, \end{aligned}$$

where the second step is by the Binomial theorem. Hence,

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{u}^\top \mathbf{x})^{2t}] \leq (2t-1)!! \|\mathbf{u}\|^{2t} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \quad (121)$$

Property 7. With the Gaussian Mixture Model, we have

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\|\mathbf{x}\|^{2t}] \leq d^t (2t-1)!! \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \quad (122)$$

Proof:

$$\begin{aligned}
& \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\|\mathbf{x}\|_2^{2t}] \\
&= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^d x_i^2 \right)^t \right] \\
&= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[d^t \left(\sum_{i=1}^d \frac{x_i^2}{d} \right)^t \right] \\
&\leq \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[d^t \sum_{i=1}^d \frac{x_i^{2t}}{d} \right] \\
&= d^{t-1} \sum_{i=1}^d \sum_{j=1}^L \int_{-\infty}^{\infty} (x_i - \mu_{ji} + \mu_{ji})^{2t} \lambda_j \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(x_i - \mu_{ji})^2}{2\sigma_{ji}^2}\right) dx_i \\
&= d^{t-1} \sum_{i=1}^d \sum_{j=1}^L \sum_{k=1}^{2t} \binom{2t}{k} \lambda_j |\mu_{ji}|^{2t-k} \cdot \begin{cases} 0, & k \text{ is odd} \\ (k-1)!! \sigma_{ji}^k, & k \text{ is even} \end{cases} \\
&\leq d^{t-1} \sum_{i=1}^d \sum_{j=1}^L \sum_{k=1}^{2t} \binom{2t}{k} \lambda_j |\mu_{ji}|^{2t-k} \sigma_j^k \cdot (2t-1)!! \\
&= d^{t-1} \sum_{i=1}^d \sum_{j=1}^L \lambda_j (|\mu_{ji}| + \sigma_{ji})^{2t} (2t-1)!! \\
&\leq d^t (2t-1)!! \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t}
\end{aligned} \tag{123}$$

In the 3rd step, we apply Jensen inequality because $f(x) = x^t$ is convex when $x \geq 0$ and $t \geq 1$. In the 4th step we apply the Binomial theorem and the result of k-order central moment of Gaussian variable.

Property 8. Under the Gaussian Mixture Model $\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ where $\boldsymbol{\Sigma}_l = \boldsymbol{\Lambda}_l^\top \mathbf{D}_l \boldsymbol{\Lambda}_l$, we have the following upper bound.

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{u}^\top \mathbf{x})^{2t}] \leq (2t-1)!! \|\mathbf{u}\|^{2t} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \tag{124}$$

Proof:

If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$, then $\mathbf{u}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{u}^\top \boldsymbol{\mu}_l, \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}) = \mathcal{N}((\boldsymbol{\Lambda}_l \mathbf{u})^\top \boldsymbol{\Lambda}_l \boldsymbol{\mu}_l, (\boldsymbol{\Lambda}_l \mathbf{u})^\top \mathbf{D}_l (\boldsymbol{\Lambda}_l \mathbf{u}))$. By Property 6, we have

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{u}^\top \mathbf{x})^{2t}] \leq (2t-1)!! \|\mathbf{u}\|^{2t} (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \tag{125}$$

Then we can derive the final result.

Property 9. The population risk function $\bar{f}(\mathbf{W})$ is defined as

$$\begin{aligned}
\bar{f}(\mathbf{W}) &= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [f_n(\mathbf{W})] \\
&= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\frac{1}{n} \sum_{i=1}^n \ell(\mathbf{W}; \mathbf{x}_i, y_i) \right] \\
&= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\ell(\mathbf{W}; \mathbf{x}_i, y_i)]
\end{aligned} \tag{126}$$

For any permutation matrix \mathbf{P} , where $\{\pi(j)\}_{j=1}^K$ is the indices permuted by \mathbf{P} , we have

$$\begin{aligned}
H(\mathbf{W}\mathbf{P}, \mathbf{x}) &= \frac{1}{K} \sum_{\pi^*(j)} \phi(\mathbf{w}_{\pi(j)}^\top \mathbf{x}) \\
&= \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^\top \mathbf{x}) \\
&= H(\mathbf{W}, \mathbf{x})
\end{aligned} \tag{127}$$

Therefore,

$$\bar{f}(\mathbf{W}) = \bar{f}(\mathbf{W}\mathbf{P}) \tag{128}$$

Based on (1) and (3), we can derive its gradient and Hessian as follows.

$$\frac{\partial \ell(\mathbf{W}; \mathbf{x}, y)}{\partial \mathbf{w}_j} = -\frac{1}{K} \frac{y - H(\mathbf{W})}{H(\mathbf{W})(1 - H(\mathbf{W}))} \phi'(\mathbf{w}_j^\top \mathbf{x}) \mathbf{x} = \zeta(\mathbf{W}) \cdot \mathbf{x} \quad (129)$$

$$\frac{\partial^2 \ell(\mathbf{W}; \mathbf{x}, y)}{\partial \mathbf{w}_j \partial \mathbf{w}_l} = \xi_{j,l} \cdot \mathbf{x} \mathbf{x}^\top \quad (130)$$

$$\xi_{j,l}(\mathbf{W}) = \begin{cases} \frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \frac{H(\mathbf{W})^2 + y - 2y \cdot H(\mathbf{W})}{H^2(\mathbf{W})(1 - H(\mathbf{W}))^2}, & j \neq l \\ \frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \frac{H(\mathbf{W})^2 + y - 2y \cdot H(\mathbf{W})}{H^2(\mathbf{W})(1 - H(\mathbf{W}))^2} - \frac{1}{K} \phi''(\mathbf{w}_j^\top \mathbf{x}) \frac{y - H(\mathbf{W})}{H(\mathbf{W})(1 - H(\mathbf{W}))}, & j = l \end{cases} \quad (131)$$

Property 10. With $D_m(\Psi)$ defined in definition 2, we have

$$(i) \quad D_m(\Psi) D_{2m}(\Psi) \leq D_{3m}(\Psi) \quad (132)$$

$$(ii) \quad (D_m(\Psi))^2 \leq D_{2m}(\Psi) \quad (133)$$

Proof:

To prove (132), we can first compare the terms $\sum_{i=1}^L \lambda_i a_i \sum_{i=1}^L \lambda_i a_i^2$ and $\sum_{i=1}^L \lambda_i a_i^3$, where $a_i \geq 1$, $i \in [L]$ and $\sum_{i=1}^L \lambda_i = 1$.

$$\begin{aligned} \sum_{i=1}^L \lambda_i a_i^3 - \sum_{i=1}^L \lambda_i a_i \sum_{i=1}^L \lambda_i a_i^2 &= \sum_{i=1}^L \lambda_i a_i \cdot (a_i^2 - \sum_{j=1}^L \lambda_j a_j^2) \\ &= \sum_{i=1}^L \lambda_i a_i \cdot \left((1 - \lambda_i) a_i^2 - \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_j^2 \right) \\ &= \sum_{i=1}^L \lambda_i a_i \cdot \left(\sum_{1 \leq j \leq L, j \neq i} \lambda_j a_i^2 - \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_j^2 \right) \\ &= \sum_{i=1}^L \lambda_i a_i \cdot \left(\sum_{1 \leq j \leq L, j \neq i} \lambda_j (a_i^2 - a_j^2) \right) \\ &= \sum_{1 \leq i, j \leq L, i \neq j} (\lambda_i \lambda_j a_i (a_i^2 - a_j^2) + \lambda_i \lambda_j a_j (a_j^2 - a_i^2)) \\ &= \sum_{1 \leq i, j \leq L, i \neq j} \lambda_i \lambda_j (a_i - a_j)^2 (a_i + a_j) \geq 0 \end{aligned} \quad (134)$$

The second to last step is because we can find the pairwise terms $\lambda_i a_i \cdot \lambda_j (a_i^2 - a_j^2)$ and $\lambda_j a_j \cdot \lambda_i (a_j^2 - a_i^2)$ in the summation that can be putted together. From (134), we can obtain

$$\sum_{i=1}^L \lambda_i a_i \sum_{i=1}^L \lambda_i a_i^2 \leq \sum_{i=1}^L \lambda_i a_i^3 \quad (135)$$

Combining (135) and the definition of $D_m(\Psi)$ in (2), we can derive (132).

Similarly, to prove (133), we can first compare the terms $(\sum_{i=1}^L \lambda_i a_i)^2$ and $\sum_{i=1}^L \lambda_i a_i^2$, where $a_i \geq 1$, $i \in [L]$ and $\sum_{i=1}^L \lambda_i = 1$.

$$\begin{aligned} \sum_{i=1}^L \lambda_i a_i^2 - \left(\sum_{i=1}^L \lambda_i a_i \right)^2 &= \sum_{i=1}^L \lambda_i a_i \cdot \left(a_i - \sum_{j=1}^L \lambda_j a_j \right) \\ &= \sum_{i=1}^L \lambda_i a_i \cdot \left((1 - \lambda_i) a_i - \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_j \right) \\ &= \sum_{i=1}^L \lambda_i a_i \cdot \left(\sum_{1 \leq j \leq L, j \neq i} \lambda_j a_i - \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_j \right) \\ &= \sum_{i=1}^L \lambda_i a_i \cdot \left(\sum_{1 \leq j \leq L, j \neq i} \lambda_j (a_i - a_j) \right) \\ &= \sum_{1 \leq i, j \leq L, i \neq j} (\lambda_i \lambda_j a_i (a_i - a_j) + \lambda_i \lambda_j a_j (a_j - a_i)) \\ &= \sum_{1 \leq i, j \leq L, i \neq j} \lambda_i \lambda_j (a_i - a_j)^2 \geq 0 \end{aligned} \quad (136)$$

The derivation of (136) is close to (134). By (136) we have

$$\left(\sum_{i=1}^L \lambda_i a_i\right)^2 \leq \sum_{i=1}^L \lambda_i a_i^2 \quad (137)$$

Combining (137) and the definition of $D_m(\Psi)$ in (2), we can derive (133).

K. Proof of Theorem 1 and Corollary 1

Theorem 1 is built upon **three lemmas**.

Lemma 1 shows that with $O(dK^5 \log^2 d)$ samples, the empirical risk function is strongly convex in the neighborhood of \mathbf{W}^* .

Lemma 2 shows that if initialized in the convex region, the gradient descent algorithm converges linearly to a critical point $\widehat{\mathbf{W}}_n$, which is close to \mathbf{W}^* .

Lemma 3 shows that the Tensor Initialization Method in Subroutine 1 initializes $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$ in the local convex region. Theorem 1 follows naturally by combining these three lemmas.

This proving approach is built upon those in [28]. One of our major technical contribution is extending Lemmas 1 and 2 to the Gaussian mixture model, while the results in [28] only apply to Standard Gaussian models. The second major contribution is a new tensor initialization method for Gaussian mixture model such that the initial point is in the convex region (see Lemma 3). Both contributions require the development of new tools, and our analyses are much more involved than those for the standard Gaussian due to the complexity introduced by the Gaussian mixture model.

To present these lemmas, the Euclidean ball $\mathbb{B}(\mathbf{W}^* \mathbf{P}^*, r)$ is used to denote the neighborhood of $\mathbf{W}^* \mathbf{P}^*$, where r is the radius of the ball.

$$\mathbb{B}(\mathbf{W}^* \mathbf{P}^*, r) = \{\mathbf{W} \in \mathbb{R}^{d \times K} : \|\mathbf{W} - \mathbf{W}^* \mathbf{P}^*\|_F \leq r\} \quad (138)$$

The radius of the convex region is

$$r := \Theta\left(\frac{C_3 \epsilon_0 \cdot \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)}{K^{\frac{7}{2}} \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8\right)^{\frac{1}{4}}}\right) \quad (139)$$

with some constant $C_3 > 0$.

