LEARNING ONE-HIDDEN-LAYER NEURAL NETWORKS ON GAUSSIAN MIXTURE MODELS WITH GUARANTEED GENERALIZABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We analyze the learning problem of fully connected neural networks with the sigmoid activation function for binary classification from the setup of model estimation. The outputs are assumed to be generated by a ground-truth neural network with the unknown parameters, and the learning objective is to estimate the ground-truth model parameters by minimizing a non-convex cross-entropy loss function of the training data. Instead of following the conventional and restrictive assumption in the literature that the input features follow the standard Gaussian distribution, this paper, for the first time, analyzes a more general and practical scenario that the input features follow a Gaussian mixture model of a finite number of Gaussian distributions of various mean and variance. We propose a gradient descent algorithm with a tensor initialization approach and show that our algorithm converges linearly to a critical point that has a diminishing distance to the ground-truth model with guaranteed generalizability. We characterize the required number of samples for successful convergence, referred to as the sample complexity, as a function of the parameters of the Gaussian mixture model. We prove analytically that when any mean or variance in the mixture model is large, or when all variances are close to zero, the sample complexity increases, and the convergence slows down, indicating a more challenging learning problem. Although focusing on one-hidden-layer neural networks, this paper provides the first theoretical analyses of the impact of the parameters of the input distributions on the learning performance.

1 INTRODUCTION

Deep Neural networks (LeCun et al., 2015) have demonstrated superior empirical performance in various applications such as speech recognition (Krizhevsky et al., 2012) and computer vision (Graves et al., 2013; He et al., 2016). Despite the numerical success, the theoretical underpin of leaning neural networks is much less investigated. One bottleneck for the wide acceptance of deep learning in critical applications is the lack of the theoretical generalization guarantee, i.e., why a model learned from the training data would achieve a high accuracy on the testing data.

One line of theoretical research on the learning performance considers the over-parameterized setting where the number of network parameters is greater than the number of training samples. (Bousquet & Elisseeff, 2002; Hardt et al., 2016; Keskar et al., 2016; Livni et al., 2014; Neyshabur et al., 2017; Rumelhart et al., 1988; Soltanolkotabi et al., 2018; Allen-Zhu et al., 2019b). (Allen-Zhu et al., 2019c; Du et al., 2019; Zou & Gu, 2019) show the deep neural network can fit the training samples in polynomial time. The optimization problem has no spurious local minima Livni et al. (2014); Zhang et al. (2016); Soltanolkotabi et al. (2018), and the global minimum of the empirical risk function can be obtained by gradient descent. (Allen-Zhu et al., 2019a; Li & Liang, 2018) analyze the generalization error by characterizing the training error and the test error separately. Still, there is no guarantee that a learned model with a small training error would have a small test error.

Another line of works analyzes the generalizability of neural networks from the perspective of model estimation (Zhong et al., 2017b;a; Zhang et al., 2019; 2020b; Fu et al., 2020; Zhang et al., 2020a). Assuming the existence of a ground-truth model that maps the input features to the output label,

these works aim to estimate the unknown network weights of a given network architecture from the training samples. If the ground-truth model can be found in training, it has the guaranteed test performance. These works focus on one-hidden-layer networks, because the analysis is already complicated for one-hidden-layer, and the complexity increases tremendously when the number of hidden layers increases. (Brutzkus & Globerson, 2017; Du et al., 2018; Ge et al., 2018; Liang et al., 2018; Li & Yuan, 2017; Shamir, 2018; Safran & Shamir, 2018; Tian, 2017) consider the ideal case of infinite number of training samples so that the training and test accuracy coincide and can be analyzed simultaneously. When the number of training samples is finite, (Zhong et al., 2017b;a) characterize the sample complexity, i.e., the required number of samples, of learning one-hidden-layer fully connected neural networks with smooth activation functions and propose a gradient descent algorithm that converges to the ground-truth model linearly. (Zhang et al., 2019; 2020b) extend the analyses to the non-smooth ReLU for fully-connected and convolutional neural networks, respectively. (Zhang et al., 2020a) analyzes the generalizability of graph neural networks for both regression and binary classification problems. (Fu et al., 2020) analyzes the cross entropy loss function for binary classification problems. Compared with other common loss functions such as the squared loss, the cross entropy loss function is harder to analyze due to the complicated forms and saturation phenomenon of its Gradient and Hessian (Fu et al., 2020).

One critical assumption of almost all works in this line is that the input features follow the standard Gaussian distribution. (Du et al., 2017) extends to rotationally invariant distributions, but the results only apply to the case of only one neuron in the hidden layer. Although these assumptions enable the algorithmic design and theoretical analysis, they do not apply to most practical datasets. Moreover, the resulting theoretical results cannot characterize the impact of input feature characteristics such as mean and variance on the learning performance, while the property of the training data is critical in the success of learning in practice.

Contributions This paper provides the *first* theoretical analysis of learning one-hidden-layer neural networks when the input distribution follows a Gaussian mixture model containing an arbitrary number of Gaussian distributions with arbitrary mean and variance. Gaussian mixture model is much more general than the standard Gaussian and has been employed to characterize the practical datasets in many applications such as data clustering and unsupervised learning (Dasgupta, 1999; Figueiredo & Jain, 2002; Jain, 2010), image classification and segmentation (Permuter et al., 2006), etc. For the binary classification problem with the cross entropy loss function, this paper proposes a gradient descent algorithm with tensor initialization to estimate the weights of the one-hidden-layer fully-connected neural network. Our algorithm converges to a critical point linearly, and the returned critical point converges to the ground-truth model at a rate of $\sqrt{d \log n / n}$, where d is the feature dimension, and n is the number of samples. We also characterize the sample complexity for accurate estimation as a function of d, the number of neurons K, and the input distribution. One interesting finding is the when any mean or variance in the Gaussian mixture model increases, the sample complexity increases, and the algorithm converges slower, indicating that it will be more challenging to learn a model with a small test error. The same phenomenon happens when all the variance in the mixture model approaches zero. To the best of our knowledge, *this is the first theoretical characterization about how the input distribution affects the learning performance*.

Notations: Vectors are bold lowercase, matrices and tensors are bold uppercase. Scalars are in normal fonts. For instance, $\boldsymbol{Z}$ is a matrix, and $\boldsymbol{z}$ is a vector. z_i denotes the i-th entry of $\boldsymbol{z}$, and $Z_{i,j}$ denotes the (i,j)-th entry of $\boldsymbol{Z}$. $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ stand for the sets of all natural numbers, all integers and all real numbers. $[K]$ $(K > 0)$ denotes the set including integers from 1 to K. $\boldsymbol{I}_d \in \mathbb{R}^{d \times d}$ and $\boldsymbol{e}_i$ represent the identity matrix in $\mathbb{R}^{d \times d}$ and the i-th standard basis vector, respectively. $\boldsymbol{Z}^\top$ denote the transpose of $\boldsymbol{Z}$, similarly for $\boldsymbol{z}$. $||\boldsymbol{z}||$ (or $||\boldsymbol{z}||_2$) stands for the ℓ-2 norm of a vector $\boldsymbol{z}$, and $||\boldsymbol{Z}||$ (or $||\boldsymbol{Z}||_2$) and $||\boldsymbol{Z}||_F$ represent the spectral norm and the Frobenius norm of the matrix $\boldsymbol{z}$, respectively. We use $\delta_i(\boldsymbol{Z})$ to denote the i-th largest singular value of $\boldsymbol{Z}$. For a positive semi-definite (PSD) matrix $\boldsymbol{A}$, we write $\boldsymbol{A} \succeq 0$. The gradient and the Hessian of a function $f(\boldsymbol{W})$ is denoted by $\nabla f(\boldsymbol{W})$ and $\nabla^2 f(\boldsymbol{W})$, respectively. Moreover, the outer product of a group of vectors $\boldsymbol{z}_i \in \mathbb{R}^{n_i}$, $i \in [l]$, is defined as $\boldsymbol{T} = \boldsymbol{z}_1 \otimes \cdots \times \boldsymbol{z}_l \in \mathbb{R}^{n_1 \times \cdots \times n_l}$ with $\boldsymbol{T}_{j_1 \cdots j_l} = (\boldsymbol{z}_1)_{j_1} \cdots (\boldsymbol{z}_l)_{j_l}$.
Given a tensor $\boldsymbol{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and three matrices $\boldsymbol{A} \in \mathbb{R}^{n_1 \times d_1}$, $\boldsymbol{B} \in \mathbb{R}^{n_2 \times d_2}$, $\boldsymbol{C} \in \mathbb{R}^{n_3 \times d_3}$, the (i_1, i_2, i_3)-th entry of the tensor $\boldsymbol{T}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ is given by

$$\sum_{i_1'}^{n_1} \sum_{i_2'}^{n_2} \sum_{i_3'}^{n_3} \boldsymbol{T}_{i_1', i_2', i_3'} \boldsymbol{A}_{i_1', i_1} \boldsymbol{B}_{i_2', i_2} \boldsymbol{C}_{i_3', i_3}. \tag{1}$$

2 PROBLEM FORMULATION

We consider a one-hidden-layer fully-connected neural network for the binary classification problem. Let $\boldsymbol{x} \in \mathbb{R}^d$ denote the input features. Let $K \geq 1$ be the number of neurons in the hidden layer. Following the setup of model estimation, see e.g., (Fu et al., 2020), we assume the existence of ground-truth weights $\boldsymbol{w}_j^* \in \mathbb{R}^d$ ($j \in [K]$) that maps the input features to output labels. Let $\boldsymbol{W}^* = [\boldsymbol{w}_1^*, ..., \boldsymbol{w}_K^*] \in \mathbb{R}^{d \times K}$ contain all the weights. Let $\delta_i(\boldsymbol{W}^*)$ denote the i-th largest singular value of $\boldsymbol{W}^*$. Let κ denote the number that $\kappa = \frac{\delta_1(\boldsymbol{W}^*)}{\delta_K(\boldsymbol{W}^*)}$, and define $\eta = \prod_{i=1}^K \frac{\delta_i(\boldsymbol{W}^*)}{\delta_K(\boldsymbol{W}^*)}$. The nonlinear activation function is the sigmoid function $\phi(x) = \frac{1}{1+\exp(-x)}$. The binary output y is generated by

$$\mathbb{P}(y = 1|\boldsymbol{x}) = H(\boldsymbol{W}^*, \boldsymbol{x}) := \frac{1}{K} \sum_{j=1}^K \phi({\boldsymbol{w}_j^*}^\top \boldsymbol{x}). \tag{2}$$

Given n pairs of training samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, the empirical risk function is

$$f_n(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{W}; \boldsymbol{x}_i, y_i) \tag{3}$$

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

$$\ell(\boldsymbol{W}; \boldsymbol{x}_i, y_i) = -y \cdot \log(H(\boldsymbol{W}, \boldsymbol{x})) - (1 - y) \cdot \log(1 - H(\boldsymbol{W}, \boldsymbol{x})). \tag{4}$$

To recover the ground truth parameters $\boldsymbol{W}^*$ from training samples, we solve the following nonconvex minimization problem

$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times K}} f_n(\boldsymbol{W}). \tag{5}$$

Here we assume the input features $\boldsymbol{x}_i$ are generated i.i.d. from the Gaussian mixture model (Pearson, 1894; Titterington et al., 1985; Hsu & Kakade, 2013), which we denote as

$$\boldsymbol{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d), \tag{6}$$

where $\mathcal{N}$ denotes the multi-variate Gaussian distribution with mean $\boldsymbol{\mu}_l \in \mathbb{R}^d$, and covariance $\sigma_l \boldsymbol{I}_d$ for $\sigma_l \in \mathbb{R}$ for all $l \in [L]$. The Gaussian mixture model can be viewed as

$$\boldsymbol{x} := \boldsymbol{\mu}_h + \boldsymbol{z}_h \in \mathbb{R}^d \tag{7}$$

where h is a discrete random variable with $\Pr(h = l) = \lambda_l$ for $l \in [L]$, and $\boldsymbol{z}_l$ follows the multivariate Gaussian $\mathcal{N}(\boldsymbol{0}, \sigma_l^2 \boldsymbol{I}_d)$ with zero mean and covariance $\sigma_l^2 \boldsymbol{I}_d$.

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

$$\boldsymbol{x} \sim \begin{cases} \sum_{l=1}^{\frac{L}{2}} \lambda_l \big(\mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d) + \mathcal{N}(-\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)\big) & L \text{ is even} \\ \lambda_1 \mathcal{N}(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I}_d) + \sum_{l=2}^{\frac{L-1}{2}} \lambda_l \big(\mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d) + \mathcal{N}(-\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)\big) & L \text{ is odd} \end{cases} \tag{8}$$

We assume without loss of generality that $\boldsymbol{\mu}_l$ belongs to the column space of $\boldsymbol{W}^*$ for all $l \in [L]$. To see this, note that an arbitrary $\boldsymbol{\mu}_l$ can be written as $\boldsymbol{\mu}_{l\|} + \boldsymbol{\mu}_{l\perp}$, where $\boldsymbol{\mu}_{l\|}$ belongs to the column space of $\boldsymbol{W}^*$, and $\boldsymbol{\mu}_{l\perp}$ is perpendicular to the column space. Then, from (2) and (7) we have

$$H(\boldsymbol{W}^*, \boldsymbol{x}) = \frac{1}{K} \sum_{j=1}^K \phi({\boldsymbol{w}_j^*}^\top (\boldsymbol{\mu}_{h\|} + \boldsymbol{\mu}_{h\perp} + \boldsymbol{z}_h)) = \frac{1}{K} \sum_{j=1}^K \phi({\boldsymbol{w}_j^*}^\top (\boldsymbol{\mu}_{h\|} + \boldsymbol{z}_h)) = H(\boldsymbol{W}^*, \boldsymbol{x}') \tag{9}$$

where $\boldsymbol{x}' \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_{l\|}, \sigma_l^2 \boldsymbol{I}_d)$. Thus, these two cases are equivalent.