Proof of Theorem 1

From Lemma 2 and Lemma 3, we know that if n is sufficiently large such that the initialization \mathbf{W}_0 by the tensor method is in the region $\mathbb{B}(\mathbf{W}^* \mathbf{P}^*, r)$, then the gradient descent method converges to a critical point $\widehat{\mathbf{W}}_n$ that is sufficiently close to \mathbf{W}^* . To achieve that, one sufficient condition is

$$\begin{aligned} \|\mathbf{W}_0 - \mathbf{W}^* \mathbf{P}^*\|_F &\leq \sqrt{K} \|\mathbf{W}_0 - \mathbf{W}^* \mathbf{P}^*\| \leq \kappa^6 K^{\frac{7}{2}} \cdot \tau^6 \sqrt{D_6(\Psi)} \sqrt{\frac{d \log n}{n}} \|\mathbf{W}^* \mathbf{P}^*\| \\ &\leq \frac{C_3 \epsilon_0 \Gamma(\Psi) \sigma_{\max}^2}{K^{\frac{7}{2}} \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8\right)^{\frac{1}{4}}} \end{aligned} \quad (140)$$

where the first inequality follows from $\|\mathbf{W}\|_F \leq \sqrt{K} \|\mathbf{W}\|$ for $\mathbf{W} \in \mathbb{R}^{d \times K}$, the second inequality comes from Lemma 3, and the third inequality comes from the requirement to be in the region $\mathbb{B}(\mathbf{W}^* \mathbf{P}^*, r)$. That is equivalent to the following condition

$$\begin{aligned} n &\geq C_0 \epsilon_0^{-2} \cdot \tau^{12} \kappa^{12} K^{14} \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8\right)^{\frac{1}{2}} \\ &\quad \cdot (\delta_1(\mathbf{W}^*))^2 D_6(\Psi) \Gamma(\Psi)^{-2} \sigma_{\max}^{-4} \cdot d \log^2 d \end{aligned} \quad (141)$$

where $C_0 = \max\{C_4, C_3^{-2}\}$. By Definition 2, we can obtain

$$\left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8\right)^{\frac{1}{2}} \leq \sqrt{D_4(\Psi) D_8(\Psi)} \sigma_{\max}^6 \quad (142)$$

From Property 10, we have that

$$\begin{aligned} &\sqrt{D_4(\Psi) D_8(\Psi)} D_6(\Psi) \\ &\leq \sqrt{D_{12}(\Psi)} \sqrt{D_{12}(\Psi)} = D_{12}(\Psi) \end{aligned} \quad (143)$$

Plugging (142), (143) into (141), we have

$$n \geq C_0 \epsilon_0^{-2} \cdot \kappa^{12} K^{14} (\sigma_{\max} \delta_1(\mathbf{W}^*))^2 \tau^{12} \Gamma(\Psi)^{-2} D_{12}(\Psi) \cdot d \log^2 d \quad (144)$$

Considering the requirements on the sample complexity in (15), (42), and (144), (144) shows a sufficient number of samples. Taking the union bound of all the failure probabilities in Lemma 1, and 3, (144) holds with probability $1 - d^{-10}$.

By Property 3.4, $\rho(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ can be lower bounded by positive and monotonically decreasing functions $\mathcal{L}_m(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ when everything else except $|\tilde{\boldsymbol{\mu}}_{l(i)}|$ is fixed, or $\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ when everything else except $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$ is fixed. Then, by replacing the lower bound of $\rho(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ with these two functions in $\Gamma(\Psi)$, we can have an upper bound of $(\sigma_{\max} \delta_1(\mathbf{W}^*))^2 \tau^{12} \Gamma(\Psi)^{-2} D_{12}(\Psi)$, denoted as $\mathcal{B}(\Psi)$.

To be more specific, when everything else except $|\tilde{\boldsymbol{\mu}}_{l(i)}|$ is fixed, $\mathcal{L}_m(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ is plugged in $\mathcal{B}(\Psi)$. Then since that $D_{12}(\Psi)$ and $\mathcal{L}_m(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ are both increasing function of $|\tilde{\boldsymbol{\mu}}_{l(i)}|$, $\mathcal{B}(\Psi)$ is an increasing function of $|\tilde{\boldsymbol{\mu}}_{l(i)}|$.

When everything else except $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$ is fixed, if $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\| = \sigma_{\max} > \zeta_s$, then $\sigma_{\max}^2 \tau^{12} D_{12}(\Psi)$ is an increasing function of $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$. Since that $\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ is a decreasing function, $\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})^{-2}$ is an increasing function of $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$. Hence, $\mathcal{B}(\Psi)$ is an increasing function of $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$. Moreover, when all $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\| < \zeta_s$ and go to 0, two decreasing functions of $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$, $\sigma_{\max}^2 \mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})^{-2}$ and $D_{12}(\Psi)$ will be the dominant term of $\mathcal{B}(\Psi)$. Therefore, $\mathcal{B}(\Psi)$ increases to infinity as all $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$'s go to 0. In sum, we can define a universe $\mathcal{B}(\Psi)$ as:

$$\mathcal{B}(\Psi) = \begin{cases} (\sigma_{\max} \delta_1(\mathbf{W}^*))^2 \tau^{12} \left(\sum_{l=1}^L \frac{\lambda_l \|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \sigma_{\max}^2} \mathcal{L}_m(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}) \right)^{-2} \\ \cdot D_{12}(\Psi), \text{ if } \mathbf{S} \text{ is fixed} \\ (\sigma_{\max} \delta_1(\mathbf{W}^*))^2 \tau^{12} \left(\sum_{l=1}^L \frac{\lambda_l \|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \sigma_{\max}^2} \mathcal{L}_s(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}) \right)^{-2} \\ \cdot D_{12}(\Psi), \text{ if } \mathbf{M} \text{ is fixed} \\ (\sigma_{\max} \delta_1(\mathbf{W}^*))^2 \tau^{12} \left(\sum_{l=1}^L \frac{\lambda_l \|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \sigma_{\max}^2} \rho(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}) \right)^{-2} \\ \cdot D_{12}(\Psi), \text{ otherwise} \end{cases} \quad (145)$$

where \mathcal{L}_m , \mathcal{L}_s and D_{12} are defined in (84), (89) and Definition 2, respectively.

Hence, we have

$$n \geq \text{poly}(\epsilon_0^{-1}, \kappa, \eta, \tau K) \mathcal{B}(\Psi) \cdot d \log^2 d \quad (146)$$

Similarly, by replacing $\rho(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ with $\mathcal{L}_m(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ when everything else except $|\tilde{\boldsymbol{\mu}}_{l(i)}|$ is fixed, or $\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ (or $\|\boldsymbol{\Sigma}_l^{-1}\| \mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ for $\|\boldsymbol{\Sigma}_l^{-1}\|^{-1} \geq 1$) when everything else except $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$ is fixed, (29) can also be transferred to another feasible upper bound. We denote the modified version of the convergence rate as $v = 1 - K^{-2} q(\Psi)$. Since that $q(\Psi)$ is a ratio between the smallest and the largest singular value of $\nabla^2 \bar{f}(\mathbf{W}^*)$, we have $q(\Psi) \in (0, 1)$. Hence, we can obtain $1 - K^{-2} q(\Psi) \in (0, 1)$ by $K \geq 1$. When everything else except $|\tilde{\boldsymbol{\mu}}_{l(i)}|$ is fixed, since that $\mathcal{L}_m(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ is monotonically decreasing and $\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2$ is increasing as $|\tilde{\boldsymbol{\mu}}_{l(i)}|$ increases, v is an increasing function of $|\tilde{\boldsymbol{\mu}}_{l(i)}|$ to 1. Similarly, when everything else except $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$ is fixed where $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\| \geq \max\{1, \zeta_s\}$, $\frac{1}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}$ decreases to 0 as $\|\boldsymbol{\Sigma}_l\|$ increases. We replace $\rho(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ by $\|\boldsymbol{\Sigma}_l^{-1}\| \mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$ and then

$$\begin{aligned} & \|\boldsymbol{\Sigma}_l^{-1}\|^{-1} \cdot \|\boldsymbol{\Sigma}_l^{-1}\| \mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}) \\ & = \mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}) \end{aligned} \quad (147)$$

is an decreasing function less than $\rho(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$. Therefore, v is an increasing function of $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|$ to 1 when $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\| \geq \max\{1, \zeta_s\}$. When everything else except all $\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\| \leq \zeta_{s'}$'s go to 0, all $\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})$'s will decrease and all $\frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}$'s will decrease to 0. Therefore, v increases to 1.

$q(\Psi)$ can then be defined as

$$q(\Psi) = \begin{cases} \Omega\left(\frac{\sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \mathcal{L}_m\left(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}\right), \\ \text{if } \mathbf{S} \text{ is fixed} \\ \Omega\left(\frac{\sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}\right), \\ \text{if } \mathbf{M} \text{ is fixed and all } \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\| \leq \zeta_{s'} \\ \Omega\left(\frac{\lambda_l \frac{1}{\eta \tau^K \kappa^2} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_i}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_i^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_i^{-1}\|^{-\frac{1}{2}}\right) + \sum_{l \neq i} r(\lambda_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \mathbf{W}^*)}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}\right), \\ \text{if } \mathbf{M} \text{ is fixed and one } \|\boldsymbol{\Sigma}_i^{\frac{1}{2}}\| \geq \max\{1, \zeta_s\} \\ \Omega\left(\frac{\sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}\right), \\ \text{otherwise} \end{cases} \quad (148)$$

where $r(\lambda_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \mathbf{W}^*) = \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)$. Note that here the $\rho(\cdot)$ function is defined in

Definition 1. $\mathcal{L}_m(\cdot)$ and $\mathcal{L}_s(\cdot)$ are defined in (84) and (89), respectively.

The bound of $\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F$ is directly from (28). We can derive that

$$\mathcal{E}_w(\Psi) = O\left(\frac{\sqrt{\sum_{j=1}^L \lambda_l (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j^{\frac{1}{2}}\|)^2}}{\sum_{j=1}^L \lambda_l \|\boldsymbol{\Sigma}_j^{-1}\|^{-1} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_j}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_j^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_j^{-1}\|^{-\frac{1}{2}}\right)}\right) \quad (149)$$

$$\mathcal{E}(\Psi) = O\left(\frac{\sum_{j=1}^L \lambda_l (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j^{\frac{1}{2}}\|)^2}{\sum_{j=1}^L \lambda_l \|\boldsymbol{\Sigma}_j^{-1}\|^{-1} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_j}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_j^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_j^{-1}\|^{-\frac{1}{2}}\right)}\right) \quad (150)$$

$$\mathcal{E}_l(\Psi) = O\left(\frac{\sqrt{\sum_{j=1}^L \lambda_l (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j^{\frac{1}{2}}\|)^2} (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)}{\sum_{j=1}^L \lambda_l \|\boldsymbol{\Sigma}_j^{-1}\|^{-1} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_j}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_j^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_j^{-1}\|^{-\frac{1}{2}}\right)}\right) \quad (151)$$

The discussion of the monotonicity of $\mathcal{E}_w(\Psi)$, $\mathcal{E}(\Psi)$ and $\mathcal{E}_l(\Psi)$ can follow the analysis of $q(\Psi)$.

We finish our proof of Theorem 1 here. The parameters $\mathcal{B}(\Psi)$, $q(\Psi)$, $\mathcal{E}_w(\Psi)$, $\mathcal{E}(\Psi)$, and $\mathcal{E}_l(\Psi)$ can be found in (145), (148), (149), (150), and (151), respectively.

Proof of Corollary 1:

The monotonicity analysis has been included in the proof of Theorem 1. In this part, we specify our proof for the results in Table I. For simplicity, we denote $\rho_l = \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)$.

When everything else except $\|\boldsymbol{\Sigma}_l\|^{\frac{1}{2}}$ is fixed, if $\|\boldsymbol{\Sigma}_l\| = o(1)$, by some basic mathematical computation, then we have

$$\begin{aligned} n_{sc} &= C_0 \epsilon_0^{-2} \cdot \eta^2 \tau^{12} \kappa^{16} K^{14} \left(\sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{2}} (\delta_1(\mathbf{W}^*))^2 D_6(\Psi) \\ &\quad \cdot \left(\frac{1}{\sum_{l=1}^L \lambda_l \|\boldsymbol{\Sigma}_l^{-1}\|^{-1} \rho_l} \right)^2 \cdot d \log^2 d \\ &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \tau, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot O\left(\lambda_L \frac{1}{\|\boldsymbol{\Sigma}_L^{\frac{1}{2}}\|^6}\right) \end{aligned} \quad (152)$$

$$\begin{aligned}
v(\Psi) &= 1 - \frac{\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\kappa^2} \rho_l}{K^2(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2)} \\
&\leq 1 - \frac{\lambda_l}{K^2 \eta \kappa^2 \tau^K} \Theta(\|\Sigma_l\|^3)
\end{aligned} \tag{153}$$

$$\begin{aligned}
\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}^*\| &\leq O\left(\frac{K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 (1 + \xi)}}{\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)} \sqrt{\frac{d \log n}{n}}\right) \\
&\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O(1 - \|\Sigma_l\|^3)
\end{aligned} \tag{154}$$

$$\begin{aligned}
\bar{f}_l(\mathbf{W}_t) &= \bar{f}_l(\mathbf{W}_t) - \bar{f}_l(\mathbf{W}^*) \\
&\leq \mathbb{E}\left[\sum_{k=1}^K \frac{\partial(\bar{f}_l(\mathbf{W}_t))}{\partial \tilde{\mathbf{w}}_k} \right]^\top (\mathbf{w}_{t(k)} - \mathbf{w}_k^*) \\
&\leq \|\mathbf{W}_t - \mathbf{W}^* \mathbf{P}^*\| (\|\boldsymbol{\mu}_l\| + \|\Sigma_l\|^{\frac{1}{2}}) \\
&\lesssim O\left(\frac{\sum_{j=1}^L \sqrt{\lambda_j} (\|\boldsymbol{\mu}_j\| + \|\Sigma_j\|^{\frac{1}{2}})}{\sum_{j=1}^L \lambda_j \|\Sigma_j^{-1}\|^{-1} \rho_j}\right) (\|\boldsymbol{\mu}_j\| + \|\Sigma_j\|^{\frac{1}{2}}) \cdot \sqrt{\frac{d \log n}{n}} \eta \kappa^2 K^2 (1 + \xi) \\
&\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O\left(\frac{1}{1 + \|\Sigma_l\|^3}\right) \\
&\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O(1) - \Theta(\|\Sigma_l\|^3),
\end{aligned} \tag{155}$$