The parameters λ_l, $\boldsymbol{\mu}_l$ and σ_l, $i \in [L]$ of the Gaussian Mixture Model are assumed to be known a priori throughout the paper. In practice, one can estimate these parameters by methods such as the EM algorithm (Redner & Walker, 1984) and the moment-based method (Hsu & Kakade, 2013).

3 PROPOSED LEARNING ALGORITHM

We propose Algorithm 1 to solve (5) and defer its theoeretical analysis to Section 4. The method starts from a initialization $\boldsymbol{W}_0 \in \mathbb{R}^{d\times K}$ computed based on the tensor initialization method (Subroutine 1) and then updates the iterates $\boldsymbol{W}_t$ using gradient descent with a step size η_0.

Algorithm 1 Our proposed learning algorithm

Input: Training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the step size η_0, iteration T
Initialization: $\boldsymbol{W}_0 \leftarrow$ Tensor initialization method via Subroutine 1
Gradient Descent: for $t = 0, 1, \cdots, T-1$

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta_0 \nabla f_n(\boldsymbol{W}_t)$$

Output: $\boldsymbol{W}_T$

Our tensor initialization method is extended from (Janzamin et al., 2014) and (Zhong et al., 2017b) for the standard Gaussian distribution. The idea is to design tensors based on the training data to estimate the column space of $\boldsymbol{W}^*$, the direction and magnitude of each $\boldsymbol{w}_j^*$, respectively. Because the existing method only applies to the standard Gaussian, we exploit the relationship between probability density functions and tensor expressions developed in (Janzamin et al., 2014) to design tensors suitable for the Gaussian mixture model. Formally,

Definition 1 *Let* $p(\boldsymbol{x}) = \sum_{l=1}^L \lambda_l (2\pi\sigma_l)^{-\frac{d}{2}} \exp(-\frac{||\boldsymbol{x}-\boldsymbol{\mu}_l||^2}{2\sigma_l^2})$ *be the probability density function of the Gaussian mixture model in (6). We define*

$$\boldsymbol{M}_j := \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I})}[y \cdot (-1)^j \frac{\nabla^{(m)} p(\boldsymbol{x})}{p(\boldsymbol{x})}],\ j = 1, 2, 3 \tag{10}$$

Let $\boldsymbol{\alpha} \in \mathbb{R}^d$ *denote an arbitrary vector. If the Gaussian Mixture Model is symmetric as in (8), then* $\boldsymbol{P}_2 := \boldsymbol{M}_3(\boldsymbol{I}_d, \boldsymbol{I}_d, \boldsymbol{\alpha})$. *Otherwise,* $\boldsymbol{P}_2 := \boldsymbol{M}_2$.

Note that $\boldsymbol{M}_i$ $(i = 1, 2, 3)$ and $\boldsymbol{P}_2$ depend on $\boldsymbol{W}^*$ through having y in the definition and cannot be directly computed. We estimate these values by sample means, denoted by $\widehat{\boldsymbol{M}}_i$ $(i = 1, 2, 3)$ and $\widehat{\boldsymbol{P}}_2$, from samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$.We also make the following assumption, which guarantees that $\boldsymbol{M}_i$ $(i = 1, 2, 3)$ and $\boldsymbol{P}_2$ are all nonzero, and thus can be leveraged to estimate $\boldsymbol{W}^*$.

Assumption 1 *The Gaussian Mixture Model in (6) satisfies the following conditions:*

1. $\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I})}[\phi'''(\boldsymbol{w}_i^{*\top}\boldsymbol{x})] \neq 0$ *for* $i \in [K]$
2. *If the Gaussian Mixture Model is not symmetric, then* $\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I})}[\phi''(\boldsymbol{w}_i^{*\top}\boldsymbol{x})] \neq 0$ *for* $i \in [K]$

Note that Assumption 1 is a very mild assumption. Moreover, as indicted in (Janzamin et al., 2014), in the rare case that some quantities $\boldsymbol{M}_i$ $(i = 1, 2, 3)$ and $\boldsymbol{P}_2$ are zero, one can construct higher-order tensors in a similar way as in Definition 1 and then estimate $\boldsymbol{W}^*$ from higher-order tensors.

The key steps of Subroutine 1 are as follows. We first use the power method to decompose $\widehat{\boldsymbol{P}}_2$ to approximate the subspace spanned by $\{\bar{\boldsymbol{w}}_1^*, \bar{\boldsymbol{w}}_2^*, \cdots, \bar{\boldsymbol{w}}_K^*\}$, denoted by $\widehat{\boldsymbol{U}}$, where $\bar{\boldsymbol{w}}_i^* = \frac{\boldsymbol{w}_i^*}{\|\boldsymbol{w}_i^*\|}$. Then, we project $\widehat{\boldsymbol{M}}_3 \in \mathbb{R}^{d\times d\times d}$ to $\widehat{\boldsymbol{R}}_3 \in \mathbb{R}^{K\times K\times K}$ using $\widehat{\boldsymbol{U}}$ to reduce the computational and sample complexity (Zhong et al., 2017b). One then applies the KCL algorithm to $\widehat{\boldsymbol{R}}_3$ to obtain vectors $\widehat{\boldsymbol{v}}_i = s_i \widehat{\boldsymbol{U}}^\top \bar{\boldsymbol{w}}_i^*,\ i \in [K]$, where $s_i \in \{1, -1\}$ is a random sign. Finally, the magnitude of $\boldsymbol{w}_i^*$'s and the signs of s_i's are determined by solving a linear system of equations using the RecMagSign method. Please refer to (Zhong et al., 2017b) and (Kuleshov et al., 2015) for more details of the power method, KCL algorithm and the RecMagsign method.

Subroutine 1 Tensor Initialization Method

Input: Partition n pairs of data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ into three subsets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$
Output:
Compute $\widehat{\boldsymbol{P}}_2$ using $\mathcal{D}_1$ and an arbitrary vector $\boldsymbol{\alpha}$
$\widehat{\boldsymbol{U}} \longleftarrow$ PowerMethod$(\widehat{\boldsymbol{P}}_2, K)$
Compute $\widehat{\boldsymbol{R}}_3 = \widehat{\boldsymbol{M}}_3(\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{U}}, \widehat{\boldsymbol{U}})$ from data set $\mathcal{D}_2$
$\{\widehat{\boldsymbol{v}}_i\}_{i\in[K]} \longleftarrow$ KCL$(\widehat{\boldsymbol{R}}_3)$
$\{\boldsymbol{W}_0\} \longleftarrow$ RecMagSign$(\widehat{\boldsymbol{U}}, \{\hat{\boldsymbol{v}}_i\}_{i\in[K]}, \mathcal{D}_3)$
Return: $\boldsymbol{W}_0$

4 Main Theoretical Results

The main idea of our analysis is to show that the empirical risk function in (3) is strongly convex in a region near $\boldsymbol{W}^*$. Then $\boldsymbol{W}_0$ returned by Subroutine is in this convex region, and the iterates returned by Algorithm 1 converge to a critical point in this region. Before formally stating our result in Theorem 1, we summarize the key implications of Theorem 1 as follow.

(1). Convergence rate and estimation accuracy: The iterates $\boldsymbol{W}_t$ converge to a critical point $\widehat{\boldsymbol{W}}_n$ linearly, and the distance between $\widehat{\boldsymbol{W}}_n$ and $\boldsymbol{W}^*$ decays in the rate of $\sqrt{d \log n / n}$. For example, when the number of samples n is $\Theta(d \log^2 d)$[1], the estimation error decays in the order of $\frac{1}{\log d}$.

(2). Sample complexity: The sample complexity for accurate estimation is $\Theta(d \log^2 d)$ where d is the feature dimension. This result is in the same order as the sample complexity with the standard Gaussian input in (Fu et al., 2020) and (Zhong et al., 2017b). It is almost order-wise optimal with respect to d with an additional multiplier of $\log^2 d$.

(3). Impact of a large mean or variance: If everything else is fixed, and at least one variance σ_l or one entry of a mean $\mu_{l(i)}$ (the ith entry of $\boldsymbol{\mu}_l$) of the Gaussian mixture model increases, the sample complexity increases to infinity and the convergence slows down. The intuition is that when some mean or variance is large, some training samples have significantly large magnitude such that the sigmoid function saturates. These training samples are not informative for the estimation of $\boldsymbol{W}^*$, and the gradient of these samples is close to zero. Therefore, the required number of samples to estimate $\boldsymbol{W}^*$ needs to increase, and the gradient descent algorithm slows down.

(4). Impact of small variances: When all variances in the Gaussian mixture model approach zero, the sample complexity increases to infinity, and the convergence slows down. The intuition is that when input data are concentrated on a few vectors, the optimization problem does not have a benign landscape.

Combining points (3) and (4), one can see that the training samples shall have a medium level of variance to enhance the learning performance in terms of reducing the sample size and speeding up the convergence. If the variance is too large, some samples become non-informative and affect the learning negatively. If the variance is too small, the learning problem becomes mathematically challenging to solve.

To formally state Theorem 1, we define the following three quantities. $\rho(\boldsymbol{\mu}, \sigma)$ is motivated by the ρ parameter for the standard Gaussian distribution in (Zhong et al., 2017b), and we generalize it to a Gaussian with an arbitrary mean and variance. We define the new quantities $\Gamma(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*)$ and $D_m(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})$ for the Gaussian mixture model.

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

[1] $f(x) = O(g(x))$ means there exist $c' > 0$ such that $f(x) \leq c' g(x)$ holds when x is sufficiently large. $f(x) = \Omega(g(x))$ means there exist $c'' > 0$ such that $f(x) \geq c'' g(x)$ holds when x is sufficiently large. $f(x) = \Theta(g(x))$ means there exists $c > 0$ and $C > 0$ such that $c g(x) \leq f(x) \leq C g(x)$ holds when x is sufficiently large.

and $\boldsymbol{u}$, respectively. Define $\rho(\boldsymbol{u}, \sigma)$ as

$$\rho(\boldsymbol{u}, \sigma) = \min_{i,j\in[d], j\neq i} \{(u_j^2+1)(\beta_0(i,\boldsymbol{u},\sigma) - \alpha_0(i,\boldsymbol{u},\sigma)^2), \beta_2(i,\boldsymbol{u},\sigma) - \frac{\alpha_2(i,\boldsymbol{u},\sigma)^2}{u_i^2+1}\} \tag{11}$$

Definition 3 *(Γ-function). With (6), (11) and κ, η defined in Section 2, we define*

$$\Gamma(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*) = \sum_{l=1}^{L} \frac{\lambda_l}{\kappa^2 \eta} \frac{\sigma_l^2}{\sigma_{\max}^2} \rho\Big(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l \delta_K(\boldsymbol{W}^*)}, \sigma_l \delta_K(\boldsymbol{W}^*)\Big) \tag{12}$$

Definition 4 *(D-function). Given the Gaussian Mixture Model in (6) and any positive integer m, define $D_m(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})$ as*

$$D_m(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) = \sum_{l=1}^{L} \lambda_l \Big(\frac{||\boldsymbol{\mu}_l||_\infty}{\sigma_l} + 1\Big)^m, \tag{13}$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_L) \in \mathbb{R}^L$, $\boldsymbol{M} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_L) \in \mathbb{R}^{d\times L}$ and $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_L) \in \mathbb{R}^L$.

Next we state our main theoretical result as follows.

Theorem 1 *Consider the binary classification problem with one hidden-layer FCN as in (2). Assume that Assumption 1 holds, then there exist constants $C_0 > 0$ and $\epsilon_0 \in (0, \frac{1}{4})$ such that as long as the sample size n satisfies*[2]

$$n \geq n_{sc} := poly(\epsilon_0^{-1}, \kappa, \tau, K)(\sigma_{\max}\delta_1)^2 \Gamma(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*)^{-2} D_{12}(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) d \log^2 d, \tag{14}$$

we have that with probability at least $1 - d^{-10}$, the iterates $\{\boldsymbol{W}_t\}_{t=1}^T$ returned by Algorithm 1 converge linearly to a critical point $\widehat{\boldsymbol{W}}_n$ with the rate of convergence v, i.e.,

$$||\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n||_F \leq v^t ||\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_n||_F, \tag{15}$$

where

$$v = 1 - \frac{\Gamma(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*)\sigma_{\max}^2}{C_0 K^2 \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2} \in (0, 1). \tag{16}$$

Moreover, the distance between $\boldsymbol{W}^*$ and $\widehat{\boldsymbol{W}}_n$ is bounded by

$$||\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*||_F \leq O\Big(\frac{K^{\frac{5}{2}} \sqrt{\sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2}}{\Gamma(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*)\sigma_{\max}^2} \cdot \sqrt{\frac{d \log n}{n}}\Big) \tag{17}$$

To characterize how the parameters of the Gaussian mixture model affects the learning performance, we quantify their impact on the sample complexity n_{sc} and the convergence rate v discussed in Theorem 1 as follows.

Corollary 1 *(Impact of the Gaussian mixture model on n_{sc} and v) When everything else is fixed except for some σ_l for any $l \in [L]$, n_{sc} in (14) is lower bounded by an increasing function of σ_l, and the function approaches infinity if σ_l increases to infinity. Similarly, when everything else is fixed, n_{sc} is lower bounded by an increasing function of $\mu_{l(i)}$ with any $l \in [L]$ and $i \in [d]$, where $\mu_{l(i)}$ is the ith entry of $\boldsymbol{\mu}_l$. Moreover, v in (16) increase to 1 as σ_l (or $\mu_{l(i)}$) increases to infinity. Furthermore, n_{sc} goes to infinity, and v goes to 1 if all σ_l go to zero for all $l \in [L]$.*

Theorem 1 provides the first theoretical guarantee of learning one-hidden-layer neural networks with the input following the Gaussian mixture model. Although we consider the sigmoid activation in this paper, our results and analysis apply to any other activations function ϕ as long as ϕ' is an even function, and ϕ, ϕ' and ϕ'' are bounded. Examples of such activation include $\tanh$ and erf.