The first inequality of (155) is by the Mean Value Theorem. The second inequality of (155) is from Property 8, and the third inequality is derived from (28, 29). The last inequality is obtained by the condition that $\|\Sigma_l\| = o(1)$. We can similarly have

$$\begin{aligned}
\bar{f}(\mathbf{W}_t) &\leq \mathbb{E}\left[\sum_{k=1}^K \frac{\partial(\bar{f}(\mathbf{W}_t))}{\partial \tilde{\mathbf{w}}_k} \right]^\top (\mathbf{w}_{t(k)} - \mathbf{w}_k^*) \\
&\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O\left(\frac{1}{1 + \|\Sigma_l\|^3}\right) \\
&\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O(1) - \Theta(\|\Sigma_l\|^3)
\end{aligned} \tag{156}$$

If $\|\Sigma_l\|^{\frac{1}{2}} = \Omega(1)$, we have

$$n_{sc} \lesssim \text{poly}(\epsilon_0^{-1}, \eta, \tau, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot O(\|\Sigma_l\|^3) \tag{157}$$

$$v(\Psi) \leq 1 - \frac{1}{K^2 \tau^K \eta \kappa^2} \Theta\left(\frac{1}{1 + \|\Sigma_l\|}\right) \tag{158}$$

$$\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}^*\|_F \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^{\frac{5}{2}} (1 + \xi) \cdot \sqrt{\|\Sigma_l\|} \tag{159}$$

$$\bar{f}_l(\mathbf{W}_t) \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot \|\Sigma_l\| \tag{160}$$

$$\bar{f}(\mathbf{W}_t) \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot \|\Sigma_l\| \tag{161}$$

When everything is fixed except $\|\boldsymbol{\mu}_l\|$, by combining (15) and (141), we have

$$n_{sc} \lesssim \text{poly}(\epsilon_0^{-1}, \eta, \tau, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot \begin{cases} O(\|\boldsymbol{\mu}_l\|^4), & \text{if } \|\boldsymbol{\mu}_l\| \leq 1 \\ O(\|\boldsymbol{\mu}_l\|^{12}), & \text{if } \|\boldsymbol{\mu}_l\| \geq 1 \end{cases} \tag{162}$$

$$v(\Psi) \leq 1 - \frac{1}{K^2 \tau^K \eta \kappa^2} \Theta\left(\frac{1}{1 + \|\boldsymbol{\mu}_l\|^2}\right) \tag{163}$$

$$\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}^*\|_F \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^{\frac{5}{2}} (1 + \xi) \cdot (1 + \|\boldsymbol{\mu}_l\|) \tag{164}$$

$$\bar{f}_i(\mathbf{W}_t) \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot (1 + \|\boldsymbol{\mu}_i\|^2) \quad (165)$$

$$\bar{f}(\mathbf{W}_t) \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot (1 + \|\boldsymbol{\mu}_i\|^2) \quad (166)$$

When everything else is fixed except $\lambda_1, \lambda_2, \dots, \lambda_L$, where $\|\boldsymbol{\Sigma}_j\| = \Omega(1)$, $j \in [L]$ and $\|\boldsymbol{\mu}_j\| = \|\boldsymbol{\mu}_i\|$, $i, j \in [L]$, if $\|\boldsymbol{\Sigma}_i\| \leq \|\boldsymbol{\Sigma}_j\|$, $j \in [L]$, we have

$$\begin{aligned} n_{sc} &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot \frac{(a_1 \lambda_i^2 + a_2 \lambda_i^{\frac{3}{2}} + a_3 \lambda_i + a_4 \lambda_i^{\frac{1}{2}} + a_5)}{(\sum_{j=1}^L \lambda_j \rho_j)^2} \\ &\leq \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot \frac{a_5}{(\sum_{j=1}^L \lambda_j \rho_j)^2} \\ &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot O((1 + \lambda_i)^{-2}) \end{aligned} \quad (167)$$

where $a_1 = (\|\boldsymbol{\mu}_i\| + \|\boldsymbol{\Sigma}_i\|^{\frac{1}{2}})^{12} / \|\boldsymbol{\Sigma}_i\|^3$, $a_2 = (\|\boldsymbol{\mu}_i\| + \|\boldsymbol{\Sigma}_i\|^{\frac{1}{2}})^8 (\sum_{j \neq i} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^8)^{\frac{1}{2}} / \|\boldsymbol{\Sigma}_i\|^3$, $a_3 = (\|\boldsymbol{\mu}_i\| / \|\boldsymbol{\Sigma}_i\|^{\frac{1}{2}} + 1)^6 (\sum_{j \neq i} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^4 \sum_{j \neq i} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^8)^{\frac{1}{2}} + (\|\boldsymbol{\mu}_i\| + \|\boldsymbol{\Sigma}_i\|^{\frac{1}{2}})^6 \sum_{j \neq i} \lambda_j (\|\boldsymbol{\mu}_j\| / \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}} + 1)^6$, $a_4 = \sum_{j \neq i} \lambda_j (\|\boldsymbol{\mu}_j\| / \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}} + 1)^6 (\|\boldsymbol{\mu}_i\| + \|\boldsymbol{\Sigma}_i\|^{\frac{1}{2}})^2 (\sum_{j \neq i} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^8)^{\frac{1}{2}}$, $a_5 = (\sum_{j \neq i} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^4 \sum_{j \neq i} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^8)^{\frac{1}{2}} \cdot \sum_{j \neq i} \lambda_j (\|\boldsymbol{\mu}_j\| / \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}} + 1)^6$. The second step of (167) is by $a_i = O(a_5)$, $i = 1, 2, 3, 4$.

$$v \leq \frac{1}{K^2 \eta \tau^K \kappa^2} \Theta\left(\frac{1}{1 + \lambda_i}\right) \quad (168)$$

$$\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \leq \text{poly}(\eta, \kappa, \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^{\frac{5}{2}} (1 + \xi) \cdot O\left(\frac{1}{1 + \sqrt{\lambda_i}}\right) \quad (169)$$

$$\bar{f}_i(\mathbf{W}_t) \leq \text{poly}(\eta, \kappa, \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O\left(\frac{1}{1 + \sqrt{\lambda_i}}\right) \quad (170)$$

$$\bar{f}(\mathbf{W}_t) \leq \text{poly}(\eta, \kappa, \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O\left(\frac{1}{1 + \lambda_i}\right) \quad (171)$$

If $\|\boldsymbol{\Sigma}_i\| \geq \|\boldsymbol{\Sigma}_j\|$, $j \in [L]$, we can similarly derive that

$$\begin{aligned} n_{sc} &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot \frac{(a_1 \lambda_i^2 + a_2 \lambda_i^{\frac{3}{2}} + a_3 \lambda_i + a_4 \lambda_i^{\frac{1}{2}} + a_5)}{(\sum_{j=1}^L \lambda_j \rho_j)^2} \\ &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot (O(1) - \Theta((1 + \lambda_i)^{-2})) \end{aligned} \quad (172)$$

$$v \leq 1 - \frac{1}{K^2 \eta \tau^K \kappa^2} \Theta\left(\frac{1}{1 + \lambda_i}\right) \quad (173)$$

$$\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \leq \text{poly}(\eta, \kappa, \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^{\frac{5}{2}} (1 + \xi) \cdot O(1 + \sqrt{\lambda_i}) \quad (174)$$

$$\bar{f}_i(\mathbf{W}_t) \leq \text{poly}(\eta, \kappa, \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O(1 + \sqrt{\lambda_i}) \quad (175)$$

$$\bar{f}(\mathbf{W}_t) \leq \text{poly}(\eta, \kappa, \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot \left(O(1) - \frac{\Theta(1)}{1 + \lambda_i}\right) \quad (176)$$

L. Proof of Lemma 1 and its supportive lemmas

We first describe the proof of Lemma 1 in Section B2. The proofs of the supportive lemmas are provided in Section L1 to L5 in sequence. The proof idea mainly follows from [28]. Lemma 6 shows the Hessian $\nabla^2 \bar{f}(\mathbf{W})$ of the population risk function is smooth. Lemma 7 illustrates that $\nabla^2 \bar{f}(\mathbf{W})$ is strongly convex in the neighborhood around $\boldsymbol{\mu}^*$. Lemma 8 shows the Hessian of the empirical risk function $\nabla^2 f_n(\mathbf{W}^*)$ is close to its population risk $\nabla^2 \bar{f}(\mathbf{W}^*)$ in the local convex region. Summing up these three lemmas, we can derive the proof of Lemma 1. Lemma 4 is used in the proof of Lemma 7. Lemma 5 is used in the proof of Lemma 8.

The analysis of the Hessian matrix of the population loss in [28] and [26] can not be extended to the Gaussian mixture model. To solve this problem, we develop new tools using some good properties of symmetric distribution and even function. Our approach can also be applied to other activations like tanh or erf. Moreover, if we directly apply the existing matrix

concentration inequalities in these works in bounding the error between the empirical loss and the population loss, the resulting sample complexity bound is loose and cannot reflect the influence of each component of the Gaussian mixture distribution. We develop a new version of Bernstein's inequality (see (237)) so that the final bound is $O(d \log^2 d)$.

[83] showed that the landscape of the empirical risk is close to that of the population risk when the number of samples is sufficiently large for the special case that $K = 1$. Focusing on Gaussian mixture models, our result explicitly shows how the parameters of the input distribution, including the proportion, mean and, variance of each component will affect the error bound between the empirical loss and the population loss in Lemma 8.

1) *Proof of Lemma 4:* Following the proof idea in Lemma D.4 of [26], we have

$$\mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[\left(\sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i) \right)^2 \right] = A_0 + B_0 \quad (177)$$

$$A_0 = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left(\sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'^2(\sigma \cdot x_i) \cdot \mathbf{x} \mathbf{x}^\top \mathbf{r}_i \right) \quad (178)$$

$$B_0 = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left(\sum_{i \neq l} \mathbf{r}_i^\top \phi'(\sigma \cdot x_i) \phi'(\sigma \cdot x_l) \cdot \mathbf{x} \mathbf{x}^\top \mathbf{r}_l \right) \quad (179)$$

In A_0 , we know that $\mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} x_j = 0$. Therefore, by some basic mathematical computation,

$$\begin{aligned} A_0 &= \sum_{i=1}^k \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[\mathbf{r}_i^\top \left(\phi'^2(\sigma \cdot x_i) \left(x_i^2 \mathbf{e}_i \mathbf{e}_i^\top + \sum_{j \neq i} x_j x_j (\mathbf{e}_i \mathbf{e}_j^\top + \mathbf{e}_j \mathbf{e}_i^\top) + \sum_{j \neq i} \sum_{l \neq i} x_j x_l \mathbf{e}_j \mathbf{e}_l^\top \right) \right) \mathbf{r}_i \right] \\ &= \sum_{i=1}^k \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[\mathbf{r}_i^\top \left(\phi'^2(\sigma \cdot x_i) \left(x_i^2 \mathbf{e}_i \mathbf{e}_i^\top + \sum_{j \neq i} x_j^2 \mathbf{e}_j \mathbf{e}_j^\top \right) \right) \mathbf{r}_i \right] \\ &= \sum_{i=1}^k \left[\mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} [\phi'^2(\sigma \cdot x_i) x_i^2] \mathbf{r}_i^\top \mathbf{e}_i \mathbf{e}_i^\top \mathbf{r}_i \right. \\ &\quad \left. + \sum_{j \neq i} \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} [x_j^2] \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I})} [\phi'^2(\sigma \cdot x_i)] \mathbf{r}_i^\top \mathbf{e}_j \mathbf{e}_j^\top \mathbf{r}_i \right] \\ &= \sum_{i=1}^k r_{ii}^2 \beta_2(i, \boldsymbol{\mu}, \sigma) + \sum_{i=1}^k \sum_{j \neq i} r_{ij}^2 \beta_0(i, \boldsymbol{\mu}, \sigma) (1 + \mu_j^2) \end{aligned} \quad (180)$$

In B_0 , $\alpha_1(i, \boldsymbol{\mu}, \sigma) = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} (x_i \phi'(x_i)) = 0$. By the equation in Page 30 of [26], we have

$$\begin{aligned} B_0 &= \sum_{i \neq l}^k \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[\mathbf{r}_i^\top \left(\phi'(\sigma \cdot x_i) \phi'(\sigma \cdot x_l) \left(x_i^2 \mathbf{e}_i \mathbf{e}_i^\top + x_l^2 \mathbf{e}_l \mathbf{e}_l^\top + x_i x_l (\mathbf{e}_i \mathbf{e}_l^\top + \mathbf{e}_l \mathbf{e}_i^\top) \right. \right. \right. \\ &\quad \left. \left. \left. + \sum_{j \neq i} x_j x_l \mathbf{e}_j \mathbf{e}_l^\top + \sum_{j \neq l} x_j x_i \mathbf{e}_j \mathbf{e}_i^\top + \sum_{j \neq i, l} \sum_{j' \neq i, l} x_j x_{j'} \mathbf{e}_j \mathbf{e}_{j'}^\top \right) \right) \mathbf{r}_l \right] \\ &= \sum_{i \neq l} r_{ii} r_{li} \alpha_2(i, \boldsymbol{\mu}, \sigma) \alpha_0(l, \boldsymbol{\mu}, \sigma) + \sum_{i \neq l} r_{ij} r_{lj} \alpha_0(i, \boldsymbol{\mu}, \sigma) \alpha_0(l, \boldsymbol{\mu}, \sigma) (1 + \mu_j^2) \end{aligned} \quad (181)$$