Note that if we scale the weights $\boldsymbol{W}^{*'} = \boldsymbol{W}^*/c$ and the input feature $\boldsymbol{x}' = c\boldsymbol{x}$ simultaneously, the output remain the same for any nonzero constant c. Therefore, the learning problem in these

[2] $\sigma_{\min}$ and $\sigma_{\max}$ denote the minimum and maximum among $\{\sigma_1, \cdots, \sigma_L\}$, respectively. $\tau = \frac{\sigma_{\max}}{\sigma_{\min}}$.

two cases should be equivalent in terms of the sample complexity and convergence rate. Theorem 1 reflects such equivalence. To see this, note that $\delta_i' = \delta_i/c$ where δ_i' is the ith largest singular value of $\boldsymbol{W}^{*'}$. Thus, $\sigma_{\max}\delta_1 = \sigma'_{\max}\delta'_1$, $\Gamma(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*) = \Gamma(\boldsymbol{\lambda}, \boldsymbol{M}', \boldsymbol{\sigma}', \boldsymbol{W}^{*\prime})$, and other parameters in (14) remain the same. Similarly, the convergence rate in (15) remains the same in both cases.

5 Numerical Experiments

We verify Theorem 1 through numerical experiments. The number of neurons is fixed as $K = 3$ as (Fu et al., 2020). We generate a ground-truth $\boldsymbol{W}^* \in \mathbb{R}^{d\times K}$ from standard Gaussian distribution. The training samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ are generated using (6) and (2). The maximum iterations of Algorithm 1 is set as 12000. We use random initialization rather than tensor initialization to reduce the computational time. We randomly select $\boldsymbol{W}_0$ from the set $\{\boldsymbol{W} \in \mathbb{R}^{d\times K} : \frac{\|\boldsymbol{W}-\boldsymbol{W}^*\|_F}{\|\boldsymbol{W}^*\|_F} \leq 0.1\}$.

5.1 Sample complexity

Consider the case that $L = 2$, $\lambda_1 = \lambda_2 = \frac{1}{2}$. Let $\boldsymbol{\mu}_1$ be an all one vector in $\mathbb{R}^d$ and let $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$. Let $\sigma_1 = \sigma_2 = 1$. We vary d and evaluate the sample complexity bound in (14) with respect to d. We randomly initialize M times and let $\widehat{\boldsymbol{W}}_n^{(m)}$ denote the output of Algorithm 1 in the mth trail. Let $\bar{\boldsymbol{W}}_n$ denote the mean values of all $\widehat{\boldsymbol{W}}_n^{(m)}$, and let $d_W = \sqrt{\sum_{m=1}^M ||\widehat{\boldsymbol{w}}_n^m - \bar{\boldsymbol{W}}_n||^2/M}$ denote the variance. An experiment is successful if $d_W \leq 10^{-4}$ and fails otherwise. M is set as 20.

We vary d and the number of samples n. For each pair of d and n, 50 independent sets of $\boldsymbol{W}^*$ and the corresponding training samples are generated. Fig. 1 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. One can see that the sample complexity is indeed almost linear in d, as predicted by (14).

We then fix $d = 5$ and study the impact on the sample complexity when the mean and variance in the Gaussian mixture model change. In Fig. 2.(a), we fix $\sigma_1 = \sigma_2 = 1$ and let $\boldsymbol{\mu}_1 = \mu \cdot \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$. μ varies from 0 to 7.5. Fig. 2.(a) shows that when the mean increases, the sample complexity increases. This coincides with our theoretical analyses in Section 4. In Fig. 2.(b), we fix $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$, and let $\sigma_1 = \sigma$ and $\sigma_2 = 1$. σ varies from $10^{-1.4}$ to $10^{1.4}$. The sample complexity increases both when σ increases and when σ approaches zero. The result matches our theoretical prediction in Section 4.

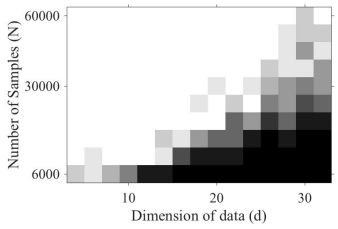

Figure 1: The sample complexity against the feature dimension d

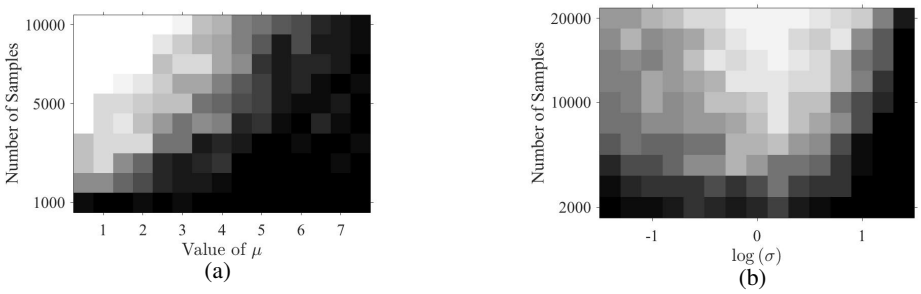

Figure 2: (a) The sample complexity when one mean changes, (b) The sample complexity when one variance changes.

5.2 CONVERGENCE ANALYSIS

We next study the convergence rate of Algorithm 1. d is fixed as 5. Fig. 3.(a) shows the impact of the mean of the Gaussian mixture model on the convergence rate. We set $\lambda_1 = \lambda_2 = 0.5$, $\boldsymbol{\mu}_1 = \mu \cdot \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$, and $\sigma_1 = \sigma_2 = 1$. The sample complexity n is set to 10000. One can see that Algorithm 1 always converges linearly when μ changes. Moreover, as μ increases, Algorithm 1 converges slower, as predicted by our theoretical analyses in Section 4. In Fig. 3.(b) shows the impact of the variance of the Gaussian mixture model. We set $\lambda_1 = \lambda_2 = 0.5$, $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$, $\sigma_1 = \sigma_2 = \sigma$. The sample complexity n is set to 50000. Among different σ we test, Algorithm 1 converges fastest when $\sigma = 1$. The convergence rate slows down when σ increases to 2 or when σ decreases to 0.5. The result is consistent with our theoretical results in Section 4.

We then verify the convergence rate v in (16), which shows that $v = 1 - \Theta(K^{-2})$. We set $\lambda_1 = \lambda_2 = 0.5$, $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$, $\sigma_1 = \sigma_2 = 1$. K ranges from 2 to 8. One can see from Fig. 4 that, as predicted, the convergence rate is almost linear in $1/K^2$.

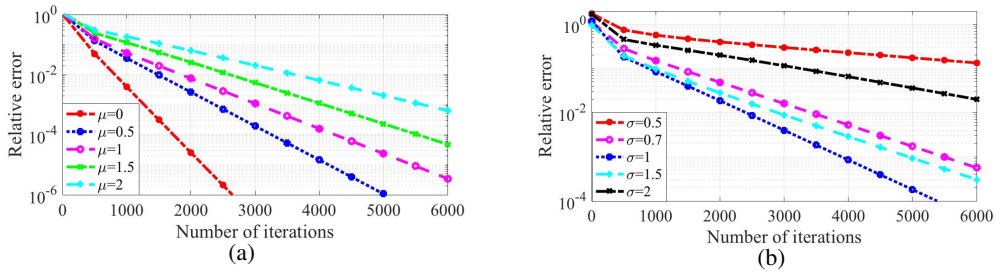

Figure 3: (a) The convergence rate with different μ, (b) The convergence rate with different σ.

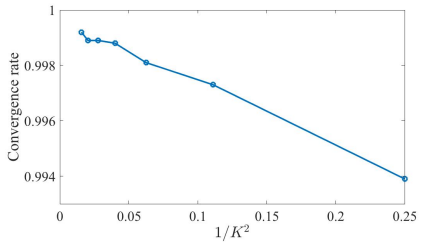

Figure 4: Convergence rate when the number of neurons K changes

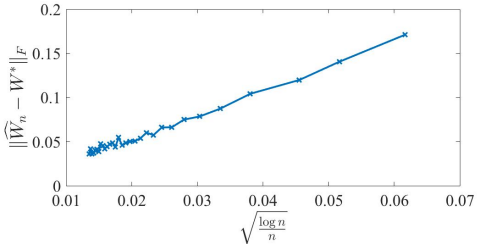

Figure 5: The relative error of the learned model with the ground-truth when n changes

We then evaluate the distance between the solution $\widehat{\boldsymbol{W}}_n$ returned by Algorithm 1 and $\boldsymbol{W}^*$, measured by $||\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*||_F$. d is set as 5. n ranges from 2×10^3 to 6×10^4. $\sigma_1 = \sigma_2 = 3$, $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$. Each point in Fig. 5 is averaged over 100 independent experiments of different $\boldsymbol{W}^*$ and the corresponding training set. $\|\boldsymbol{W}^*\|_F$ is normalized to 1. The error is indeed linear in $\sqrt{\log(n)/n}$, as predicted by (15).

6 CONCLUSIONS

This paper analyzes the theoretical performance guarantee of learning one-hidden-layer neural networks for binary classification when the input follows the Gaussian mixture model. We develop an algorithm that converges linearly to a model that has a diminishing difference with the ground-truth model that has guaranteed generalizability. We also analyze the impact of the input distribution from the perspective of sample complexity and convergence rate. This is the first theoretical characterization of the impact of input distributions on the learning performance. Future works include the analysis of multi-layer neural networks and multi-class classification.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6158–6169, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6155–6166, 2019b.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019c.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 605–614. JMLR. org, 2017.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pp. 634–644. IEEE, 1999.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685, 2019.

Simon S. Du, Jason D. Lee, and Yuandong Tian. When is a convolutional filter easy to learn? *arXiv preprint, http://arxiv.org/abs/1709.06129*, 2017.

Simon S Du, Jason D Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pp. 1338–1347, 2018.

Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

Haoyu Fu, Yuejie Chi, and Yingbin Liang. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE Transactions on Signal Processing*, 68:3225–3235, 2020.

Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BkwHObbRZ.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. IEEE, 2013.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pp. 11–20, 2013.

Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*, 2014.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.

Volodymyr Kuleshov, Arun Chaganty, and Percy Liang. Tensor factorization via matrix factorization. In *Artificial Intelligence and Statistics*, pp. 507–516, 2015.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with ReLU activation. In *Advances in Neural Information Processing Systems*, pp. 597–607. 2017.

Shiyu Liang, Ruoyu Sun, Jason D Lee, and R Srikant. Adding one neuron can eliminate all bad local minima. In *Advances in Neural Information Processing Systems*, pp. 4355–4365, 2018.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in neural information processing systems*, pp. 855–863, 2014.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *Ann. Statist.*, 46(6A):2747–2774, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Haim Permuter, Joseph Francos, and Ian Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695–706, 2006.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3), 1988.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pp. 4430–4438, 2018.

Ohad Shamir. Distribution-specific hardness of learning neural networks. *The Journal of Machine Learning Research*, 19(1):1135–1163, 2018.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3404–3413. JMLR. org, 2017.

D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley,, 1985.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. Fast learning of graph neural networks with guaranteed generalizability: One-hidden-layer case. *arXiv preprint arXiv:2006.14117*, 2020a.

Shuai Zhang, Meng Wang, Jinjun Xiong, Sijia Liu, and Pin-Yu Chen. Improved linear convergence of training cnns with generalizability guarantees: A one-hidden-layer case. *IEEE Transactions on Neural Networks and Learning Systems*, 2020b.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1524–1534. PMLR, 2019.

Kai Zhong, Zhao Song, and Inderjit S Dhillon. Learning non-overlapping convolutional neural networks with multiple kernels. *arXiv preprint arXiv:1711.03440*, 2017a.

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4140–4149, 2017b. URL https://arxiv.org/pdf/1706.03175.pdf.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 2055–2064, 2019.

A PRELIMINARIES

In this section, we introduce some properties that will be used in proving the main results.

Property 1 *$\rho(\boldsymbol{u}, \sigma)$ in Definition 2 satisfies the following properties,*

1. *$\rho(\boldsymbol{u}, \sigma) > 0$ for any $\boldsymbol{u} \in \mathbb{R}^d$ and $\sigma \neq 0$.*

2. *When everything else except $|u_i|$ is fixed, $\rho(\boldsymbol{u}, \sigma)$ is upper bounded by a function of $|u_i|$, $(u_j^2 + 1)\beta_0(i, \boldsymbol{u}, \sigma)$, which is monotonically decreasing to 0.*

3. *When everything else except σ is fixed, $\rho(\frac{\boldsymbol{u}}{\sigma}, \sigma)$ is upper bounded by a function of σ, $(\frac{u_j^2}{\sigma^2} + 1)\beta_0(i, \mathbf{0}, \sigma)$, which is monotonically decreasing to 0.*

4. *$\lim_{\sigma \to \infty} \rho(\frac{\boldsymbol{u}}{\sigma}, \sigma) = 0$.*

5. *When everything else except σ is fixed, $\rho(\frac{\boldsymbol{u}}{\sigma}, \sigma)$ is upper bounded by a positive constant, $\mathcal{C}_0(\boldsymbol{u}, \epsilon_s, \delta_s)$ for some $\epsilon_s > 0$ and $\delta_s > 0$.*

6. *When everything else except $|u_i|$ is fixed, $\rho({\boldsymbol{W}^*}^\top \boldsymbol{u}, \sigma)$ is upper bounded by a positive constant, $\mathcal{C}_1({\boldsymbol{W}^*}^\top \boldsymbol{u}|_{u_i=0}, \sigma, \epsilon_m, \delta_m)$ for some $\epsilon_m > 0$ and $\delta_m > 0$.*

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

$$\mathbb{E}_{z_i \sim \mathcal{N}(u_i,1)}[\phi'(\sigma \cdot z_i)] \leq \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i,1)}[\phi'^2(\sigma \cdot z_i)]} \tag{18}$$

$$\begin{aligned} \mathbb{E}_{z_i \sim \mathcal{N}(u_i,1)}[\phi'(\sigma \cdot z_i) z_i \cdot z_i] &\leq \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i,1)}[\phi'^2(\sigma \cdot z_i) z_i^2]} \cdot \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i,1)}[z_i^2]} \\ &= \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i,1)}[\phi'^2(\sigma \cdot z_i) z_i^2]} \cdot \sqrt{u_i^2 + 1} \end{aligned} \tag{19}$$

The equalities of the (18) and (19) hold if and only if ϕ' is a constant function. Since that ϕ is the sigmoid function, the equalities of (18) and (19) cannot hold.
By the definition of $\rho(\boldsymbol{u}, \sigma)$ in Definition 2, we have $\beta_0(i, \boldsymbol{u}, \sigma) - \alpha_0^2(i, \boldsymbol{u}, \sigma) > 0$ and $\beta_2(i, \boldsymbol{u}, \sigma) - \frac{\alpha_2^2(i, \boldsymbol{u}, \sigma)}{u_i^2 + 1} > 0$. Therefore,

$$\rho(\boldsymbol{u}, \sigma) > 0 \tag{20}$$

(2) By the definition of $\beta_0(i, \boldsymbol{\mu}, \sigma)$, we have

$$
\begin{aligned}
\frac{\partial \beta_0(i, \boldsymbol{u}, \sigma)}{\partial u_i} &= \frac{\partial}{\partial u_i} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i)(2\pi)^{-\frac{1}{2}} \exp(-\frac{(z_i - u_i)^2}{2}) dz_i \\
&= \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i)(2\pi)^{-\frac{1}{2}} \exp(-\frac{(z_i - u_i)^2}{2})(z_i - u_i) dz_i \\
&= \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot (x_i + u_i))(2\pi)^{-\frac{1}{2}} \exp(-\frac{x_i^2}{2}) x_i dx_i \quad x_i = z_i - u_i \\
&= \int_{0}^{\infty} (2\pi)^{-\frac{1}{2}} \exp(-\frac{x_i^2}{2}) x_i \cdot \phi'^2(\sigma \cdot (x_i + u_i)) dx_i \\
&\quad + \int_{-\infty}^{0} (2\pi)^{-\frac{1}{2}} \exp(-\frac{x_i^2}{2}) x_i \cdot \phi'^2(\sigma \cdot (x_i + u_i)) dx_i \\
&= \int_{0}^{\infty} (2\pi)^{-\frac{1}{2}} \exp(-\frac{x_i^2}{2}) x_i \cdot \phi'^2(\sigma \cdot (x_i + u_i)) dx_i \\
&\quad + \int_{-\infty}^{0} (2\pi)^{-\frac{1}{2}} \exp(-\frac{t_i^2}{2})(-t_i) \cdot \phi'^2(\sigma \cdot (-t_i + u_i)) d(-t_i) \;\; t_i = -x_i \\
&= \int_{0}^{\infty} (2\pi)^{-\frac{1}{2}} \exp(-\frac{x_i^2}{2}) x_i \cdot \phi'^2(\sigma \cdot (x_i + u_i)) dx_i \\
&\quad + \int_{+\infty}^{0} (2\pi)^{-\frac{1}{2}} \exp(-\frac{t_i^2}{2}) t_i \cdot \phi'^2(\sigma \cdot (-t_i + u_i)) dt_i \\
&= \int_{0}^{\infty} (2\pi)^{-\frac{1}{2}} \exp(-\frac{x_i^2}{2}) x_i \cdot \Big(\phi'^2(\sigma \cdot (x_i + u_i)) - \phi'^2(\sigma \cdot (-x_i + u_i))\Big) dx_i
\end{aligned}
\tag{21}
$$

From the definition of ϕ'^2, we can obtain that when $u_i > 0$, $\phi'^2(\sigma \cdot (x_i + u_i)) < \phi'^2(\sigma \cdot (-x_i + u_i))$ for any $x_i > 0$. When $u_i < 0$, $\phi'^2(\sigma \cdot (x_i + u_i)) > \phi'^2(\sigma \cdot (-x_i + u_i))$ for any $x_i > 0$. Therefore, we have that when $u_i > 0$, $\frac{\partial \beta_0(i, \boldsymbol{u}, \sigma)}{\partial u_i} < 0$ and when $u_i < 0$, $\frac{\partial \beta_0(i, \boldsymbol{u}, \sigma)}{\partial u_i} > 0$. We can then conclude that if $|u_i|$ increases, $(u_j^2 + 1)\beta_0(i, \boldsymbol{u}, \sigma)$ will decrease, $(j \neq i)$. By the definition of $\rho(\boldsymbol{u}, \sigma)$, we can derive $\rho(\boldsymbol{u}, \sigma) \leq (u_j^2 + 1)(\beta_0(i, \boldsymbol{u}, \sigma))$. Moreover, we have

$$
\begin{aligned}
\lim_{|u_i| \to \infty} \beta_0(i, \boldsymbol{u}, \sigma) &= \lim_{|u_i| \to \infty} \mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'^2(\sigma \cdot z_i)] \\
&= \lim_{|u_i| \to \infty} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i)(2\pi)^{-\frac{1}{2}} \exp(-\frac{(z_i - u_i)^2}{2}) dz_i \\
&= \lim_{|u_i| \to \infty} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot u_i \cdot x_i)(2\pi \frac{1}{u_i^2})^{-\frac{1}{2}} \exp(-\frac{(x_i - 1)^2}{2\frac{1}{u_i^2}}) dx_i \quad \text{Let } z_i = u_i \cdot x_i \\
&= \lim_{|u_i| \to \infty} \int_{-\infty}^{\infty} \phi'(\sigma \cdot u_i \cdot x_i) \cdot \delta(x_i - 1) dx_i \\
&= \lim_{|u_i| \to \infty} \phi'(\sigma \cdot u_i \cdot 1) \\
&= 0
\end{aligned}
\tag{22}
$$

The third to last line of (22) comes from that if the standard deviation of a Gaussian distribution goes to zero, then its probability density function will become a Dirac delta function (δ function) centered at the mean value. The second to last line of (22) follows from that $\int_{-\infty}^{\infty} f(x) \cdot \delta(x - t) dx = f(t)$. Hence, $(u_j^2 + 1)\beta_0(i, \boldsymbol{u}, \sigma)$ decrease to zero as $|u_i|$ goes to infinity.

(3) From the conclusion of monotonicity of $\beta_0(i, \boldsymbol{u}, \sigma)$ against u_i in (2), we know $\beta_0(i, \frac{\boldsymbol{u}}{\sigma}, \sigma) \leq$

$\beta_0(i, \mathbf{0}, \sigma)$. Then we have that for any $i \in [d]$

$$\begin{aligned}
&\frac{\partial(\frac{u_j^2}{\sigma^2}+1)\beta_0(i,\mathbf{0},\sigma)}{\partial u_i} = \frac{\partial}{\partial\sigma}(\frac{u_j^2}{\sigma^2}+1)\int_{-\infty}^{\infty}\phi'^2(\sigma\cdot z_i)(2\pi)^{-\frac{1}{2}}\exp(-\frac{z_i^2}{2})dz_i \\
=&\int_{-\infty}^{\infty}(\frac{u_j^2}{\sigma^2}+1)\cdot 2\phi'(\sigma\cdot z_i)\phi''(\sigma\cdot z_i)z_i(2\pi)^{-\frac{1}{2}}\exp(-\frac{z_i^2}{2})dz_i \\
&-\frac{2u_j^2}{\sigma^3}\int_{-\infty}^{\infty}\phi'^2(\sigma\cdot z_i)(2\pi)^{-\frac{1}{2}}\exp(-\frac{z_i^2}{2})dz_i
\end{aligned} \tag{23}$$

Since that $\phi''(\sigma\cdot z_i)z_i \le 0$ in the first term of the last step in (23), $\frac{\partial(\frac{u_j^2}{\sigma^2}+1)\beta_0(i,\mathbf{0},\sigma)}{\partial\sigma} \le 0$. By the fact that $\rho(\frac{\boldsymbol{u}}{\sigma},\sigma) \le (\frac{u_j^2}{\sigma^2}+1)\beta_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma) \le (\frac{u_j^2}{\sigma^2}+1)\beta_0(i,\mathbf{0},\sigma)$, we find a monotonically decreasing function of σ. From (22), we also have $\lim_{|u_i|\to\infty}(\frac{u_j^2}{\sigma^2}+1)\beta_0(i,\mathbf{0},\sigma) = 0$. Therefore, (3) can be derived.

(4) Similar to (22), we can derive

$$\begin{aligned}
&\lim_{\sigma\to\infty}\beta_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma) - \alpha_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)^2 \\
=&\lim_{\sigma\to\infty}\int_{-\infty}^{\infty}\phi'^2(\sigma\cdot z_i)(2\pi)^{-\frac{1}{2}}\exp(-\frac{(z_i-\frac{u_i}{\sigma})^2}{2})dz_i \\
&-\Big(\int_{-\infty}^{\infty}\phi'(\sigma\cdot z_i)(2\pi)^{-\frac{1}{2}}\exp(-\frac{(z_i-\frac{u_i}{\sigma})^2}{2})dz_i\Big)^2 \\
=&\lim_{\sigma\to\infty}\int_{-\infty}^{\infty}\phi'^2(\sigma\cdot\frac{u_i}{\sigma}x_i)(2\pi\frac{\sigma^2}{u_i^2})^{-\frac{1}{2}}\exp(-\frac{(x_i-1)^2}{2\frac{\sigma^2}{u_i^2}})dx_i \\
&-\Big(\int_{-\infty}^{\infty}\phi'(\sigma\cdot\frac{u_i}{\sigma}x_i)(2\pi\frac{\sigma^2}{u_i^2})^{-\frac{1}{2}}\exp(-\frac{(z_i-1)^2}{2\frac{\sigma^2}{u_i^2}})dz_i\Big)^2 \quad x_i = \frac{u_i}{\sigma}x_i \\
=&\lim_{\sigma\to\infty}\int_{-\infty}^{\infty}\phi'^2(\sigma\cdot\frac{u_i}{\sigma}x_i)\delta(x_i-1)dx_i - \Big(\int_{-\infty}^{\infty}\phi'(\sigma\cdot\frac{u_i}{\sigma}x_i)\delta(x_i-1)dx_i\Big)^2 \\
=&\phi'^2(u_i) - (\phi'(u_i))^2 = 0
\end{aligned} \tag{24}$$

$$\begin{aligned}
\lim_{\sigma\to\infty}\frac{\partial}{\partial\sigma}\Big(\beta_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)-\alpha_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)^2\Big) &= \lim_{\sigma\to\infty}\frac{\partial}{\partial\sigma}\beta_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma) - 2\alpha_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)\frac{\partial}{\partial\sigma}\alpha_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma) \\
&= \lim_{\sigma\to\infty}2\phi'(u_i)\phi''(u_i)\frac{u_i}{\sigma} - 2\phi'(u_i)\cdot\phi''(u_i)\frac{u_i}{\sigma} = 0
\end{aligned} \tag{25}$$

From (24) and (25), we have

$$\begin{aligned}
&\lim_{\sigma\to\infty}(\frac{u_j^2}{\sigma^2}+1)(\beta_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)-\alpha_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)^2) \\
=&\lim_{\sigma\to\infty}\frac{u_j^2}{\sigma^2}(\beta_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)-\alpha_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)^2) \\
=&\lim_{\sigma\to\infty}\frac{u_j^2}{2\sigma}\frac{\partial}{\partial\sigma}\Big(\beta_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)-\alpha_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)^2\Big) \\
=&\lim_{\sigma\to\infty}\frac{u_j^2}{2}\frac{\partial^2}{\partial\sigma^2}\Big(\beta_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)-\alpha_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)^2\Big) \\
=&\lim_{\sigma\to\infty}\frac{u_j^2}{2}\Big(2\phi''^2(u_i)\frac{u_i^2}{\sigma^2}+2\phi'(u_i)\phi^{(3)}(u_i)\frac{u_i^2}{\sigma^2}-2\phi''^2(u_i)\frac{u_i^2}{\sigma^2}-2\phi'(u_i)\phi'''(u_i)\frac{u_i^2}{\sigma^2}\Big) = 0
\end{aligned} \tag{26}$$

The second and the third step of (26) comes from the L'Hopital's rule. Since that $0 < \rho(\frac{\boldsymbol{u}}{\sigma},\sigma) \le (\frac{u_j^2}{\sigma^2}+1)(\beta_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)-\alpha_0(i,\frac{\boldsymbol{u}}{\sigma},\sigma)^2)$, (4) can be derived.

(5) From (4), we have there exists an $\epsilon_s > 0$ and its corresponding $\delta_s > 0$ such that when $0 < \sigma < \delta_s$, $0 < \rho(\frac{\boldsymbol{u}}{\sigma},\sigma) < \epsilon_s$. From (3), we have when $\sigma \in [\delta_s, +\infty)$, $\rho(\frac{\boldsymbol{u}}{\sigma},\sigma) \le (\frac{u_j^2}{\delta_s^2}+1)\beta_0(i,\mathbf{0},\delta_s)$.