Therefore,

$$\begin{aligned} A_0 + B_0 &= \sum_{i=1}^k \left(r_{ii} \frac{\alpha_2(i, \boldsymbol{\mu}, \sigma)}{\sqrt{1 + \mu_i^2}} + \sum_{l \neq i} r_{li} \alpha_0(l, \boldsymbol{\mu}, \sigma) \sqrt{1 + \mu_i^2} \right)^2 - \sum_{i=1}^k r_{ii}^2 \frac{\alpha_2^2(i, \boldsymbol{\mu}, \sigma)}{1 + \mu_i^2} \\ &\quad - \sum_{i=1}^k \sum_{l \neq i} r_{li}^2 \alpha_0(l, \boldsymbol{\mu}, \sigma)^2 (1 + \mu_i^2) + \sum_{i=1}^k r_{ii}^2 \beta_2(i, \boldsymbol{\mu}, \sigma) + \sum_{i=1}^k \sum_{j \neq i} r_{ij}^2 \beta_0(i, \boldsymbol{\mu}, \sigma) (1 + \mu_j^2) \\ &\geq \sum_{i=1}^k r_{ii}^2 \left(\beta_2(i, \boldsymbol{\mu}, \sigma) - \frac{\alpha_2^2(i, \boldsymbol{\mu}, \sigma)}{1 + \mu_i^2} \right) + \sum_{i=1}^k \sum_{j \neq i} r_{ij}^2 \left(\beta_0(i, \boldsymbol{\mu}, \sigma) - \alpha_0^2(i, \boldsymbol{\mu}, \sigma) \right) (1 + \mu_j^2) \\ &\geq \rho(\boldsymbol{\mu}, \sigma) \|\mathbf{R}\|_F^2 \end{aligned} \quad (182)$$

2) *Proof of Lemma 5*: Following the equation (92) in Lemma 8 of [28] and by (131)

$$\|\nabla^2 \ell(\mathbf{W}) - \nabla^2 \ell(\mathbf{W}')\| \leq \sum_{j=1}^K \sum_{l=1}^K |\xi_{j,l}(\mathbf{W}) - \xi_{j,l}(\mathbf{W}')| \cdot \|\mathbf{x}\mathbf{x}^\top\| \quad (183)$$

By Lagrange's inequality, we have

$$|\xi_{j,l}(\mathbf{W}) - \xi_{j,l}(\mathbf{W}')| \leq (\max_k |T_{j,k,l}|) \cdot \|\mathbf{x}\| \cdot \sqrt{K} \|\mathbf{W} - \mathbf{W}'\|_F \quad (184)$$

From Lemma 6, we know

$$\max_k |T_{j,k,l}| \leq C_7 \quad (185)$$

By Property 7, we have

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\|\mathbf{x}\|^{2t}] \leq d^t (2t-1)!! \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^{2t} \quad (186)$$

Therefore, for some constant $C_{12} > 0$

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sup_{\mathbf{W} \neq \mathbf{W}'} \frac{\|\nabla^2 \ell(\mathbf{W}) - \nabla^2 \ell(\mathbf{W}')\|}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \leq K^{\frac{5}{2}} \mathbb{E}[\|\mathbf{x}\|_2^3] \\ & \leq K^{\frac{5}{2}} \sqrt{d \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2} \sqrt{3d^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^4} \\ & = C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^4} \end{aligned} \quad (187)$$

3) *Proof of Lemma 6*: Let $\mathbf{a} = (\mathbf{a}_1^\top, \dots, \mathbf{a}_K^\top)^\top \in \mathbb{R}^{dK}$. Let $\Delta_{j,l} \in \mathbb{R}^{d \times d}$ be the (j, l) -th block of $\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P}) \in \mathbb{R}^{dK \times dK}$. By definition,

$$\|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| = \max_{\|\mathbf{a}\|=1} \sum_{j=1}^K \sum_{l=1}^K \mathbf{a}_j^\top \Delta_{j,l} \mathbf{a}_l \quad (188)$$

Denote $\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_K) \in \mathbb{R}^{K \times K}$. By the mean value theorem and (131),

$$\begin{aligned} \Delta_{j,l} &= \frac{\partial^2 \bar{f}(\mathbf{W})}{\partial \mathbf{w}_j \partial \mathbf{w}_l} - \frac{\partial^2 \bar{f}(\mathbf{W}^* \mathbf{P})}{\partial \mathbf{w}_j^* \partial \mathbf{w}_l^*} = \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \mathbf{I}_d)} [(\xi_{j,l}(\mathbf{W}) - \xi_{j,l}(\mathbf{W}^* \mathbf{P})) \cdot \mathbf{x}\mathbf{x}^\top] \\ &= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sum_{k=1}^K \left\langle \frac{\partial \xi_{j,l}(\mathbf{W}')}{\partial \mathbf{w}'_k}, \mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k \right\rangle \cdot \mathbf{x}\mathbf{x}^\top \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sum_{k=1}^K \langle T_{j,l,k} \cdot \mathbf{x}, \mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k \rangle \cdot \mathbf{x}\mathbf{x}^\top \right] \end{aligned} \quad (189)$$

where $\mathbf{W}' = \gamma \mathbf{W} + (1 - \gamma) \mathbf{W}^* \mathbf{P}$ for some $\gamma \in (0, 1)$ and $T_{j,l,k}$ is defined such that $\frac{\partial \xi_{j,l}(\mathbf{W}')}{\partial \mathbf{w}'_k} = T_{j,l,k} \cdot \mathbf{x} \in \mathbb{R}^d$. Then we provide an upper bound for $\xi_{j,l}$. Since that $y = 1$ or 0 , we first compute the case in which $y = 1$. From (131) we can obtain

$$\xi_{j,l}(\mathbf{W}) = \begin{cases} \frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \cdot \frac{1}{H^2(\mathbf{W})}, & j \neq l \\ \frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \cdot \frac{1}{H^2(\mathbf{W})} - \frac{1}{K} \phi''(\mathbf{w}_j^\top \mathbf{x}) \cdot \frac{1}{H(\mathbf{W})}, & j = l \end{cases} \quad (190)$$

We can bound $\xi_{j,l}(\mathbf{W})$ by bounding each component of (190). Note that we have

$$\frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \cdot \frac{1}{H^2(\mathbf{W})} \leq \frac{1}{K^2} \frac{\phi(\mathbf{w}_j^\top \mathbf{x}) \phi(\mathbf{w}_l^\top \mathbf{x}) (1 - \phi(\mathbf{w}_j^\top \mathbf{x})) (1 - \phi(\mathbf{w}_l^\top \mathbf{x}))}{\frac{1}{K^2} \phi(\mathbf{w}_j^\top \mathbf{x}) \phi(\mathbf{w}_l^\top \mathbf{x})} \leq 1 \quad (191)$$

$$\frac{1}{K} \phi''(\mathbf{w}_j^\top \mathbf{x}) \cdot \frac{1}{H(\mathbf{W})} \leq \frac{1}{K} \frac{\phi(\mathbf{w}_j^\top \mathbf{x}) (1 - \phi(\mathbf{w}_j^\top \mathbf{x})) (1 - 2\phi(\mathbf{w}_j^\top \mathbf{x}))}{\frac{1}{K} \phi(\mathbf{w}_j^\top \mathbf{x})} \leq 1 \quad (192)$$

where (191) holds for any $j, l \in [K]$. The case $y = 0$ can be computed with the same upper bound by substituting $(1 - H(\mathbf{W})) = \frac{1}{K} \sum_{j=1}^K (1 - \phi(\mathbf{w}_j^\top \mathbf{x}))$ for $H(\mathbf{W})$ in (190), (191) and (192). Therefore, there exists a constant $C_9 > 0$, such that

$$|\xi_{j,l}(\mathbf{W})| \leq C_9 \quad (193)$$

We then need to calculate $T_{j,l,k}$. Following the analysis of $\xi_{j,l}(\mathbf{W})$, we only consider the case of $y = 1$ here for simplicity.

$$T_{j,l,k} = \frac{-2}{K^3 H^3(\mathbf{W}')} \phi'(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_l^\top \mathbf{x}) \phi'(\mathbf{w}'_k^\top \mathbf{x}), \quad \text{where } j, l, k \text{ are not equal to each other} \quad (194)$$

$$T_{j,j,k} = \begin{cases} \frac{-2}{K^3 H^3(\mathbf{W}')} \phi'(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_k^\top \mathbf{x}) + \frac{1}{K^2 H^2(\mathbf{W}')} \phi''(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_k^\top \mathbf{x}), & j \neq k \\ \frac{-2}{K^3 H^3(\mathbf{W}')} (\phi'(\mathbf{w}'_j^\top \mathbf{x}))^3 + \frac{3}{K^2 H^2(\mathbf{W}')} \phi''(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_j^\top \mathbf{x}) - \frac{\phi'''(\mathbf{w}'_j^\top \mathbf{x})}{KH(\mathbf{W}')}, & j = k \end{cases} \quad (195)$$

$$\begin{aligned} \mathbf{a}_j^\top \Delta_{j,l} \mathbf{a}_l &= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{k=1}^K T_{j,l,k} \langle \mathbf{x}, \mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k \rangle \right) \cdot (\mathbf{a}_j^\top \mathbf{x}) (\mathbf{a}_l^\top \mathbf{x}) \right] \\ &\leq \sqrt{\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sum_{k=1}^K T_{j,k,l}^2 \right] \cdot \mathbb{E} \left[\sum_{k=1}^K (\langle \mathbf{x}, \mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k \rangle (\mathbf{a}_j^\top \mathbf{x}) (\mathbf{a}_l^\top \mathbf{x}))^2 \right]} \\ &\leq \sqrt{\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sum_{k=1}^K T_{j,k,l}^2 \right]} \sqrt{\sum_{k=1}^K \sqrt{\mathbb{E}[(\mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k)^\top \mathbf{x}]^4} \cdot \sqrt{\mathbb{E}[(\mathbf{a}_j^\top \mathbf{x})^4 (\mathbf{a}_l^\top \mathbf{x})^4]}} \\ &\leq C_8 \sqrt{\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sum_{k=1}^K T_{j,k,l}^2 \right]} \sqrt{\sum_{k=1}^K \|\mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k\|_2^2 \cdot \|\mathbf{a}_j\|_2^2 \cdot \|\mathbf{a}_l\|_2^2} \\ &\quad \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \end{aligned} \quad (196)$$

for some constant $C_8 > 0$. All the three inequalities of (196) are derived from Cauchy-Schwarz inequality. Note that we have

$$\begin{aligned} \left| \frac{-2}{K^3 H^3(\mathbf{W}')} (\phi'(\mathbf{w}'_j^\top \mathbf{x}))^2 \phi'(\mathbf{w}'_k^\top \mathbf{x}) \right| &\leq \frac{2\phi^2(\mathbf{w}'_j^\top \mathbf{x})(1 - \phi(\mathbf{w}'_j^\top \mathbf{x}))^2 \phi(\mathbf{w}'_k^\top \mathbf{x})(1 - \phi(\mathbf{w}'_k^\top \mathbf{x}))}{K^3 \frac{1}{K^3} \phi^2(\mathbf{w}'_j^\top \mathbf{x}) \phi(\mathbf{w}'_k^\top \mathbf{x})} \\ &= 2(1 - \phi(\mathbf{w}'_j^\top \mathbf{x}))^2 (1 - \phi(\mathbf{w}'_k^\top \mathbf{x})) \leq 2 \end{aligned} \quad (197)$$

$$\left| \frac{-2}{K^3 H^3(\mathbf{W}')} \phi'(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_l^\top \mathbf{x}) \phi'(\mathbf{w}'_k^\top \mathbf{x}) \right| \leq 2 \quad (198)$$

$$\begin{aligned} &\left| \frac{3}{K^2 H^2(\mathbf{W}')} \phi''(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_k^\top \mathbf{x}) \right| \\ &\leq \left| \frac{3\phi(\mathbf{w}'_j^\top \mathbf{x})(1 - \phi(\mathbf{w}'_j^\top \mathbf{x}))(1 - 2\phi(\mathbf{w}'_j^\top \mathbf{x}))\phi(\mathbf{w}'_k^\top \mathbf{x})(1 - \phi(\mathbf{w}'_k^\top \mathbf{x}))}{K^2 \frac{1}{K^2} \phi(\mathbf{w}'_j^\top \mathbf{x}) \phi(\mathbf{w}'_k^\top \mathbf{x})} \right| \\ &= \left| 3(1 - \phi(\mathbf{w}'_j^\top \mathbf{x}))(1 - 2\phi(\mathbf{w}'_j^\top \mathbf{x}))(1 - \phi(\mathbf{w}'_k^\top \mathbf{x})) \right| \leq 3 \end{aligned} \quad (199)$$

$$\left| \frac{\phi'''(\mathbf{w}'_j^\top \mathbf{x})}{KH(\mathbf{W}')} \right| \leq \left| \frac{\phi(\mathbf{w}'_j^\top \mathbf{x})(1 - \phi(\mathbf{w}'_j^\top \mathbf{x}))(1 - 6\phi(\mathbf{w}'_j^\top \mathbf{x}) + 6\phi^2(\mathbf{w}'_j^\top \mathbf{x}))}{K \frac{1}{K} \phi(\mathbf{w}'_j^\top \mathbf{x})} \right| \leq 1 \quad (200)$$

Therefore, by combining (194), (195) and (197) to (200), we have

$$|T_{j,l,k}| \leq C_7 \quad \Rightarrow \quad T_{j,l,k}^2 \leq C_7^2, \quad \forall j, l, k \in [K], \quad (201)$$

for some constants $C_7 > 0$. By (188), (189), (196), (201) and the Cauchy-Schwarz's Inequality, we have

$$\begin{aligned} &\|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \\ &\leq C_8 \sqrt{C_7^2 K} \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \\ &\quad \cdot \max_{\|\mathbf{a}_l\|=1} \sum_{j=1}^K \sum_{l=1}^K \|\mathbf{a}_j\|_2 \|\mathbf{a}_l\|_2 \\ &\leq C_8 \sqrt{C_7^2 K} \cdot \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \cdot \left(\sum_{j=1}^K \|\mathbf{a}_j\| \right)^2 \\ &\leq C_8 \sqrt{C_7^2 K^3} \cdot \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \end{aligned} \quad (202)$$

Hence, we have

$$\begin{aligned} & \|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \\ & \leq C_5 K^{\frac{3}{2}} \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \end{aligned} \quad (203)$$

for some constant $C_5 > 0$.

4) *Proof of Lemma 7* : From [28], we know

$$\begin{aligned} \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P}) & \succeq \min_{\|\mathbf{a}\|=1} \frac{4}{K^2} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{j=1}^K \phi'(\mathbf{w}_{\pi^*(j)}^\top \mathbf{x}) (\mathbf{a}_{\pi^*(j)}^\top \mathbf{x}) \right)^2 \right] \cdot \mathbf{I}_{dK} \\ & = \min_{\|\mathbf{a}\|=1} \frac{4}{K^2} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{j=1}^K \phi'(\mathbf{w}_j^\top \mathbf{x}) (\mathbf{a}_j^\top \mathbf{x}) \right)^2 \right] \cdot \mathbf{I}_{dK} \end{aligned} \quad (204)$$

with $\mathbf{a} = (\mathbf{a}_1^\top, \dots, \mathbf{a}_K^\top)^\top \in \mathbb{R}^{dK}$, where \mathbf{P} is a specific permutation matrix and $\{\pi^*(j)\}_{j=1}^K$ is the indices permuted by \mathbf{P} . Similarly,

$$\begin{aligned} \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P}) & \preceq \left(\max_{\|\mathbf{a}\|=1} \mathbf{a}^\top \nabla^2 \bar{f}(\mathbf{W}^*) \mathbf{a} \right) \cdot \mathbf{I}_{dK} \leq C_4 \cdot \max_{\|\mathbf{a}\|=1} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sum_{j=1}^K (\mathbf{a}_{\pi^*(j)}^\top \mathbf{x})^2 \right] \cdot \mathbf{I}_{dK} \\ & = C_4 \cdot \max_{\|\mathbf{a}\|=1} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sum_{j=1}^K (\mathbf{a}_j^\top \mathbf{x})^2 \right] \cdot \mathbf{I}_{dK} \end{aligned} \quad (205)$$

for some constant $C_4 > 0$. By applying Property 8, we can derive the upper bound in (205) as

$$C_4 \cdot \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sum_{j=1}^K (\mathbf{a}_j^\top \mathbf{x})^2 \right] \cdot \mathbf{I}_{dK} \leq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot \mathbf{I}_{dK} \quad (206)$$

To find a lower bound for (204), we can first transfer the expectation of the Gaussian Mixture Model to the weight sum of the expectations over general Gaussian distributions.

$$\begin{aligned} & \min_{\|\mathbf{a}\|=1} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{j=1}^K \phi'(\mathbf{w}_j^\top \mathbf{x}) (\mathbf{a}_j^\top \mathbf{x}) \right)^2 \right] \\ & = \min_{\|\mathbf{a}\|=1} \sum_{l=1}^L \lambda_l \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{j=1}^K \phi'(\mathbf{w}_j^\top \mathbf{x}) (\mathbf{a}_j^\top \mathbf{x}) \right)^2 \right] \end{aligned} \quad (207)$$

Denote $\mathbf{U} \in \mathbb{R}^{d \times k}$ as the orthogonal basis of \mathbf{W}^* . For any vector $\mathbf{a}_i \in \mathbb{R}^d$, there exists two vectors $\mathbf{b}_i \in \mathbb{R}^k$ and $\mathbf{c}_i \in \mathbb{R}^{d-k}$ such that

$$\mathbf{a}_i = \mathbf{U} \mathbf{b}_i + \mathbf{U}_\perp \mathbf{c}_i \quad (208)$$

where $\mathbf{U}_\perp \in \mathbb{R}^{d \times (d-k)}$ denotes the complement of \mathbf{U} . We also have $\mathbf{U}_\perp^\top \boldsymbol{\mu}_l = 0$ by Property 1. Plugging (208) into RHS of (207), and then we have

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^K \mathbf{a}_i^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^\top \mathbf{x}) \right)^2 \right] \\ & = \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^K (\mathbf{U} \mathbf{b}_i + \mathbf{U}_\perp \mathbf{c}_i)^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^\top \mathbf{x}) \right)^2 \right] = A + B + C \end{aligned} \quad (209)$$

$$A = \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^K \mathbf{b}_i^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^\top \mathbf{x}) \right)^2 \right] \quad (210)$$

$$\begin{aligned}
C &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[2 \left(\sum_{i=1}^K \mathbf{c}_i^\top \mathbf{U}_\perp^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right) \cdot \left(\sum_{i=1}^K \mathbf{b}_i^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right) \right] \\
&= \sum_{i=1}^K \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[2 \mathbf{c}_i^\top \mathbf{U}_\perp^\top \mathbf{x} \right] \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\mathbf{b}_j^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \phi'(\mathbf{w}_j^{*\top} \mathbf{x}) \right] \\
&= \sum_{i=1}^K \sum_{j=1}^K \left[2 \mathbf{c}_i^\top \mathbf{U}_\perp^\top \boldsymbol{\mu}_l \right] \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\mathbf{b}_j^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \phi'(\mathbf{w}_j^{*\top} \mathbf{x}) \right] = 0
\end{aligned} \tag{211}$$

where the last step is by $\mathbf{U}_\perp^\top \boldsymbol{\mu}_l = 0$ by Property 1.

$$\begin{aligned}
B &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^K \mathbf{c}_i^\top \mathbf{U}_\perp^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right)^2 \right] \\
&= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[(\mathbf{t}^\top \mathbf{s})^2 \right] \quad \text{by defining } \mathbf{t} = \sum_{i=1}^k \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \mathbf{c}_i \in \mathbb{R}^{d-K} \text{ and } \mathbf{s} = \mathbf{U}_\perp^\top \mathbf{x} \\
&= \sum_{i=1}^K \mathbb{E}[t_i^2 s_i^2] + \sum_{i \neq j} \mathbb{E}[t_i t_j s_i s_j] \\
&= \sum_{i=1}^K \mathbb{E}[t_i^2] \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{lk}^2 + \left(\sum_{i=1}^K \mathbb{E}[t_i^2] (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_i^2 + \sum_{i \neq j} \mathbb{E}[t_i t_j] (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_i \cdot (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_j \right) \\
&= \mathbb{E} \left[\sum_{i=1}^{d-K} t_i^2 \cdot \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{lk}^2 \right] + \mathbb{E}[(\mathbf{t}^\top \mathbf{U}_\perp^\top \boldsymbol{\mu}_l)^2] = \mathbb{E} \left[\sum_{i=1}^{d-K} t_i^2 \cdot \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{lk}^2 \right]
\end{aligned} \tag{212}$$

The last step is by $\mathbf{U}_\perp^\top \boldsymbol{\mu}_l = 0$. The 4th step is because that s_i is independent of t_i , thus $\mathbb{E}[t_i t_j s_i s_j] = \mathbb{E}[t_i t_j] \mathbb{E}[s_i s_j]$

$$\mathbb{E}[s_i s_j] = \begin{cases} (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_i \cdot (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_j, & \text{if } i \neq j \\ (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_i^2 + \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{lk}^2, & \text{if } i = j \end{cases} \tag{213}$$

Since $\left(\sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i) \right)^2$ is an even function for any $\mathbf{r}_i \in \mathbb{R}^d$, $i \in [k]$, so from Property 5 we have

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i) \right)^2 \right] = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2} \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) + \frac{1}{2} \mathcal{N}(-\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i) \right)^2 \right] \tag{214}$$

Combining Lemma 4 and Property 5, we next follow the derivation for the standard Gaussian distribution in Page 36 of [26] and generalize the result to a Gaussian distribution with an arbitrary mean and variance as follows.

$$\begin{aligned}
A &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^K \mathbf{b}_i^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right)^2 \right] \\
&\geq \int (2\pi)^{-\frac{K}{2}} |\mathbf{U}^\top \boldsymbol{\Sigma}_l \mathbf{U}|^{-\frac{1}{2}} \left[\left(\sum_{i=1}^K \mathbf{b}_i^\top \mathbf{z} \cdot \phi'(\mathbf{v}_i^\top \mathbf{z}) \right)^2 \right] \exp\left(-\frac{1}{2} \|\boldsymbol{\Sigma}_l^{-1}\| \|\mathbf{z} - \mathbf{U}^\top \boldsymbol{\mu}_l\|^2\right) d\mathbf{z} \\
&= \int (2\pi)^{-\frac{K}{2}} |\mathbf{U}^\top \boldsymbol{\Sigma}_l \mathbf{U}|^{-\frac{1}{2}} \left[\left(\sum_{i=1}^K \mathbf{b}_i^\top \mathbf{V}^{\dagger\top} \mathbf{s} \cdot \phi'(s_i) \right)^2 \right] \exp\left(-\frac{1}{2} \|\boldsymbol{\Sigma}_l^{-1}\| \|\mathbf{V}^{\dagger\top} \mathbf{s} - \mathbf{U}^\top \boldsymbol{\mu}_l\|^2\right) |\det(\mathbf{V}^\dagger)| d\mathbf{s} \\
&\geq \int (2\pi)^{-\frac{K}{2}} |\mathbf{U}^\top \boldsymbol{\Sigma}_l \mathbf{U}|^{-\frac{1}{2}} \left[\left(\sum_{i=1}^k \mathbf{b}_i^\top \mathbf{V}^{\dagger\top} \mathbf{s} \cdot \phi'(s_i) \right)^2 \right] \exp\left(-\frac{\|\boldsymbol{\Sigma}_l^{-1}\| \|\mathbf{s} - \mathbf{V}^\top \mathbf{U}^\top \boldsymbol{\mu}_l\|^2}{2\delta_K^2(\mathbf{W}^*)}\right) |\det(\mathbf{V}^\dagger)| d\mathbf{s} \\
&\geq \int (2\pi)^{-\frac{K}{2}} |\mathbf{U}^\top \boldsymbol{\Sigma}_l \mathbf{U}|^{-\frac{1}{2}} \left[\left(\sum_{i=1}^k \mathbf{b}_i^\top \mathbf{V}^{\dagger\top} (\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}) \mathbf{g} \cdot \phi'(\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \cdot g_i) \right)^2 \right] \\
&\quad \cdot \exp\left(-\frac{\|\mathbf{g} - \frac{\sqrt{\|\boldsymbol{\Sigma}_l^{-1}\| \mathbf{W}^{*\top} \boldsymbol{\mu}_l}}{\delta_K(\mathbf{W}^*)}\|^2}{2}\right) |\det(\mathbf{V}^\dagger)| \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{K}{2}} \delta_K^K(\mathbf{W}^*) d\mathbf{g} \\
&= \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\tau^K \eta} \mathbb{E}_{\mathbf{g}} \left[\left(\sum_{i=1}^K (\mathbf{b}_i^\top \mathbf{V}^{\dagger\top} \delta_K(\mathbf{W}^*) \mathbf{g} \cdot \phi'(\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*) \cdot g_i) \right)^2 \right] \\
&\geq \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\tau^K \kappa^2 \eta} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)}, \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)\right) \|\mathbf{b}\|^2.
\end{aligned} \tag{215}$$

The second step is by letting $\mathbf{z} = \mathbf{U}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{U}^\top \boldsymbol{\mu}_l, \mathbf{U}^\top \boldsymbol{\Sigma}_l \mathbf{U})$, $\mathbf{y}^\top \mathbf{U}^\top \boldsymbol{\Sigma}_l^{-1} \mathbf{U} \mathbf{y} \leq \|\boldsymbol{\Sigma}_l^{-1}\| \|\mathbf{y}\|^2$ for any $\mathbf{y} \in \mathbb{R}^K$. The third step is by letting $\mathbf{s} = \mathbf{V}^\top \mathbf{z}$. The last to second step follows from $\mathbf{g} = \frac{\mathbf{s}}{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)}$, where $\mathbf{g} \sim \mathcal{N}\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)}, \mathbf{I}_K\right)$ and the last inequality is by Lemma 4. Similarly, we extend the derivation in Page 37 of [26] for the standard Gaussian distribution to a general Gaussian distribution as follows.