Therefore, when $\sigma > 0$, $\rho(\frac{\boldsymbol{u}}{\sigma}, \sigma) < \mathcal{C}_0(\boldsymbol{u}, \epsilon_s, \delta_s) := \max\{\epsilon_s, (\frac{u_j^2}{\delta_s^2} + 1)\beta_0(i, \mathbf{0}, \delta_s)\}$ which is a positive constant independent of σ.
(6) Since that $\rho(\boldsymbol{W}^{*\top}\boldsymbol{u}, \sigma)$ is a continuous function of u_i, we have

$$\lim_{|u_i| \to 0} \rho(\boldsymbol{W}^{*\top}\boldsymbol{u}, \sigma) = \rho(\boldsymbol{W}^{*\top}\boldsymbol{u}, \sigma)|_{u_i=0} \tag{27}$$

From (2), we can obtain

$$\lim_{|u_i| \to \infty} \rho(\boldsymbol{W}^{*\top}\boldsymbol{u}, \sigma) = 0 \tag{28}$$

Combining (27) and (28), and following the proof in (5), we can derive $\rho(\boldsymbol{W}^{*\top}\boldsymbol{u}, \sigma)$ is upper bounded by a positive function $\mathcal{C}_1(\boldsymbol{W}^{*\top}\boldsymbol{u}|_{u_i=0}, \sigma, \epsilon_m, \delta_m)$ which is independent of u_i

Property 2 *With the notation in (6), if a function $f(\boldsymbol{x})$ is an even function, then*

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d)}[f(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d)}[f(\boldsymbol{x})] \tag{29}$$

Proof:
Denote

$$g(\boldsymbol{x}) = f(\boldsymbol{x})(2\pi\sigma^2)^{-\frac{d}{2}} \exp(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}||^2}{2\sigma^2}) \tag{30}$$

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})}[f(\boldsymbol{x})] &= \int_{\boldsymbol{x} \in \mathbb{R}^d} g(\boldsymbol{x}) d\boldsymbol{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \cdots, x_d) dx_1 \cdots dx_d \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{\infty}^{-\infty} g(x_1, x_2, \cdots, x_d) d(-x_1) dx_2 \cdots dx_d \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(-x_1, x_2 \cdots, x_d) dx_1 dx_2 \cdots dx_d \\
&= \int_{\boldsymbol{x} \in \mathbb{R}^d} g(-\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})(2\pi\sigma^2)^{-\frac{d}{2}} \exp(-\frac{||\boldsymbol{x} + \boldsymbol{\mu}||^2}{2\sigma^2}) d\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(-\boldsymbol{\mu}, \sigma^2 \cdot \boldsymbol{I}_d)}[f(\boldsymbol{x})]
\end{aligned} \tag{31}$$

Therefore, we have

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d)}[f(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d)}[f(\boldsymbol{x})] \tag{32}$$

Property 3 *Under Gaussian Mixture Model $\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)$, we have the following upper bound.*

$$\mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)}[(\boldsymbol{u}^\top \boldsymbol{x})^{2t}] \leq (2t-1)!!||\boldsymbol{u}||^{2t} \sum_{l=1}^{L} \lambda_l (||\boldsymbol{\mu}_l||_\infty + \sigma_l)^{2t} \tag{33}$$

Proof:
The main idea is to find an upper bound with symmetric distribution assumption first, and then apply Property 2 to extend the conclusion to the general case.
(a) If the Mixed-Gaussian distribution is symmetric and $L = 2$, i.e. $\boldsymbol{x} \sim \frac{1}{2}\Big(\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d) + \mathcal{N}(-\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_d)\Big)$, then we first need to analyse the distribution of $\boldsymbol{u}^\top \boldsymbol{x}$ by computing the moment

generating function

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\left(\mathcal{N}(\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)+\mathcal{N}(-\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)\right)}[\exp(t\boldsymbol{u}^\top\boldsymbol{x})] = \mathbb{E}[\exp(t\sum_{i=1}^{d}u_ix_i)] = \prod_{i=1}^{d}\mathbb{E}[\exp(tu_ix_i)] \\
=&\prod_{i=1}^{d}\{\sum_{j=1}^{2}\frac{1}{2}\int_{-\infty}^{\infty}\exp(tu_ix_i)\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x_i-(-1)^j\mu_i)^2}{2\sigma^2})dx_i\} \\
=&\prod_{i=1}^{d}\{\sum_{j=1}^{2}\frac{1}{2}\exp(tu_i(-1)^j\mu_i) \\
&\cdot\int_{-\infty}^{\infty}\exp\left(tu_i(x_i-(-1)^j\mu_i)\right)\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x_i-(-1)^j\mu_i)^2}{2\sigma^2})dx_i\} \\
=&\prod_{i=1}^{d}\{\frac{1}{2}\exp(-tu_i\mu_i)+\frac{1}{2}\sigma^2u_i^2t^2+\frac{1}{2}\exp(tu_i\mu_i+\frac{1}{2}\sigma^2u_i^2t^2)\} \\
:=&\sum_{i=1}^{2^d}\frac{1}{2^d}\exp(t\mu_i'+\frac{1}{2}t^2\sigma'^2)
\end{aligned}
\tag{34}
$$

which is the Moment Generating Function of $\sum_{i=1}^{2^d}\frac{1}{2^d}\mathcal{N}(\mu_i',\sigma'^2)$. The last step of (34) is by expanding the multiplication of d terms. Specifically, let $\{\boldsymbol{s}^i\}_{i=1}^{2^d}$ denote all 2^d vectors in $\mathbb{R}^d$ taking values from 0 and 1. Let s_k^i ($k\in[d]$) denote the k-th entry of $\boldsymbol{s}^i$. We define $\mu_i'=\sum_{k=1}^{d}(-1)^{s_k^i}u_k\mu_k\in\mathbb{R}$ for $i\in[2^d]$, and $\sigma'=\sigma||\boldsymbol{u}||\in\mathbb{R}$, where u_k and μ_k are the k-th entry of the vector $\boldsymbol{u}$ and $\boldsymbol{\mu}$, respectively. Then we can derive the first few steps of $\mathbb{E}[(\boldsymbol{u}^\top\boldsymbol{x})^{2t}]$

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\left(\mathcal{N}(\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)+\mathcal{N}(-\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)\right)}[(\boldsymbol{u}^\top\boldsymbol{x})^{2t}] \\
=&\int_{-\infty}^{\infty}y^{2t}\sum_{i=1}^{2^d}\frac{1}{2^d}\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{(y-\mu_i')^2}{2\sigma'^2}}dy \\
=&\sum_{i=1}^{2^d}\frac{1}{2^d}\int_{-\infty}^{\infty}(y-\mu_i'+\mu_i')^{2t}\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{(y-\mu_i')^2}{2\sigma'^2}}dy \\
=&\sum_{i=1}^{2^d}\frac{1}{2^d}\int_{-\infty}^{\infty}\sum_{p=0}^{2t}\binom{2t}{p}\mu_i'^{2t-p}(y-\mu')^p\frac{1}{\sqrt{2\pi}\sigma'}e^{-\frac{(y-\mu_i')^2}{2\sigma'^2}}dy \\
=&\sum_{i=1}^{2^d}\frac{1}{2^d}\sum_{p=0}^{2t}\binom{2t}{p}\mu_i'^{2t-p}\cdot\begin{cases}0, & p\text{ is odd}\\ (p-1)!!\sigma'^2, & p\text{ is even}\end{cases} \\
=&\sum_{i=1}^{2^d}\frac{1}{2^d}\sum_{k=0}^{t}\binom{2t}{2k}\mu_i'^{2t-2k}\sigma'^{2k}(2k-1)!! \\
=&\frac{1}{2^d}\sum_{k=0}^{t}\binom{2t}{2k}\sigma'^{2k}(2k-1)!!\sum_{i=1}^{2^d}\mu_i'^{2t-2k}
\end{aligned}
\tag{35}
$$

The first step is by the distribution of $\boldsymbol{u}^\top\boldsymbol{x}$ we obtain from (34). The third step follows from the binomial expansion. The forth step results from the calculation of high-order moment of Gaussian distribution. The second to last step is derived from the inverse of binomial expansion. The last step is due to the substitution of summation. To compute the inner summation in the last step of (35), we

have

$$
\begin{aligned}
&\sum_{i=1}^{2^d} {\mu_i'}^{2t} \\
=&\sum_{i=1}^{2^d} (u_1(-1)^{s_1^i}\mu_1 + u_2(-1)^{s_2^i}\mu_2 + ... + u_d(-1)^{s_d^i}\mu_d)^{2t} \\
=&\sum_{i=1}^{2^d} \sum_{p_1^{(i)}+\cdots+p_d^{(i)}=2t} \frac{(2t)!}{p_1^{(i)}!p_2^{(i)}!...p_d^{(i)}!}(u_1(-1)^{s_1^i}\mu_1)^{p_1^{(i)}}...(u_d(-1)^{s_d^i}\mu_d)^{p_d^{(i)}} \\
=&\sum_{i=1}^{2^d} \sum_{p_1^{(i)}+\cdots+p_d^{(i)}=2t} \frac{(2t)!}{p_1^{(i)}!p_2^{(i)}!...p_d^{(i)}!}(u_1\mu_1)^{p_1^{(i)}}...(u_d\mu_d)^{p_d^{(i)}} \qquad \text{all the } p_i \text{ are even} \\
=&\sum_{i=1}^{2^d} \sum_{p_1^{(i)}+\cdots+p_d^{(i)}=2t} \frac{(2t)!}{p_1^{(i)}!p_2^{(i)}!...p_d^{(i)}!}(u_1^2\mu_1^2)^{q_1^{(i)}}...(u_d^2\mu_d^2)^{q_d^{(i)}} \qquad q_j^{(i)} = \frac{p_j^{(i)}}{2} \\
\leq&\sum_{i=1}^{2^d} \max \frac{(2t)!}{p_1^{(i)}!p_2^{(i)}!...p_d^{(i)}!}\Big/\frac{(t)!}{q_1^{(i)}!q_2^{(i)}!...q_d^{(i)}!}\} \sum_{\sum_{h=1}^d q_h^{(i)}=t} \frac{(t)!}{q_1^{(i)}!q_2^{(i)}!...q_d^{(i)}!}(u_1^2\mu_1^2)^{q_1^{(i)}}...(u_d^2\mu_d^2)^{q_d^{(i)}} \\
\leq&\sum_{i=1}^{2^d} \max\{\frac{(2t)!}{p_1^{(i)}!p_2^{(i)}!...p_d^{(i)}!}\Big/\frac{(t)!}{q_1^{(i)}!q_2^{(i)}!...q_d^{(i)}!}\} \cdot (u_1^2\mu_1^2 + \cdots + u_d^2\mu_d^2)^t \\
\leq&\sum_{i=1}^{2^d} \max\{\frac{(2t)!}{p_1^{(i)}!p_2^{(i)}!...p_d^{(i)}!}\Big/\frac{(t)!}{q_1^{(i)}!q_2^{(i)}!...q_d^{(i)}!}\} \cdot (u_1^2 + ... + u_d^2)^t \cdot \max_j\{|\mu_j|\}^{2t} \\
\leq&2^d||\boldsymbol{u}||^{2t} \cdot \max_j\{|\mu_j|\}^{2t} \cdot (2t-1)!!
\end{aligned} \tag{36}
$$

Firstly we explain the third step. For any odd p_ℓ, there is a term $a_0 = (u_\ell(-1)^{s_\ell^i}\mu_\ell)^{p_\ell} \cdot \prod_{k\neq\ell}(u_k(-1)^{s_k^i}\mu_k)^{p_k}$ among the expansion of ${\mu_i'}^{2t}$, whose corresponding vector $\boldsymbol{s}^i$ is $(s_i^i, \cdots, s_\ell^i, \cdots, s_d^i)$. We can find a μ_j' such that its corresponding vector is $(s_1^i, \cdots, 1-s_\ell^i, \cdots, s_d^i)$, which is only different from the tuple of μ_i' in the ℓ-th entry. Therefore, in the expansion of ${\mu_j'}^{2t}$, there exists a term $a_0' = (u_\ell(-1)^{1-s_\ell^i}\mu_\ell)^{p_\ell} \cdot \prod_{k\neq\ell}(u_k(-1)^{s_k^i}\mu_k)^{p_k}$ that can be cancelled out by a_0. Therefore, there will be no odd power terms left. The third to last step of (36) is by the inverse binomial expansion. The second to last step is by the inequality $\sum_{i=1}^N a_i b_i \leq \max\{b_i\} \cdot \sum_{i=1}^N a_i$, where a_i and b_i are positive. The last step is because

$$
\begin{aligned}
&\frac{(2t)!}{p_1^{(i)}!p_2^{(i)}!...p_d^{(i)}!}\Big/\frac{(t)!}{q_1^{(i)}!q_2^{(i)}!...q_d^{(i)}!} \\
=&\frac{(2t)!}{t!} \cdot \frac{\frac{p_{d_1}^{(i)}}{2}\frac{p_{d_2}^{(i)}}{2}!\cdots\frac{p_{d_m}^{(i)}}{2}!}{p_{d_1}^{(i)}!p_{d_2}^{(i)}!\cdots p_{d_m}^{(i)}!} \\
\leq&\frac{(2t)!}{t!} \cdot (\frac{1}{2})^m \\
\leq&\frac{(2t)!}{t!} \cdot (\frac{1}{2})^t = (2t-1)!!
\end{aligned} \tag{37}
$$

In the first equality of (37), $p_{d_1}, ..., p_{d_m}$ denote all the positive p_i. Thus, we have $\sum_{i=1}^m p_{d_i} = 2t$ where $p_{d_i} \geq 2$. Therefore, $m \leq \frac{2t}{2} = t$ which is used in the second inequality. Therefore, combining

(36), we can continue the derivation of (35) as follows.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\left(\mathcal{N}(\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)+\mathcal{N}(-\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)\right)}[(\boldsymbol{u}^\top\boldsymbol{x})^{2t}]\\
=&\frac{1}{2^d}\sum_{k=0}^{t}\binom{2t}{2k}\sigma_i'^{2k}(2k-1)!!\sum_{i=1}^{2^d}\mu_i'^{2t-2k}\\
\leq&\sum_{k=0}^{t}\binom{2t}{2k}\cdot\sigma'^{2k}(2k-1)!!||\boldsymbol{u}||^{2t}\cdot\max_j\{|\mu_j|\}^{2t-2k}(2t-1-2k)!!\\
\leq&(2t-1)!!||\boldsymbol{u}||^{2t}(||\boldsymbol{\mu}||_\infty+\sigma')^{2t}
\end{aligned}\tag{38}$$

The last step is because that

$$(2t-1-2k)!!(2k-1)!!\leq(2t-1-2k)!!\underbrace{(2t-1)(2t-3)\cdots(2t-2k+1)}_{k\text{ terms}}=(2t-1)!!$$

(b) From Property 2, since that $(\boldsymbol{u}^\top\boldsymbol{x})^{2t}$ is an even function, we have a result for a general Gaussian distribution

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)}[(\boldsymbol{u}^\top\boldsymbol{x})^{2t}]=\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\sigma^2\boldsymbol{I}_d)}[(\boldsymbol{u}^\top\boldsymbol{x})^{2t}]\\
\leq&(2t-1)!!||\boldsymbol{u}||^{2t}(||\boldsymbol{\mu}||_\infty+\sigma)^{2t}
\end{aligned}\tag{39}$$

Therefore, if there are L components in the Gaussian Mixture Model, then

$$\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[(\boldsymbol{u}^\top\boldsymbol{x})^{2t}]\leq(2t-1)!!||\boldsymbol{u}||^{2t}\sum_{l=1}^{L}\lambda_l(||\boldsymbol{\mu}_l||_\infty+\sigma_l)^{2t}\tag{40}$$

Property 4 *With Gaussian Mixture Model (7), we have*

$$\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[||\boldsymbol{x}||^{2t}]\leq d^t(2t-1)!!\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^{2t}\tag{41}$$

Proof:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[||\boldsymbol{x}||_2^{2t}]\\
=&\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[(\sum_{i=1}^{d}x_i^2)^t]\\
=&\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[d^t(\sum_{i=1}^{d}\frac{x_i^2}{d})^t]\\
\leq&\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[d^t\sum_{i=1}^{d}\frac{x_i^{2t}}{d}]\\
=&d^{t-1}\sum_{i=1}^{d}\sum_{j=1}^{L}\int_{-\infty}^{\infty}(x_i-\mu_{ji}+\mu_{ji})^{2t}\lambda_j\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x_i-\mu_{ji})^2}{2\sigma_j^2})dx_i\\
=&d^{t-1}\sum_{i=1}^{d}\sum_{j=1}^{L}\sum_{k=1}^{2t}\binom{2t}{k}\lambda_j|\mu_{ji}|^{2t-k}\cdot\begin{cases}0, & k\text{ is odd}\\(k-1)!!\sigma_j^k, & k\text{ is even}\end{cases}\\
\leq&d^{t-1}\sum_{i=1}^{d}\sum_{j=1}^{L}\sum_{k=1}^{2t}\binom{2t}{k}\lambda_j|\mu_{ji}|^{2t-k}\sigma_j^k\cdot(2t-1)!!\\
=&d^{t-1}\sum_{i=1}^{d}\sum_{j=1}^{L}\lambda_j(|\mu_{ji}|+\sigma_j)^{2t}(2t-1)!!\\
\leq&d^t(2t-1)!!\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}\|_\infty+\sigma_l)^{2t}
\end{aligned}\tag{42}$$

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 5 *The population risk function $f(\boldsymbol{W})$ is defined as*

$$\begin{aligned} f(\boldsymbol{W}) &= \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[f_n(\boldsymbol{W})] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{W};\boldsymbol{x}_i,y_i)\Big] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\ell(\boldsymbol{W};\boldsymbol{x}_i,y_i)] \end{aligned} \tag{43}$$

Based on (2), (3) and (4), we can derive its gradient and Hessian as follows.

$$\frac{\partial\ell(\boldsymbol{W};\boldsymbol{x},y)}{\partial\boldsymbol{w}_j} = -\frac{1}{K}\frac{y-H(\boldsymbol{W})}{H(\boldsymbol{W})(1-H(\boldsymbol{W}))}\phi'(\boldsymbol{w}_j^\top\boldsymbol{x})\boldsymbol{x} = \zeta(\boldsymbol{W})\cdot\boldsymbol{x} \tag{44}$$

$$\frac{\partial^2\ell(\boldsymbol{W};\boldsymbol{x},y)}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_l} = \xi_{j,l}\cdot\boldsymbol{x}\boldsymbol{x}^\top \tag{45}$$

$$\xi_{j,l}(\boldsymbol{W}) = \begin{cases} \frac{1}{K^2}\phi'(\boldsymbol{w}_j^\top\boldsymbol{x})\phi'(\boldsymbol{w}_l^\top x)\frac{H(\boldsymbol{W})^2+y-2y\cdot H(\boldsymbol{W})}{H^2(\boldsymbol{W})(1-H(\boldsymbol{W}))^2}, & j\neq l \\ \frac{1}{K^2}\phi'(\boldsymbol{w}_j^\top\boldsymbol{x})\phi'(\boldsymbol{w}_l^\top\boldsymbol{x})\frac{H(\boldsymbol{W})^2+y-2y\cdot H(\boldsymbol{W})}{H^2(\boldsymbol{W})(1-H(\boldsymbol{W}))^2} - \frac{1}{K}\phi''(\boldsymbol{w}_j^\top\boldsymbol{x})\frac{y-H(\boldsymbol{W})}{H(\boldsymbol{W})(1-H(\boldsymbol{W}))}, & j=l \end{cases} \tag{46}$$

Property 6 *With $D_m(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})$ defined in (4), we have*

$$(i)\ D_m(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})D_{2m}(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma}) \leq D_{3m}(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma}) \tag{47}$$

$$(ii)\ \big(D_m(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})\big)^2 \leq D_{2m}(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma}) \tag{48}$$

Proof:
To prove (47), 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 $\sum_{i=1}^{L}\lambda_i = 1$.

$$\begin{aligned} \sum_{i=1}^{L}\lambda_i a_i^3 - \sum_{i=1}^{L}\lambda_i a_i\sum_{i=1}^{L}\lambda_i a_i^2 &= \sum_{i=1}^{L}\lambda_i a_i\cdot\big(a_i^2 - \sum_{j=1}^{L}\lambda_j a_j^2\big) \\ &= \sum_{i=1}^{L}\lambda_i a_i\cdot\big((1-\lambda_i)a_i^2 - \sum_{1\leq j\leq L, j\neq i}\lambda_j a_j^2\big) \\ &= \sum_{i=1}^{L}\lambda_i a_i\cdot\big(\sum_{1\leq j\leq L, j\neq i}\lambda_j a_i^2 - \sum_{1\leq j\leq L, j\neq i}\lambda_j a_j^2\big) \\ &= \sum_{i=1}^{L}\lambda_i a_i\cdot\big(\sum_{1\leq j\leq L, j\neq i}\lambda_j(a_i^2 - a_j^2)\big) \\ &= \sum_{1\leq i,j\leq L, i\neq j}\big(\lambda_i\lambda_j a_i(a_i^2-a_j^2) + \lambda_i\lambda_j a_j(a_j^2-a_i^2)\big) \\ &= \sum_{1\leq i,j\leq L, i\neq j}\lambda_i\lambda_j(a_i-a_j)^2(a_i+a_j) \geq 0 \end{aligned} \tag{49}$$

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 (49), we can obtain

$$\sum_{i=1}^{L}\lambda_i a_i\sum_{i=1}^{L}\lambda_i a_i^2 \leq \sum_{i=1}^{L}\lambda_i a_i^3 \tag{50}$$

Combining (50) and the definition of $D_m(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})$ in (4), we can derive (47).
Similarly, to prove (48), 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 $\sum_{i=1}^L \lambda_i = 1$.

$$\begin{aligned}
\sum_{i=1}^L \lambda_i a_i^2 - (\sum_{i=1}^L \lambda_i a_i)^2 &= \sum_{i=1}^L \lambda_i a_i \cdot \big(a_i - \sum_{j=1}^L \lambda_j a_j\big) \\
&= \sum_{i=1}^L \lambda_i a_i \cdot \big((1-\lambda_i) a_i - \sum_{1\leq j\leq L, j\neq i} \lambda_j a_j\big) \\
&= \sum_{i=1}^L \lambda_i a_i \cdot \big(\sum_{1\leq j\leq L, j\neq i} \lambda_j a_i - \sum_{1\leq j\leq L, j\neq i} \lambda_j a_j\big) \\
&= \sum_{i=1}^L \lambda_i a_i \cdot \big(\sum_{1\leq j\leq L, j\neq i} \lambda_j (a_i - a_j)\big) \\
&= \sum_{1\leq i,j\leq L, i\neq j} \big(\lambda_i \lambda_j a_i (a_i - a_j) + \lambda_i \lambda_j a_j (a_j - a_i)\big) \\
&= \sum_{1\leq i,j\leq L, i\neq j} \lambda_i \lambda_j (a_i - a_j)^2 \geq 0
\end{aligned} \tag{51}$$

The derivation of (51) is close to (49). By (51) we have

$$(\sum_{i=1}^L \lambda_i a_i)^2 \leq \sum_{i=1}^L \lambda_i a_i^2 \tag{52}$$

Combining (52) and the definition of $D_m(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})$ in (4), we can derive (48).

B Proof of Theorem 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 $\boldsymbol{W}^*$. Lemma 2 shows that if initialized in the convex region, that the gradient descent algorithm converges linearly to a critical point $\widehat{\boldsymbol{W}}_n$, that is close to $\boldsymbol{W}^*$. Lemma 3 shows that the Tensor Initialization Method in Subroutine 1 initializes $\boldsymbol{W}_0 \in \mathbb{R}^{d\times K}$ in the local convex region. Theorem 1 follows naturally by combining these three lemmas.

This proving approach is built upon those in (Fu et al., 2020). One of our major technical contribution is extending Lemmas 1 and 2 to the Gaussian mixture model, while the results in (Fu et al., 2020) 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 $\mathbb{B}(\boldsymbol{W}^*, r)$ is used to denote the neighborhood of $\boldsymbol{W}^*$, where r is the radius of the ball.

$$\mathbb{B}(\boldsymbol{W}^*, r) = \{\boldsymbol{W} \in \mathbb{R}^{d\times K} : ||\boldsymbol{W} - \boldsymbol{W}^*||_F \leq r\} \tag{53}$$

The radius of the convex region is

$$r := \frac{C_3 \epsilon_0 \cdot \sum_{l=1}^L \lambda_l \frac{\sigma_l^2}{\eta\kappa^2} \rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l \delta_K(\boldsymbol{W}^*)}, \sigma_l \delta_K(\boldsymbol{W}^*))}{K^{\frac{7}{2}} \Big(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^8\Big)^{\frac{1}{4}}} \tag{54}$$

with some constant $C_3 > 0$.

Lemma 1 *(Strongly local convexity) Consider the classification model with FCN (2) and the sigmoid activation function. There exists a constant C such that as long as the sample size*

$$n \geq C_1 \epsilon_0^{-2} \cdot \big(\sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}\|_\infty + \sigma_l)^2\big)^2 \big(\sum_{l=1}^{L} \lambda_l \frac{\sigma_l^2}{\eta \kappa^2} \rho(\frac{\boldsymbol{W}^{*\top} \boldsymbol{\mu}_l}{\sigma_l \delta_K(\boldsymbol{W}^*)}, \sigma_l \delta_K(\boldsymbol{W}^*))\big)^{-2} dK^5 \log^2 d \tag{55}$$

for some constant $C_1 > 0$ and $\epsilon_0 \in (0, \frac{1}{4})$, we have for all $\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^*, r_{FCN})$,

$$\begin{aligned} &\Omega\Big(\frac{1-2\epsilon_0}{K^2} \sum_{l=1}^{L} \lambda_l \frac{\sigma_l^2}{\eta \kappa^2} \rho(\frac{\boldsymbol{W}^{*\top} \boldsymbol{\mu}_l}{\sigma_l \delta_K(\boldsymbol{W}^*)}, \sigma_l \delta_K(\boldsymbol{W}^*))\Big) \cdot \boldsymbol{I}_{dK} \\ &\preceq \nabla^2 f_n(\boldsymbol{W}) \preceq C_2 \sum_{l=1}^{L} \lambda_l (||\boldsymbol{\mu}_l||_\infty + \sigma_l)^2 \cdot \boldsymbol{I}_{dK} \end{aligned} \tag{56}$$

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

Lemma 2 *(Linear convergence of gradient descent) Assume the conditions in Lemma 1 hold. If the local convexity holds, there exists a critical point in $\mathbb{B}(\boldsymbol{W}^*, r)$ for some constant $C_3 > 0$ and $\epsilon_0 \in (0, \frac{1}{2})$, such that*

$$||\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*||_F \leq O\Big(\frac{K^{\frac{5}{2}} \sqrt{\sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}\|_\infty + \sigma_l)^2}}{\sum_{l=1}^{L} \lambda_l \frac{\sigma_l^2}{\eta \kappa^2} \rho(\frac{\boldsymbol{W}^{*\top} \boldsymbol{\mu}_l}{\sigma_l \delta_K(\boldsymbol{W}^*)}, \sigma_l \delta_K(\boldsymbol{W}^*))} \sqrt{\frac{d \log n}{n}}\Big) \tag{57}$$

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

$$||\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n||_F \leq v^t ||\boldsymbol{W}_0 - \widehat{\boldsymbol{W}}_n||_F \tag{58}$$

for some constant $C_4 > 0$.