$$B = \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{ik}^2 \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\|\mathbf{t}\|^2] \geq \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)}, \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)\right) \|\mathbf{c}\|^2 \tag{216}$$

Combining (209) - (212), (215) and (216), we have

$$\min_{\|\mathbf{a}\|=1} \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^k \mathbf{a}_i^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right)^2 \right] \geq \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right). \tag{217}$$

For the Gaussian Mixture Model $\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$, we have

$$\begin{aligned}
&\min_{\|\mathbf{a}\|=1} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{i=1}^k \mathbf{a}_i^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right)^2 \right] \\
&\geq \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)
\end{aligned} \tag{218}$$

Therefore,

$$\begin{aligned}
&\frac{4}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right) \cdot \mathbf{I}_{dK} \\
&\leq \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P}) \leq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot \mathbf{I}_{dK}
\end{aligned} \tag{219}$$

From (19) in Lemma 6, since that we have the condition $\|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \leq r$ and (139), we can obtain

$$\begin{aligned}
&\|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \\
&\leq C_5 K^{\frac{3}{2}} \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \\
&\leq \frac{4\epsilon_0}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right),
\end{aligned} \tag{220}$$

where $\epsilon_0 \in (0, \frac{1}{4})$. Then we have

$$\begin{aligned} \|\nabla^2 \bar{f}(\mathbf{W})\| &\geq \|\nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| - \|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \\ &\geq \frac{4(1-\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right) \end{aligned} \quad (221)$$

$$\begin{aligned} \|\nabla^2 \bar{f}(\mathbf{W})\| &\leq \|\nabla^2 \bar{f}(\mathbf{W}^*)\| + \|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \\ &\leq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 + \frac{4}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-\frac{1}{2}}\|}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-\frac{1}{2}}\|\right) \\ &\lesssim C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \end{aligned} \quad (222)$$

The last inequality of (222) holds since $C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 = \Omega(\max_l \{\|\boldsymbol{\Sigma}_l\|\})$, $\frac{4}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right) = O\left(\frac{\max_l \{\|\boldsymbol{\Sigma}_l\|\}}{K^2}\right)$ and $\Omega(\max_l \{\|\boldsymbol{\Sigma}_l\|\}) \geq O\left(\frac{\max_l \{\|\boldsymbol{\Sigma}_l\|\}}{K^2}\right)$. Combining (221) and (222), we have

$$\begin{aligned} &\frac{4(1-\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right) \cdot \mathbf{I} \\ &\preceq \nabla^2 \bar{f}(\mathbf{W}) \preceq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \sigma_l)^2 \cdot \mathbf{I} \end{aligned} \quad (223)$$

5) *Proof of Lemma 8:* Let N_ϵ be the ϵ -covering number of the Euclidean ball $\mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$. It is known that $\log N_\epsilon \leq dK \log(\frac{3r}{\epsilon})$ from [76]. Let $\mathcal{W}_\epsilon = \{\mathbf{W}_1, \dots, \mathbf{W}_{N_\epsilon}\}$ be the ϵ -cover set with N_ϵ elements. For any $\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$, let $j(\mathbf{W}) = \arg \min_{j \in [N_\epsilon]} \|\mathbf{W} - \mathbf{W}_j(\mathbf{W})\|_F \leq \epsilon$ for all $\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$.

Then for any $\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$, we have

$$\begin{aligned} &\|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \\ &\leq \frac{1}{n} \left\| \sum_{i=1}^n [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] \right\| \\ &\quad + \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] \right\| \\ &\quad + \left\| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right\| \end{aligned} \quad (224)$$

Hence, we have

$$\mathbb{P}\left(\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \geq t\right) \leq \mathbb{P}(A_t) + \mathbb{P}(B_t) + \mathbb{P}(C_t) \quad (225)$$

where A_t , B_t and C_t are defined as

$$A_t = \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{1}{n} \left\| \sum_{i=1}^n [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \quad (226)$$

$$B_t = \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \quad (227)$$

$$\begin{aligned} C_t = &\left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left\| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] \right. \right. \\ &\left. \left. - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \end{aligned} \quad (228)$$

Then we bound $\mathbb{P}(A_t)$, $\mathbb{P}(B_t)$, and $\mathbb{P}(C_t)$ separately.

1) **Upper bound on $\mathbb{P}(B_t)$.** By Lemma 6 in [28], we obtain

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right\| \\ & \leq 2 \sup_{\mathbf{v} \in \mathbf{V}_{\frac{1}{4}}} \left| \left\langle \mathbf{v}, \left(\frac{1}{n} \sum_{i=1}^n \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right) \mathbf{v} \right\rangle \right| \end{aligned} \quad (229)$$

where $\mathbf{V}_{\frac{1}{4}}$ is a $\frac{1}{4}$ -cover of the unit-Euclidean-norm ball $\mathbb{B}(\mathbf{0}, 1)$ with $\log |\mathbf{V}_{\frac{1}{4}}| \leq dK \log 12$. Taking the union bound over \mathcal{W}_ϵ and $\mathbf{V}_{\frac{1}{4}}$, we have

$$\begin{aligned} \mathbb{P}(B_t) & \leq \mathbb{P} \left(\sup_{\mathbf{W} \in \mathcal{W}_\epsilon, \mathbf{v} \in \mathbf{V}_{\frac{1}{4}}} \left| \frac{1}{n} \sum_{i=1}^n G_i \right| \geq \frac{t}{6} \right) \\ & \leq \exp(dK (\log \frac{3r}{\epsilon} + \log 12)) \sup_{\mathbf{W} \in \mathcal{W}_\epsilon, \mathbf{v} \in \mathbf{V}_{\frac{1}{4}}} \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n G_i \right| \geq \frac{t}{6} \right) \end{aligned} \quad (230)$$

where $G_i = \left\langle \mathbf{v}, (\nabla^2 \ell(\mathbf{W}, \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}, \mathbf{x}_i)]) \mathbf{v} \right\rangle$ and $\mathbb{E}[G_i] = 0$. Here $\mathbf{v} = (\mathbf{u}_1^\top, \dots, \mathbf{u}_K^\top)^\top \in \mathbb{R}^{dK}$.

$$\begin{aligned} |G_i| & = \left| \sum_{j=1}^K \sum_{l=1}^K \left[\xi_{j,l} \mathbf{u}_j^\top \mathbf{x} \mathbf{x}^\top \mathbf{u}_l - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\xi_{j,l} \mathbf{u}_j^\top \mathbf{x} \mathbf{x}^\top \mathbf{u}_l) \right] \right| \\ & \leq C_9 \cdot \left[\sum_{j=1}^K (\mathbf{u}_j^\top \mathbf{x})^2 + \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right] \end{aligned} \quad (231)$$

for some $C_9 > 0$. The first step of (231) is by (130). The last step is by (193) and the Cauchy-Schwarz's Inequality.

$$\begin{aligned} \mathbb{E}[|G_i|^p] & \leq \sum_{l=1}^p \binom{p}{l} C_9 \cdot \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\left(\sum_{j=1}^K (\mathbf{u}_j^\top \mathbf{x})^2 \right)^l \right] \\ & \quad \cdot \left(\sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^{p-l} \\ & = \sum_{l=1}^p \binom{p}{l} C_9 \cdot \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sum_{l_1 + \dots + l_K = l} \frac{l!}{\prod_{j=1}^K l_j!} \prod_{j=1}^K (\mathbf{u}_j^\top \mathbf{x})^{2l_j} \right] \\ & \quad \cdot \left(\sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^{p-l} \\ & = \sum_{l=1}^p \binom{p}{l} C_9 \cdot \left[\sum_{l_1 + \dots + l_K = l} \frac{l!}{\prod_{j=1}^K l_j!} \prod_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^{2l_j} \right] \\ & \quad \cdot \left(\sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^{p-l} \\ & = C_9 \cdot \sum_{l=1}^p \binom{p}{l} \left(\sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^l \\ & \quad \cdot \left(\sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^{p-l} \\ & = C_9 \cdot \left(\sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^p \\ & \leq C_9 \cdot \left(\sum_{j=1}^K 1!! \|\mathbf{u}_j\|^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^p \\ & \leq C_9 \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^p \end{aligned} \quad (232)$$

where the first step is by the triangle inequality and the Binomial theorem, and the second step comes from the Multinomial theorem. The second to last inequality in (232) results from Property 8. The last inequality is because $\mathbf{v} \in \mathbf{V}_{\frac{1}{4}}$, $\sum_{j=1}^K \|u_j\|^2 = \|\mathbf{v}\|^2 \leq 1$.

$$\begin{aligned} \mathbb{E}[\exp(\theta G_i)] &= 1 + \theta \mathbb{E}[G_i] + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|G_i|^p]}{p!} \\ &\leq 1 + \sum_{p=2}^{\infty} \frac{|e\theta|^p}{p^p} C_9 \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^p \\ &\leq 1 + C_9 \cdot |e\theta|^2 \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^2 \end{aligned} \quad (233)$$

where the first inequality holds from $p! \geq (\frac{p}{e})^p$ and (232), and the third line holds provided that

$$\max_{p \geq 2} \left\{ \frac{|e\theta|^{(p+1)}}{(p+1)^{(p+1)}} \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^{p+1} \right. \\ \left. \frac{|e\theta|^p}{p^p} \cdot \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^p \right\} \leq \frac{1}{2} \quad (234)$$

Note that the quantity inside the maximization in (234) achieves its maximum when $p = 2$, because it is monotonously decreasing. Therefore, (234) holds if $\theta \leq \frac{27}{4e} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2$. Then

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n G_i \geq \frac{t}{6}\right) &= \mathbb{P}\left(\exp\left(\theta \sum_{i=1}^n G_i\right) \geq \exp\left(\frac{n\theta t}{6}\right)\right) \leq e^{-\frac{n\theta t}{6}} \prod_{i=1}^n \mathbb{E}[\exp(\theta G_i)] \\ &\leq \exp\left(C_{10}\theta^2 n \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^2 - \frac{n\theta t}{6}\right) \end{aligned} \quad (235)$$

for some constant $C_{10} > 0$. The first inequality follows from Markov's Inequality. When $\theta = \min\left\{\frac{t}{12C_{10} \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^2}, \frac{27}{4e} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right\}$, we have a modified Bernstein's Inequality for the Gaussian Mixture Model as follows

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n G_i \geq \frac{t}{6}\right) &\leq \exp\left(\max\left\{-\frac{C_{10}nt^2}{144 \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^2}, \right. \right. \\ &\quad \left. \left. - C_{11}n \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot t\right\}\right) \end{aligned} \quad (236)$$

for some constant $C_{11} > 0$. We can obtain the same bound for $\mathbb{P}\left(-\frac{1}{n} \sum_{i=1}^n G_i \geq \frac{t}{6}\right)$ by replacing G_i as $-G_i$. Therefore, we have

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n G_i\right| \geq \frac{t}{6}\right) &\leq 2 \exp\left(\max\left\{-\frac{C_{10}nt^2}{144 \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^2}, \right. \right. \\ &\quad \left. \left. - C_{11}n \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot t\right\}\right) \end{aligned} \quad (237)$$

Thus, as long as

$$t \geq C_6 \cdot \max\left\{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log \frac{36r}{\epsilon} + \log \frac{4}{\delta}}{n}}, \frac{dK \log \frac{36r}{\epsilon} + \log \frac{4}{\delta}}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 n}\right\} \quad (238)$$

for some large constant $C_6 > 0$, we have $\mathbb{P}(B_t) \leq \frac{\delta}{2}$.

2) **Upper bound on $\mathbb{P}(A_t)$ and $\mathbb{P}(C_t)$.** From Lemma 5, we can obtain

$$\begin{aligned}
& \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x})] \right| \\
& \leq \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{\left| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x})] \right|}{\|\mathbf{W} - \mathbf{W}_j(\mathbf{W})\|_F} \\
& \quad \cdot \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\mathbf{W} - \mathbf{W}_j(\mathbf{W})\|_F \\
& \leq C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4} \cdot \epsilon
\end{aligned} \tag{239}$$

Therefore, C_t holds if

$$t \geq C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4} \cdot \epsilon \tag{240}$$

We can bound the A_t as below.

$$\begin{aligned}
& \mathbb{P}\left(\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{1}{n} \left\| \sum_{i=1}^n [\nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right) \\
& \leq \frac{3}{t} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{1}{n} \left\| \sum_{i=1}^n [\nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right\| \right] \\
& = \frac{3}{t} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left\| \nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) \right\| \right] \\
& \leq \frac{3}{t} \mathbb{E} \left[\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{\left\| \nabla^2 \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) \right\|}{\|\mathbf{W} - \mathbf{W}_j(\mathbf{W})\|_F} \right] \cdot \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\mathbf{W} - \mathbf{W}_j(\mathbf{W})\|_F \\
& \leq \frac{C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4} \cdot \epsilon}{t},
\end{aligned} \tag{241}$$

where the first inequality is by Markov's inequality, and the last inequality comes from Lemma 5. Thus, taking

$$t \geq \frac{C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4} \cdot \epsilon}{\delta} \tag{242}$$

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

3) Final step

Let $\epsilon = \frac{\delta}{C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4} \cdot ndK}$ and $\delta = d^{-10}$, then from (238) and (242) we need

$$\begin{aligned}
t & > \max\left\{ \frac{1}{ndK}, C_6 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right. \\
& \quad \cdot \sqrt{\frac{dK \log(36rnd^{\frac{25}{2}} K^{\frac{7}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4)} + \log \frac{4}{\delta}}}{n}}, \\
& \quad \left. \frac{dK \log(36rnd^{\frac{25}{2}} K^{\frac{7}{2}} \cdot \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4)} + \log \frac{4}{\delta}}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 n} \right\}
\end{aligned} \tag{243}$$

So by setting $t = \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}}$, as long as $n \geq C' \cdot dK \log dK$, we have

$$\mathbb{P}\left(\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \geq C_6 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}} \right) \leq d^{-10} \tag{244}$$

M. Proof of Lemma 2 and its supportive lemmas

1) Proof of Lemma 9: Note that $\nabla \tilde{f}_n(\mathbf{W}) = \nabla f_n(\mathbf{W}) + \frac{1}{n} \sum_{i=1}^n \nu_i$, $\nabla \tilde{f}(\mathbf{W}) = \nabla \bar{f}(\mathbf{W}) + \mathbb{E}[\nu_i] = \nabla \bar{f}(\mathbf{W})$. Therefore, we have

$$\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^*, P, r)} \|\nabla \tilde{f}_n(\mathbf{W}) - \nabla \tilde{f}(\mathbf{W})\| \leq \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^*, P, r)} \|\nabla f_n(\mathbf{W}) - \nabla \bar{f}(\mathbf{W})\| + \left\| \frac{1}{n} \sum_{i=1}^n \nu_i \right\| \quad (245)$$

Then, similar to the idea of the proof of Lemma 8, we adopt an ϵ -covering net of the ball $\mathbb{B}(\mathbf{W}^*, r)$ to build a relationship between any arbitrary point in the ball and the points in the covering set. We can then divide the distance between $\nabla f_n(\mathbf{W})$ and $\nabla \bar{f}(\mathbf{W})$ into three parts, similar to (224). (246) to (248) can be derived in a similar way as (226) to (228), with “ ∇^2 ” replaced by “ ∇ ”. Then we need to bound $\mathbb{P}(A'_t)$, $\mathbb{P}(B'_t)$ and $\mathbb{P}(C'_t)$ respectively, where A'_t , B'_t and C'_t are defined below.

$$A'_t = \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^*, P, r)} \frac{1}{n} \left\| \sum_{i=1}^n [\nabla \ell(\mathbf{W}; \mathbf{x}_i) - \nabla \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \quad (246)$$

$$B'_t = \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^*, P, r)} \left\| \frac{1}{n} \sum_{i=1}^n \nabla \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \quad (247)$$

$$C'_t = \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^*, P, r)} \left\| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}; \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \quad (248)$$

(a) Upper bound of $\mathbb{P}(B'_t)$. Applying Lemma 3 in [83], we have

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n \nabla \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)] \right\| \\ & \leq 2 \sup_{\mathbf{v} \in V_{\frac{1}{2}}} \left| \left\langle \frac{1}{n} \sum_{i=1}^n \nabla \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}_j(\mathbf{W}); \mathbf{x}_i)], \mathbf{v} \right\rangle \right| \end{aligned} \quad (249)$$

Define $G'_i = \left\langle \mathbf{v}, (\nabla \ell(\mathbf{W}, \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}, \mathbf{x}_i)]) \right\rangle$. Here $\mathbf{v} \in \mathbb{R}^d$. To compute $\nabla \ell(\mathbf{W}, \mathbf{x}_i)$, we require the derivation in Property 9. Then we can have an upper bound of $\zeta(\mathbf{W})$ in (129).

$$\zeta(\mathbf{W}) = \begin{cases} \left| -\frac{1}{K} \frac{1}{H(\mathbf{W})} \phi'(\mathbf{w}_j^\top \mathbf{x}) \right| \leq \frac{\phi(\mathbf{w}_j^\top \mathbf{x})(1 - \phi(\mathbf{w}_j^\top \mathbf{x}))}{K \cdot \frac{1}{K} \phi(\mathbf{w}_j^\top \mathbf{x})} \leq 1, & y = 1 \\ \left| \frac{1}{K} \frac{1}{1 - H(\mathbf{W})} \phi'(\mathbf{w}_j^\top \mathbf{x}) \right| \leq \frac{\phi(\mathbf{w}_j^\top \mathbf{x})(1 - \phi(\mathbf{w}_j^\top \mathbf{x}))}{K \cdot \frac{1}{K} (1 - \phi(\mathbf{w}_j^\top \mathbf{x}))} \leq 1, & y = 0 \end{cases} \quad (250)$$

Then we have an upper bound of G'_i .

$$\begin{aligned} |G'_i| &= \left| \zeta_{j,l} \mathbf{v}^\top \mathbf{x} - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\zeta \mathbf{v}^\top \mathbf{x}] \right| \\ &\leq |\mathbf{v}^\top \mathbf{x}| + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [|\mathbf{v}^\top \mathbf{x}|] \end{aligned} \quad (251)$$

Following the idea of (232) and (233), and by $\mathbf{v} \in V_{\frac{1}{2}}$, we have

$$\mathbb{E}[|G'_i|^p] \leq O\left(\left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right)^{\frac{p}{2}}\right) \quad (252)$$

$$\mathbb{E}[\exp(\theta G'_i)] \leq 1 + O\left(|e\theta^2| \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right) \quad (253)$$

where (253) holds if $\theta \leq \frac{2\tau}{4e} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|)^2}$. Following the derivation of (230) and (235) to (238), we have

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n G'_i\right| \geq \frac{t}{6}\right) \\ & \leq 2 \exp\left(\max\left\{-\frac{C_{14} n t^2}{144 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}, -C_{15} n \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \cdot t\right\}\right) \end{aligned} \quad (254)$$

for some constant $C_{14} > 0$ and $C_{15} > 0$. Moreover, we can obtain $\mathbb{P}(B'_t) \leq \frac{\delta}{2}$ as long as

$$t \geq C_{13} \cdot \max\left\{ \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{dK \log \frac{18r}{\epsilon} + \log \frac{4}{\delta}}{n}}, \frac{dK \log \frac{18r}{\epsilon} + \log \frac{4}{\delta}}{\sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot n}} \right\} \quad (255)$$

(b) For the upper bound of $\mathbb{P}(A'_t)$ and $\mathbb{P}(C'_t)$, we can first derive

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{\|\nabla \ell(\mathbf{W}, \mathbf{x}) - \nabla \ell(\mathbf{W}', \mathbf{x})\|}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \\ & \leq \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{|\zeta(\mathbf{W}) - \zeta(\mathbf{W}')| \cdot \|\mathbf{x}\|}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \\ & \leq \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{\max_{1 \leq j, l \leq K} \{|\xi_{j,l}(\mathbf{W}'')|\} \cdot \|\mathbf{x}\|^2 \sqrt{K} \|\mathbf{W} - \mathbf{W}'\|_F}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \\ & \leq \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[\sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{C_9 \cdot \|\mathbf{x}\|^2 \sqrt{K} \|\mathbf{W} - \mathbf{W}'\|_F}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \\ & \leq C_9 \cdot 3\sqrt{K}d \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \end{aligned} \quad (256)$$

The first inequality is by (129). The second inequality is by the Mean Value Theorem. The third step is by (193). The last inequality is by Property 7. Therefore, following the steps in part (2) of Lemma 8, we can conclude that C'_t holds if

$$t \geq 3C_9 \cdot \sqrt{K}d \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot \epsilon \quad (257)$$

Moreover, from (242) in Lemma 8 we have that

$$t \geq \frac{18C_9 \cdot \sqrt{K}d \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2 \cdot \epsilon}{\delta} \quad (258)$$

ensures $\mathbb{P}(A'_t) \leq \frac{\delta}{2}$. Therefore, let $\epsilon = \frac{\delta}{18C_9 \cdot \sqrt{K}d \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2 \cdot \epsilon \cdot ndK}$, $\delta = d^{-10}$ and $t = C_{13} \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|)^2} \sqrt{\frac{d \log n}{n}}$, if $n \geq C'' \cdot dK \log dK$ for some constant $C'' > 0$, we have

$$\mathbb{P}\left(\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla f_n(\mathbf{W}) - \nabla \bar{f}(\mathbf{W})\| \geq C_{13} \cdot \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|)^2} \sqrt{\frac{d \log n}{n}} \leq d^{-10} \right) \quad (259)$$

By Hoeffding's inequality in [76] and Property 2, we have

$$\begin{aligned} & \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \|\nu_i\|_F \geq C_{13} \cdot \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{dK \log n}{n}} \xi \right) \\ & \lesssim \exp\left(-C_{13}^2 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \frac{\xi^2 dK \log n}{dK \xi^2}\right) \\ & \lesssim d^{-10} \end{aligned} \quad (260)$$

Therefore,

$$\begin{aligned} & \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla \tilde{f}_n(\mathbf{W}) - \nabla \tilde{f}(\mathbf{W})\| \\ & \leq C_{13} \cdot \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{d \log n}{n}} + \frac{1}{n} \sum_{i=1}^n \|\nu_i\| \\ & \leq C_{13} \cdot \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{d \log n}{n}} + \frac{1}{n} \sum_{i=1}^n \|\nu_i\|_F \\ & \leq C_{13} \cdot \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{d \log n}{n}} (1 + \xi) \end{aligned} \quad (261)$$

N. Proof of Lemma 3 and its supportive lemmas

We need Lemma 10 to Lemma 14, which are stated in Section D, for the proof of Lemma 3. Section D2 summarizes the proof of Lemma 3. The proofs of Lemma 10 to Lemma 12 are provided in Section N1 to Section N3. Lemma 13 and Lemma 14 are cited from [26]. Although [26] considers the standard Gaussian distribution, the proofs of Lemma 13 and 14 hold for any data distribution. Therefore, these two lemmas can be applied here directly.

The tensor initialization in [26] only holds for the standard Gaussian distribution. We exploit a more general definition of tensors from [82] for the tensor initialization in our algorithm. We also develop new error bounds for the initialization.

1) *Proof of Lemma 10:* From Assumption 1, if the Gaussian Mixture Model is a symmetric probability distribution defined in (64), then by Definition 3, we have

$$\begin{aligned}
& \|\widehat{\mathbf{Q}}_3(\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha}) - \mathbf{Q}_3(\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha})\| \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \left[y_i \cdot p(\mathbf{x})^{-1} \sum_{l=1}^L \lambda_l (2\pi |\boldsymbol{\Sigma}_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l)\right) \right. \right. \\
&\quad \cdot \left. \left(((\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1})^{\otimes 3} - ((\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1}) \tilde{\otimes} \boldsymbol{\Sigma}_l^{-1} \right) \right] (\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha}) \\
&\quad - \mathbb{E} \left[y \cdot p(\mathbf{x})^{-1} \sum_{l=1}^L \lambda_l (2\pi |\boldsymbol{\Sigma}_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l)\right) \right. \\
&\quad \cdot \left. \left(((\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1})^{\otimes 3} - ((\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1}) \tilde{\otimes} \boldsymbol{\Sigma}_l^{-1} \right) \right] (\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha}) \Big\|
\end{aligned} \tag{262}$$

Following [26], $\tilde{\otimes}$ is defined such that for any $\mathbf{v} \in \mathbb{R}^{d_1}$ and $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$,

$$\mathbf{v} \tilde{\otimes} \mathbf{Z} = \sum_{i=1}^{d_2} (\mathbf{v} \otimes \mathbf{z}_i \otimes \mathbf{z}_i + \mathbf{z}_i \otimes \mathbf{v} \otimes \mathbf{z}_i + \mathbf{z}_i \otimes \mathbf{z}_i \otimes \mathbf{v}), \tag{263}$$

where \mathbf{z}_i is the i -th column of \mathbf{Z} . By Definition 3, we have

$$\begin{aligned}
& \left\| \left[y \cdot p(\mathbf{x})^{-1} \sum_{l=1}^L \lambda_l (2\pi |\boldsymbol{\Sigma}_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l)\right) \right. \right. \\
&\quad \cdot \left. \left(((\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1})^{\otimes 3} - ((\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1}) \tilde{\otimes} \boldsymbol{\Sigma}_l^{-1} \right) \right] (\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha}) \Big\| \\
&\lesssim \left\| \frac{\sum_{l=1}^L \lambda_l (2\pi |\boldsymbol{\Sigma}_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l)\right) \cdot ((\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1})^{\otimes 2} (\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l))}{\sum_{l=1}^L \lambda_l (2\pi |\boldsymbol{\Sigma}_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l)\right)} \right\| \\
&\lesssim \|\sigma_{\min}^{-6}(\mathbf{x}^\top \boldsymbol{\alpha}) \mathbf{x} \mathbf{x}^\top\|
\end{aligned} \tag{264}$$