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

Lemma 3 *(Tensor initialization) For classification model, with $D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})$ defined in Definition 4, we have that if the sample size*

$$n \geq \kappa^8 K^4 \tau^{12} D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) \cdot d \log^2 d, \tag{59}$$

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

$$||\boldsymbol{W}_0 - \boldsymbol{W}^*|| \lesssim \kappa^6 K^3 \cdot \tau^6 \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})} \sqrt{\frac{d \log n}{n}} ||\boldsymbol{W}^*|| \tag{60}$$

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

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

$$\begin{aligned} ||\boldsymbol{W}_0 - \boldsymbol{W}^*||_F &\leq \sqrt{K} ||\boldsymbol{W}_0 - \boldsymbol{W}^*|| \leq \kappa^6 K^{\frac{7}{2}} \cdot \tau^6 \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})} \sqrt{\frac{d \log n}{n}} ||\boldsymbol{W}^*|| \\ &\leq \frac{C_3 \epsilon_0 \Gamma(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*) \sigma_{\max}^2}{K^{\frac{7}{2}} \Big(\sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^4 \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^8\Big)^{\frac{1}{4}}} \end{aligned} \tag{61}$$

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

$$\begin{aligned} n \geq & C_0 \epsilon_0^{-2} \cdot \tau^{12} \kappa^{12} K^{14} \Big(\sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^4 \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^8\Big)^{\frac{1}{2}} (\delta_1(\boldsymbol{W}^*))^2 D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) \\ & \cdot \Gamma(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*)^{-2} \sigma_{\max}^{-4} \cdot d \log^2 d \end{aligned} \tag{62}$$

where $C_0 = \max\{C_4, C_3^{-2}\}$. By the definition 4, we can obtain

$$\Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^8\Big)^{\frac{1}{2}} \le \sqrt{D_4(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})D_8(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})}\sigma_{\max}^6 \tag{63}$$

From Property 6, we have that

$$\begin{aligned}&\sqrt{D_4(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})D_8(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})}D_6(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})\\ \le&\sqrt{D_{12}(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})}\sqrt{D_{12}(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})} = D_{12}(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})\end{aligned} \tag{64}$$

Plugging (63), (64) into (62), we have

$$n \ge n_{\text{sc}} = C_0\epsilon_0^{-2}\cdot\kappa^{12}K^{14}(\sigma_{\max}\delta_1(\boldsymbol{W}^*))^2\tau^{12}\Gamma(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma},\boldsymbol{W}^*)^{-2}D_{12}(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})\cdot d\log^2 d \tag{65}$$

Considering the requirements on the sample complexity in (55), (59) and (65), (65) shows a sufficient number of samples. Taking the union bound of all the failure probabilities in Lemma 1, and 3, Theorem 1 holds with probability $1-d^{-10}$.

B.1 Proof of Corollary 1

(1) When some σ_l varies, let $l = L$ without loss of generality, then we have

$$\begin{aligned}&\sum_{l=1}^{L-1}\frac{\lambda_l}{\kappa^2\eta}\frac{\sigma_l^2}{\sigma_{\max}^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))\\ \le&\Gamma(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma},\boldsymbol{W}^*)\\ \le&\sum_{l=1}^{L-1}\frac{\lambda_l}{\kappa^2\eta}\frac{\sigma_l^2}{\sigma_{\max}^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))+\frac{\lambda_L}{\kappa^2\eta}\mathcal{C}(\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l,\epsilon_s',\delta_s')\\ :=&\frac{1}{\sigma_{\max}^2}\cdot q_0'+\frac{\lambda_L}{\kappa^2\eta}\mathcal{C}(\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l,\epsilon_s',\delta_s')\\ \le&O(1):=q_0\end{aligned} \tag{66}$$

for some $\epsilon_s' > 0$, $\delta_s' > 0$, $q_0 > 0$ and $q_0' > 0$ when σ_L is sufficiently large. The first step of (66) follows from Property 1.1 and the second step follows from Property 1.5. Then we can derive from (65) that

$$n_{\text{sc}} \ge poly(\epsilon_0^{-1},\kappa,K,q_0^{-1})d\log^2 d\cdot(\sigma_{\max}\delta_1(\boldsymbol{W}^*))^2\tau^{12}D_{12}(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma},\boldsymbol{W}^*) \tag{67}$$

The RHS of (67) is an increasing function of σ_L when σ_l is sufficiently large. Moreover, when σ_l goes to infinity, n_{sc} will increase to infinity. Meanwhile, we can derive from (16) that

$$\begin{aligned}v \ge& 1-\frac{q_0'+\frac{\lambda_L}{\kappa^2\eta}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_L\delta_K(\boldsymbol{W}^*)},\sigma_L\delta_K(\boldsymbol{W}^*))\sigma_{\max}^2}{C_0K^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2}\\ \ge& 1-\frac{q_0'}{C_0K^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2}-\frac{1}{\kappa^2\eta}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_L\delta_K(\boldsymbol{W}^*)},\sigma_L\delta_K(\boldsymbol{W}^*))\cdot O(\frac{1}{K^2})\end{aligned} \tag{68}$$

The second step of (68) is because that $\frac{\sigma_{\max}^2}{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2} \le \frac{1}{\lambda_L}$ when σ_L is large enough. Hence, by Property 1.3, the lower bound of the convergence rate v in (68) is also an increasing function of σ_l. When σ_l increases to infinity, v will converge to 1.

(2) When some $\mu_{L(i)}$ varies for some $i \in [d]$, let $l = L$ without loss of generality, then we have

$$\begin{aligned}&\sum_{l=1}^{L-1}\frac{\lambda_l}{\kappa^2\eta}\frac{\sigma_l^2}{\sigma_{\max}^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))\\ \le&\Gamma(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma},\boldsymbol{W}^*)\\ \le&\sum_{l=1}^{L-1}\frac{\lambda_l}{\kappa^2\eta}\frac{\sigma_l^2}{\sigma_{\max}^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))+\frac{\lambda_L\sigma_L^2}{\kappa^2\eta\sigma_{\max}^2}\mathcal{C}_1(\boldsymbol{W}^{*\top}\boldsymbol{\mu}_L|_{\mu_{L(i)}=0},\sigma_L,\epsilon_m',\delta_m'):=q_1\end{aligned} \tag{69}$$

for some $\epsilon'_m > 0$ and $\delta'_m > 0$. The first step of (69) follows from Property 1.1 and the second step follows from Property 1.6. Then we can derive from (65) that

$$n_{\text{sc}} \geq poly(\epsilon_0^{-1}, \kappa, , \tau, K, q_1^{-1}, \sigma_{\max}\delta_1(\boldsymbol{W}^*))d\log^2 d \cdot D_{12}(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*) \tag{70}$$

The RHS of (70) is an increasing function of $\mu_{L(i)}$. Moreover, when σ_l goes to infinity, n_{sc} will increase to infinity. Meanwhile, we can derive from (16) that

$$v \geq 1 - \frac{q_1\sigma_{\max}^2}{C_0K^2\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2} \tag{71}$$

Therefore, the lower bound of the convergence rate v in (71) is also an increasing function of $\mu_{L(i)}$. When $\mu_{L(i)}$ increases to infinity, v will converge to 1.
(3) If all σ_l go to zero for all $l \in [d]$. Then from (66) we have

$$\Gamma(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*) \leq \frac{1}{\kappa^2\eta}\mathcal{C}(\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l, \epsilon'_s, \delta'_s) := q_2 \tag{72}$$

for some $\epsilon'_s > 0$ and $\delta'_s > 0$. The we can derive from (65) and (16) that

$$n_{\text{sc}} \geq poly(\epsilon_0^{-1}, \kappa, \tau, K, q_2^{-1}) \cdot d\log^2 d \cdot (\sigma_{\max}\delta_1(\boldsymbol{W}^*))^2 D_{12}(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}, \boldsymbol{W}^*) \tag{73}$$

$$v \geq 1 - \frac{q_2\sigma_{\max}^2}{C_0K^2\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2} \tag{74}$$

From (73) and (74), we have that the sample complexity goes to infinity and the convergence rate goes to 1 as all σ_l go to zero for all $l \in [L]$.

C Proof of Lemma 1

We first state some important lemmas used in proof in Section C.1 and describe the proof in Section C.2. The proofs of these lemmas are provided in Section C.3 to C.7 in sequence. The proof idea mainly follows from [Fu et al. (2020)]. Lemma 6 shows the Hessian $\nabla^2 f(\boldsymbol{W})$ of the population risk function is smooth. Lemma 7 illustrates that $\nabla^2 f(\boldsymbol{W})$ is strongly convex in the neighborhood around $\boldsymbol{\mu}^*$. Lemma 8 shows the Hessian of the empirical risk function $\nabla^2 f_n(\boldsymbol{W}^*)$ is close to its population risk $\nabla^2 f(\boldsymbol{W}^*)$ 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.

C.1 Useful Lemmas in the proof of Lemma 1

Lemma 4

$$\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}\Big[(\sum_{i=1}^k \boldsymbol{p}_i^\top\boldsymbol{x}\cdot\phi'(\sigma\cdot x_i))^2\Big] \geq \rho(\boldsymbol{\mu},\sigma)||\boldsymbol{P}||_F^2\,, \tag{75}$$

where $\rho(\boldsymbol{\mu}, \sigma)$ is defined in Definition 2.

Lemma 5 *With the FCN model (2) and the Gaussian Mixture Model (7), for some constant $C_{12} > 0$, we have*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\sup_{\boldsymbol{W}\neq\boldsymbol{W}'\in\mathbb{B}(\boldsymbol{W}^*,r)}\frac{||\nabla^2\ell(\boldsymbol{W},\boldsymbol{x})-\nabla^2\ell(\boldsymbol{W}',\boldsymbol{x})||}{||\boldsymbol{W}-\boldsymbol{W}'||_F}\Big]\\ \leq& C_{12}\cdot d^{\frac{3}{2}}K^{\frac{5}{2}}\sqrt{\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}\|_\infty+\sigma_l)^2\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}\|_\infty+\sigma_l)^4}\end{aligned} \tag{76}$$

Lemma 6 *(Hessian smoothness of population loss)In the FCN model (2), assume $||\boldsymbol{w}_k^*||_2 \leq 1$ for all k. Then for some constant $C_5 > 0$, we have*

$$||\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^*)|| \leq C_5\cdot K^{\frac{3}{2}}\cdot\Big(\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}\|_\infty+\sigma_l)^4\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\sigma_l)^8\Big)^{\frac{1}{4}}\cdot||\boldsymbol{W}-\boldsymbol{W}^*||_F \tag{77}$$

Lemma 7 *(Local strong convexity of population loss) In the FCN model (2), if $||\boldsymbol{W}-\boldsymbol{W}^*||_F \leq r$ for an $\epsilon_0 \in (0, \frac{1}{4})$, then for some constant $C_4 > 0$,*

$$\frac{4(1-\epsilon_0)}{K^2}\sum_{l=1}^{L}\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))\cdot\boldsymbol{I}_{dK} \preceq \nabla^2 f(\boldsymbol{W}) \preceq C_4\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\cdot\boldsymbol{I}_{dK} \tag{78}$$

Lemma 8 *In the FCN model (2), as long as $n \geq C' \cdot dK\log dK$ for some constant $C' > 0$, we have*

$$\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r_{FCN})}||\nabla^2 f_n(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W})|| \leq C_6\cdot\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}\|_\infty+\sigma_l)^2\sqrt{\frac{dK\log n}{n}}) \tag{79}$$

with probability at least $1-d^{-10}$ for some constant $C_6 > 0$.

C.2 Proof of Lemma 1

From Lemma 7 and 8, with probability at least $1-d^{-10}$,

$$\begin{aligned}\nabla^2 f_n(\boldsymbol{W}) &\succeq \nabla^2 f(\boldsymbol{W}) - ||\nabla^2 f(\boldsymbol{W})-\nabla^2 f_n(\boldsymbol{W})||\cdot\boldsymbol{I}\\ &\succeq \Omega\Big(\frac{1-\epsilon_0}{K^2}\sum_{l=1}^{L}\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))\Big)\cdot\boldsymbol{I}\\ &- O\Big(C_6\cdot\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|+\sigma_l)^2\sqrt{\frac{dK\log n}{n}}\Big)\cdot\boldsymbol{I}\end{aligned} \tag{80}$$

As long as the sample complexity is set to satisfy

$$C_6\cdot\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\sqrt{\frac{dK\log n}{n}} \leq \frac{\epsilon_0}{K^2}\sum_{l=1}^{L}\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*)) \tag{81}$$

i.e.,

$$n \geq C_1\epsilon_0^{-2}\cdot\big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}\|_\infty+\sigma_l)^2\big)^2\big(\sum_{l=1}^{L}\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))\Big)^{-2}dK^5\log^2 d \tag{82}$$

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

$$\nabla^2 f_n(\boldsymbol{W}) \succeq \Omega\Big(\frac{1-2\epsilon_0}{K^2}\sum_{l=1}^{L}\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))\Big)\cdot\boldsymbol{I} \tag{83}$$

By (78) and (79), we can also derive the upper bound as follows,

$$\begin{aligned}||\nabla^2 f_n(\boldsymbol{W})|| &\leq ||\nabla^2 f(\boldsymbol{W})|| + ||\nabla^2 f_n(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W})||\\ &\leq C_4\cdot\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2 + C_6\cdot\sum_{1=1}\lambda_l(\|\boldsymbol{\mu}\|_\infty+\sigma_l)^2\sqrt{\frac{dK\log n}{n}}\\ &\leq C_2\cdot\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\end{aligned} \tag{84}$$

for some constant $C_2 > 0$. Combining (83) and (84), we have

$$\Omega\Big(\frac{1-2\epsilon_0}{K^2}\sum_{l=1}^{L}\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))\cdot\boldsymbol{I} \preceq \nabla^2 f_n(\boldsymbol{W}) \preceq C_2\sum_{l=1}^{L}\lambda_l(||\boldsymbol{\mu}_l||_\infty+\sigma_l)^2\cdot\boldsymbol{I} \tag{85}$$

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

C.3 PROOF OF LEMMA 4

Following the proof idea in Lemma D.4 of (Zhong et al., 2017b), we have

$$\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}\Big[(\sum_{i=1}^{k}\boldsymbol{p}_i^\top\boldsymbol{x}\cdot\phi'(\sigma\cdot x_i))^2\Big]=A_0+B_0 \tag{86}$$

$$A_0=\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}\Big(\sum_{i=1}^{k}\boldsymbol{p}_i^\top\boldsymbol{x}\cdot\phi'^2(\sigma\cdot x_i)\cdot\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{p}_i\Big) \tag{87}$$

$$B_0=\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}\Big(\sum_{i\neq l}\boldsymbol{p}_i^\top\phi'(\sigma\cdot x_i)\phi'(\sigma\cdot x_l)\cdot\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{p}_l\Big) \tag{88}$$