The first step of (264) is because $(\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1})^{\otimes 2} (\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l))$ is the dominant term of the entire expression, and $y \leq 1$. The second step is because the expression can be considered as a normalized weighted summation of $((\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1})^{\otimes 2} (\boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l))$ and $(\mathbf{x}^\top \boldsymbol{\alpha}) \mathbf{x} \mathbf{x}^\top$ is its dominant term. Define $S_m(\mathbf{x}) = (-1)^m \frac{\nabla_{\mathbf{x}}^m p(\mathbf{x})}{p(\mathbf{x})}$, where $p(\mathbf{x})$ is the probability density function of the random variable \mathbf{x} . From Definition 3, we can verify that

$$\mathbf{Q}_j = \mathbb{E}[y \cdot S_m(\mathbf{x})] \quad j \in \{1, 2, 3\} \tag{265}$$

Then define $Gp_i = \langle \mathbf{v}, ([y_i \cdot S_3(\mathbf{x}_i)](\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \mathbb{E}[y_i \cdot S_3(\mathbf{x}_i)](\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})) \mathbf{v} \rangle$, where $\|\mathbf{v}\| = 1$, then $\mathbb{E}[Gp_i] = 0$. Similar to the proof of (231), (232), and (233) in Lemma 8, we have

$$|Gp_i|^p \lesssim |\sigma_{\min}^{-6}(\mathbf{x}_i^\top \boldsymbol{\alpha}) (\mathbf{x}_i^\top \mathbf{v})^2 + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-6}(\mathbf{x}_i^\top \boldsymbol{\alpha}) (\mathbf{x}_i^\top \mathbf{v})^2]|^p \tag{266}$$

$$\begin{aligned}
\mathbb{E}[|Gp_i|^p] &\lesssim (\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-6}(\mathbf{x}_i^\top \boldsymbol{\alpha}) (\mathbf{x}_i^\top \mathbf{v})^2])^p \\
&\leq \sigma_{\min}^{-6p} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{x}^\top \boldsymbol{\alpha})^2]^{\frac{p}{2}} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{x}^\top \mathbf{v})^4]^{\frac{p}{2}} \\
&\leq \tau^{6p} \sqrt{D_2(\Psi) D_4(\Psi)^p}
\end{aligned} \tag{267}$$

$$\begin{aligned}
\mathbb{E}[\exp(\theta Gp_i)] &\lesssim 1 + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|Gp_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty} \frac{|e\theta|^p \tau^{6p} (D_2(\Psi) D_4(\Psi))^{\frac{p}{2}}}{p^p} \\
&\lesssim 1 + \theta^2 \tau^{12} D_2(\Psi) D_4(\Psi)
\end{aligned} \tag{268}$$

Hence, similar to the derivation of (235), we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Gp_i \geq t\right) \leq \exp\left(-n\theta t + C_{16}n\theta^2(\tau^6 \sqrt{D_2(\Psi)D_4(\Psi)})^2\right) \quad (269)$$

for some constant $C_{16} > 0$. Let $\theta = \frac{t}{2C_{16}(\tau^6 \sqrt{D_2(\Psi)D_4(\Psi)})^2}$ and $t = \delta_1^2(\mathbf{W}^*) \cdot (\tau^6 \sqrt{D_2(\Psi)D_4(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}}$, then we have

$$\|\widehat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\| \leq \delta_1(\mathbf{W}^*)^2 \cdot (\tau^6 \sqrt{D_2(\Psi)D_4(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}} \quad (270)$$

with probability at least $1 - 2n^{-\Omega(\delta_1^4(\mathbf{W}^*)d)}$.

If the Gaussian Mixture Model is not a symmetric distribution which is defined in (64), we would have a similar result as follows.

$$\|\widehat{\mathbf{Q}}_2 - \mathbf{Q}_2\| = \left\| \frac{1}{n} \sum_{i=1}^n [y_i \cdot S_2(\mathbf{x}_i)] - \mathbb{E}[y \cdot S_2(\mathbf{x})] \right\| \quad (271)$$

$$\|y_i \cdot S_2(\mathbf{x}_i)\| \lesssim \|\sigma_{\min}^{-4} \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*\top} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\top\| \quad (272)$$

Then define $Gp'_i = \langle \mathbf{v}, ([y_i \cdot S_2(\mathbf{x}_i)] - \mathbb{E}[y_i \cdot S_2(\mathbf{x}_i)]) \mathbf{v} \rangle$, where $\|\mathbf{v}\| = 1$, then $\mathbb{E}[Gp'_i] = 0$. Similar to the proof of (231), (232) and (233) in Lemma 8, we have

$$|Gp'_i|^p \lesssim |\sigma_{\min}^{-4}(\mathbf{x}_i^\top \mathbf{v})|^2 + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-4}(\mathbf{x}_i^\top \mathbf{v})^2]^p \quad (273)$$

$$\mathbb{E}[|Gp'_i|^p] \lesssim (\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-4}(\mathbf{x}_i^\top \mathbf{v})^2])^p \leq \tau^{4p} D_2(\Psi)^p \quad (274)$$

$$\begin{aligned} \mathbb{E}[\exp(\theta Gp'_i)] &\lesssim 1 + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|Gp'_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty} \frac{|e\theta|^p \tau^{4p} D_2(\Psi)^p}{p^p} \\ &\lesssim 1 + \theta^2 \tau^8 D_2(\Psi)^2 \end{aligned} \quad (275)$$

Hence, similar to the derivation of (235), we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Gp_i \geq t\right) \leq \exp\left(-n\theta t + C_{17}n\theta^2(\tau^4 D_2(\Psi))^2\right) \quad (276)$$

for some constant $C_{17} > 0$. Let $\theta = \frac{t}{2C_{17}(\tau^4 D_2(\Psi))^2}$ and $t = \delta_1^2(\mathbf{W}^*) \cdot (\tau^4 D_2(\Psi)) \cdot \sqrt{\frac{d \log n}{n}}$, then we have

$$\begin{aligned} \|\widehat{\mathbf{Q}}_2 - \mathbf{Q}_2\| &\lesssim \delta_1^2(\mathbf{W}^*) \cdot \tau^4 D_2(\Psi) \cdot \sqrt{\frac{d \log n}{n}} \\ &\lesssim \sqrt{\frac{d \log n}{n}} \cdot \delta_1^2(\mathbf{W}^*) \cdot \tau^6 \sqrt{D_2(\Psi)D_4(\Psi)} \end{aligned} \quad (277)$$

with probability at least $1 - 2n^{-\Omega(\delta_1^4(\mathbf{W}^*)d)}$.

2) *Proof of Lemma 11:* We consider each component of $y = \frac{1}{K} \sum_{i=1}^K \phi(\mathbf{w}_i^{*\top} \mathbf{x})$. Define $\mathbf{T}_i(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{K \times K \times K}$ such that

$$\mathbf{T}_i(\mathbf{x}) = [\phi(\mathbf{w}_i^{*\top} \mathbf{x}) \cdot S_3(\mathbf{x})](\widehat{\mathbf{U}}, \widehat{\mathbf{U}}, \widehat{\mathbf{U}}) \quad (278)$$

We flatten $\mathbf{T}_i(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{K \times K \times K}$ along the first dimension to obtain the function $\mathbf{B}_i(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{K \times K^2}$. Similar to the derivation of the last step of Lemma E.8 in [26], we can obtain $\|\mathbf{T}_i(\mathbf{x})\| \leq \|\mathbf{B}_i(\mathbf{x})\|$. By (262), we have

$$\|\mathbf{B}_i(\mathbf{x})\| \lesssim \sigma_{\min}^{-6} \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*\top} \mathbf{x}_i) (\widehat{\mathbf{U}}^\top \mathbf{x})^3 \quad (279)$$

Define $Gr_i = \langle \mathbf{v}, \mathbf{B}_i(\mathbf{x}_i) \rangle - \mathbb{E}[\langle \mathbf{B}_i(\mathbf{x}_i), \mathbf{v} \rangle]$, where $\|\mathbf{v}\| = 1$, so $\mathbb{E}[Gr_i] = 0$. Similar to the proof of (231), (232) and (233) in Lemma 8, we have

$$|Gr_i|^p \lesssim |\sigma_{\min}^{-6}(\mathbf{v}^\top \widehat{\mathbf{U}}^\top \mathbf{x})^3 + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-6}(\mathbf{v}^\top \widehat{\mathbf{U}}^\top \mathbf{x})^3]|^p \quad (280)$$

$$\mathbb{E}[|Gr_i|^p] \lesssim (\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-6}(\mathbf{v}^\top \widehat{\mathbf{U}}^\top \mathbf{x})^3])^p \lesssim \tau^{6p} \sqrt{D_6(\Psi)}^p \quad (281)$$

$$\begin{aligned} \mathbb{E}[\exp(\theta Gr_i)] &\lesssim 1 + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|Gr_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty} \frac{|\theta|^p \tau^{6p} D_6(\Psi)^{\frac{p}{2}}}{p^p} \\ &\leq 1 + \theta^2 (\tau^{12} \sqrt{D_6(\Psi)})^2 \end{aligned} \quad (282)$$

Hence, similar to the derivation of (235), we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Gr_i \geq t\right) \leq \exp\left(-n\theta t + C_{18} \theta^2 (\tau^6 \sqrt{D_6(\Psi)})^2\right) \quad (283)$$

for some constant $C_{18} > 0$. Let $\theta = \frac{t}{C_{18} (\tau^6 \sqrt{D_6(\Psi)})^2}$ and $t = \delta_1^2(\mathbf{W}^*) \cdot (\tau^6 \sqrt{D_6(\Psi)}) \cdot \sqrt{\frac{\log n}{n}}$, then we have

$$\|\widehat{\mathbf{R}}_3 - \mathbf{R}_3\| \lesssim \delta_1(\mathbf{W}^*)^2 \cdot (\tau^6 \sqrt{D_6(\Psi)}) \cdot \sqrt{\frac{\log n}{n}} \quad (284)$$

with probability at least $1 - 2n^{-\Omega(\delta_1^4(\mathbf{W}^*))}$.

3) *Proof of Lemma 12:* From Definition 3, we have

$$\|\widehat{\mathbf{Q}}_1 - \mathbf{Q}_1\| = \left\| \frac{1}{n} \sum_{i=1}^n [y_i \cdot S_1(\mathbf{x}_i)] - \mathbb{E}[y \cdot S_1(\mathbf{x})] \right\|. \quad (285)$$

Based on Definition 3,

$$\begin{aligned} \left\| [y_i \cdot S_1(\mathbf{x}_i)] \right\| &\lesssim \left\| \frac{\sum_{l=1}^L \lambda_l \lambda_l (2\pi \prod_{k=1}^d \sigma_{lk}^2)^{-\frac{d}{2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l)) \cdot (\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1}}{\sum_{l=1}^L \lambda_l \lambda_l (2\pi \prod_{k=1}^d \sigma_{lk}^2)^{-\frac{d}{2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l))} \right\| \\ &\lesssim \left\| \sigma_{\min}^{-2} \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^* \top \mathbf{x}_i) \mathbf{x}_i \right\| \end{aligned} \quad (286)$$

Define $Gq_i = \langle \mathbf{v}, ([y_i \cdot S_1(\mathbf{x}_i)] - \mathbb{E}[y_i \cdot S_1(\mathbf{x}_i)]) \mathbf{v} \rangle$, where $\|\mathbf{v}\| = 1$, so $\mathbb{E}[Gq_i] = 0$. Similar to the proof of (231), (232), and (233) in Lemma 8, we have

$$|Gq_i|^p \lesssim |\sigma_{\min}^{-2}(\mathbf{x}_i^\top \mathbf{v}) + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-2}(\mathbf{x}_i^\top \mathbf{v})]|^p \quad (287)$$

$$\mathbb{E}[|Gq_i|^p] \lesssim (\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-2}(\mathbf{x}_i^\top \mathbf{v})])^p \leq \tau^{2p} \sqrt{D_2(\Psi)}^p \quad (288)$$

$$\begin{aligned} \mathbb{E}[\exp(\theta Gq_i)] &\lesssim 1 + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|Gq_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty} \frac{|\theta|^p \tau^{2p} D_2(\Psi)^{\frac{p}{2}}}{p^p} \\ &\leq 1 + \theta^2 (\tau^2 \sqrt{D_2(\Psi)})^2 \end{aligned} \quad (289)$$

Hence, similar to the derivation of (235), we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Gq_i \geq t\right) \leq \exp\left(-n\theta t + C_{19} \theta^2 (\tau^2 \sqrt{D_2(\Psi)})^2\right) \quad (290)$$

for some constant $C_{19} > 0$. Let $\theta = \frac{t}{C_{19} (\tau^2 \sqrt{D_2(\Psi)})^2}$ and $t = (\tau^2 \sqrt{D_2(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}}$, then we have

$$\|\widehat{\mathbf{Q}}_1 - \mathbf{Q}_1\| \lesssim (\tau^2 \sqrt{D_2(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}} \quad (291)$$

with probability at least $1 - 2n^{-\Omega(d)}$.