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

$$\begin{aligned}
A_0=&\sum_{i=1}^{k}\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}\Big[\boldsymbol{p}_i^\top\Big(\phi'^2(\sigma\cdot x_i)\Big(x_i^2\boldsymbol{e}_i\boldsymbol{e}_i^\top+\sum_{j\neq i}x_ix_j(\boldsymbol{e}_i\boldsymbol{e}_j^\top\\
&+\boldsymbol{e}_j\boldsymbol{e}_i^\top)+\sum_{j\neq i}\sum_{l\neq i}x_jx_l\boldsymbol{e}_j\boldsymbol{e}_l^\top\Big)\Big)\boldsymbol{p}_i\Big]\\
=&\sum_{i=1}^{k}\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}\Big[\boldsymbol{p}_i^\top\Big(\phi'^2(\sigma\cdot x_i)\Big(x_i^2\boldsymbol{e}_i\boldsymbol{e}_i^\top+\sum_{j\neq i}x_j^2\boldsymbol{e}_j\boldsymbol{e}_j^\top\Big)\Big)\boldsymbol{p}_i\Big]\\
=&\sum_{i=1}^{k}\Big[\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}[\phi'^2(\sigma\cdot x_i)x_i^2]\boldsymbol{p}_i^\top\boldsymbol{e}_i\boldsymbol{e}_i^\top\boldsymbol{p}_i\\
&+\sum_{j\neq i}\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}[x_j^2]\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I})+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I})}[\phi'^2(\sigma\cdot x_i)]\boldsymbol{p}_i^\top\boldsymbol{e}_j\boldsymbol{e}_j^\top\boldsymbol{p}_i\Big]\\
=&\sum_{i=1}^{k}p_{ii}^2\beta_2(i,\boldsymbol{\mu},\sigma)+\sum_{i=1}^{k}\sum_{j\neq i}p_{ij}^2\beta_0(i,\boldsymbol{\mu},\sigma)(1+\mu_j^2)
\end{aligned} \tag{89}$$

In B_0, $\alpha_1(i,\boldsymbol{\mu},\sigma)=\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}(x_i\phi'(x_i))=0$. By the equation in Page 30 of (Zhong et al., 2017b), we have

$$\begin{aligned}
B_0=&\sum_{i\neq l}^{k}\mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu},\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu},\boldsymbol{I}_d)}\Big[\boldsymbol{p}_i^\top\Big(\phi'(\sigma\cdot x_i)\phi'(\sigma\cdot x_l)\Big(x_i^2\boldsymbol{e}_i\boldsymbol{e}_i^\top+x_l^2\boldsymbol{e}_l\boldsymbol{e}_l^\top+x_ix_l(\boldsymbol{e}_i\boldsymbol{e}_l^\top+\\
&\boldsymbol{e}_l\boldsymbol{e}_i^\top)+\sum_{j\neq i}x_jx_l\boldsymbol{e}_j\boldsymbol{e}_l^\top+\sum_{j\neq l}x_jx_i\boldsymbol{e}_j\boldsymbol{e}_i^\top+\sum_{j\neq i,l}\sum_{j'\neq i,l}x_jx_{j'}\boldsymbol{e}_j\boldsymbol{e}_{j'}^\top\Big)\Big)\boldsymbol{p}_l\Big]\\
=&\sum_{i\neq l}p_{ii}p_{li}\alpha_2(i,\boldsymbol{\mu},\sigma)\alpha_0(l,\boldsymbol{\mu},\sigma)+\sum_{i\neq l}p_{ij}p_{lj}\alpha_0(i,\boldsymbol{\mu},\sigma)\alpha_0(l,\boldsymbol{\mu},\sigma)(1+\mu_j^2)
\end{aligned} \tag{90}$$

Therefore,

$$\begin{aligned}
A_0+B_0=&\sum_{i=1}^{k}\Big(p_{ii}\frac{\alpha_2(i,\boldsymbol{\mu},\sigma)}{\sqrt{1+\mu_i^2}}+\sum_{l\neq i}p_{li}\alpha_0(l,\boldsymbol{\mu},\sigma)\sqrt{1+\mu_i^2}\Big)^2-\sum_{i=1}^{k}p_{ii}^2\frac{\alpha_2^2(i,\boldsymbol{\mu},\sigma)}{1+\mu_i^2}\\
&-\sum_{i=1}^{k}\sum_{l\neq i}p_{li}^2\alpha_0(l,\boldsymbol{\mu},\sigma)^2(1+\mu_i^2)+\sum_{i=1}^{k}p_{ii}^2\beta_2(i,\boldsymbol{\mu},\sigma)+\sum_{i=1}^{k}\sum_{j\neq i}p_{ij}^2\beta_0(i,\boldsymbol{\mu},\sigma)(1+\mu_j^2)\\
\geq&\sum_{i=1}^{k}p_{ii}^2\Big(\beta_2(i,\boldsymbol{\mu},\sigma)-\frac{\alpha_2^2(i,\boldsymbol{\mu},\sigma)}{1+\mu_i^2}\Big)+\sum_{i=1}^{k}\sum_{j\neq i}p_{ij}^2\Big(\beta_0(i,\boldsymbol{\mu},\sigma)-\alpha_0^2(i,\boldsymbol{\mu},\sigma)\Big)(1+\mu_j^2)\\
\geq&\rho(\boldsymbol{\mu},\sigma)||P||_F^2
\end{aligned} \tag{91}$$

C.4 PROOF OF LEMMA 5

Following the equation (92) in Lemma 8 of (Fu et al., 2020) and by (46)

$$||\nabla^2\ell(\boldsymbol{W})-\nabla^2\ell(\boldsymbol{W}')||\leq\sum_{j=1}^{K}\sum_{l=1}^{K}|\xi_{j,l}(\boldsymbol{W})-\xi_{j,l}(\boldsymbol{W}')|\cdot||\boldsymbol{x}\boldsymbol{x}^\top|| \tag{92}$$

By Lagrange's inequality, we have

$$|\xi_{j,l}(\boldsymbol{W})-\xi_{j,l}(\boldsymbol{W}')|\leq(\max_k|T_{j,k,l}|)\cdot||\boldsymbol{x}||\cdot\sqrt{K}||\boldsymbol{W}-\boldsymbol{W}'||_F \tag{93}$$

From Lemma 6, we know

$$\max_k|T_{j,k,l}|\leq C_7 \tag{94}$$

By Property 4, we have

$$\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[||\boldsymbol{x}||^{2t}||]\leq d^t(2t-1)!!\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^{2t} \tag{95}$$

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

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\sup_{\boldsymbol{W}\neq\boldsymbol{W}'}\frac{||\nabla^2\ell(\boldsymbol{W})-\nabla^2\ell(\boldsymbol{W}')||}{||\boldsymbol{W}-\boldsymbol{W}'||_F}]\leq K^{\frac{5}{2}}\mathbb{E}[||\boldsymbol{x}||_2^3]\\ \leq&K^{\frac{5}{2}}\sqrt{d\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}\|_\infty+\sigma_l)^2}\sqrt{3d^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4}\\ =&C_{12}\cdot d^{\frac{3}{2}}K^{\frac{5}{2}}\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4}\end{aligned} \tag{96}$$

C.5 PROOF OF LEMMA 6

Let $\boldsymbol{a}=(\boldsymbol{a}_1^\top,\cdots,\boldsymbol{a}_K^\top)^\top\in\mathbb{R}^{dK}$. Let $\Delta_{j,l}\in\mathbb{R}^{d\times d}$ be the (j,l)-th block of $\nabla^2 f(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W}^*)\in\mathbb{R}^{dK\times dK}$. By definition,

$$||\nabla^2 f(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W}^*)||=\max_{||\boldsymbol{a}||=1}\sum_{j=1}^{K}\sum_{l=1}^{K}\boldsymbol{a}_j^\top\Delta_{j,l}\boldsymbol{a}_l \tag{97}$$

By the mean value theorem and (46),

$$\begin{aligned}\Delta_{j,l}=\frac{\partial^2 f(\boldsymbol{W})}{\partial\boldsymbol{w}_j\partial\boldsymbol{w}_l}-\frac{\partial^2 f(\boldsymbol{W}^*)}{\partial\boldsymbol{w}_j^*\partial\boldsymbol{w}_l^*}&=\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[(\xi_{j,l}(\boldsymbol{W})-\xi_{j,l}(\boldsymbol{W}^*))\cdot\boldsymbol{x}\boldsymbol{x}^\top]\\ &=\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\sum_{k=1}^{K}\left\langle\frac{\partial\xi_{j,l}(\boldsymbol{W}')}{\partial\boldsymbol{w}_k'},\boldsymbol{w}_k-\boldsymbol{w}_k^*\right\rangle\cdot\boldsymbol{x}\boldsymbol{x}^\top]\\ &=\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\sum_{k=1}^{K}\langle T_{j,l,k}\cdot\boldsymbol{x},\boldsymbol{w}_k-\boldsymbol{w}_k^*\rangle\cdot\boldsymbol{x}\boldsymbol{x}^\top]\end{aligned} \tag{98}$$

where $\boldsymbol{W}'=\gamma\boldsymbol{W}+(1-\gamma)\boldsymbol{W}^*$ for some $\gamma\in(0,1)$ and $T_{j,l,k}$ is defined such that $\frac{\partial\xi_{j,l}(\boldsymbol{W}')}{\partial\boldsymbol{w}_k'}=T_{j,l,k}\cdot x\in\mathbb{R}^d$. 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 (46) we can obtain

$$\xi_{j,l}(\boldsymbol{W})=\begin{cases}\frac{1}{K^2}\phi'(\boldsymbol{w}_j^\top\boldsymbol{x})\phi'(\boldsymbol{w}_l^\top\boldsymbol{x})\cdot\frac{1}{H^2(\boldsymbol{W})}, & j\neq l\\ \frac{1}{K^2}\phi'(\boldsymbol{w}_j^\top\boldsymbol{x})\phi'(\boldsymbol{w}_l^\top\boldsymbol{x})\cdot\frac{1}{H^2(\boldsymbol{W})}-\frac{1}{K}\phi''(\boldsymbol{w}_j^\top\boldsymbol{x})\cdot\frac{1}{H(\boldsymbol{W})}, & j=l\end{cases} \tag{99}$$

We can bound $\xi_{j,l}(\boldsymbol{W})$ by bounding each components of (99). Note that we have

$$\frac{1}{K^2}\phi'(\boldsymbol{w}_j^\top \boldsymbol{x})\phi'(\boldsymbol{w}_l^\top \boldsymbol{x}) \cdot \frac{1}{H^2(\boldsymbol{W})} \leq \frac{1}{K^2}\frac{\phi(\boldsymbol{w}_j^\top \boldsymbol{x})\phi(\boldsymbol{w}_l^\top \boldsymbol{x})(1-\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))(1-\phi(\boldsymbol{w}_l^\top \boldsymbol{x}))}{\frac{1}{K^2}\phi(\boldsymbol{w}_j^\top \boldsymbol{x})\phi(\boldsymbol{w}_l^\top \boldsymbol{x})} \leq 1 \tag{100}$$

$$\frac{1}{K}\phi''(\boldsymbol{w}_j^\top \boldsymbol{x}) \cdot \frac{1}{H(\boldsymbol{W})} \leq \frac{1}{K}\frac{\phi(\boldsymbol{w}_j^\top \boldsymbol{x})(1-\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))(1-2\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))}{\frac{1}{K}\phi(\boldsymbol{w}_j^\top \boldsymbol{x})} \leq 1 \tag{101}$$

where (100) holds for any $j, l \in [K]$. The case $y = 0$ can be computed with the same upper bound by substituting $(1-H(\boldsymbol{W})) = \frac{1}{K}\sum_{j=1}^{K}(1-\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))$ for $H(\boldsymbol{W})$ in (99), (100) and (101). Therefore, there exists a constant $C_9 > 0$, such that

$$|\xi_{j,l}(\boldsymbol{W})| \leq C_9 \tag{102}$$

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

$$T_{j,l,k} = \frac{-2}{K^3 H^3(\boldsymbol{W}')}\phi'(\boldsymbol{w}'^\top_j \boldsymbol{x})\phi'(\boldsymbol{w}'^\top_l \boldsymbol{x})\phi'(\boldsymbol{w}'^\top_k \boldsymbol{x}), \quad \text{where } j, l, k \text{ are not equal to each other} \tag{103}$$

$$T_{j,j,k} = \begin{cases} \frac{-2}{K^3 H^3(\boldsymbol{W}')}\phi'(\boldsymbol{w}'^\top_j \boldsymbol{x})\phi'(\boldsymbol{w}'^\top_j \boldsymbol{x})\phi'(\boldsymbol{w}'^\top_k \boldsymbol{x}) + \frac{1}{K^2 H^2(\boldsymbol{W}')}\phi''(\boldsymbol{w}'^\top_j \boldsymbol{x})\phi'(\boldsymbol{w}'^\top_k \boldsymbol{x}), & j \neq k \\ \frac{-2}{K^3 H^3(\boldsymbol{W}')}(\phi'(\boldsymbol{w}'^\top_j \boldsymbol{x}))^3 + \frac{3}{K^2 H^2(\boldsymbol{W}')}\phi''(\boldsymbol{w}'^\top_j \boldsymbol{x})\phi'(\boldsymbol{w}'^\top_j \boldsymbol{x}) - \frac{\phi'''(\boldsymbol{w}'^\top_j \boldsymbol{x})}{KH(\boldsymbol{W}')}, & j = k \end{cases} \tag{104}$$

$$\begin{aligned}
\boldsymbol{a}_j^\top \Delta_{j,l}\boldsymbol{a}_l &= \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I})}[(\sum_{k=1}^{K} T_{j,l,k}\langle \boldsymbol{x}, \boldsymbol{w}_k - \boldsymbol{w}_k^*\rangle) \cdot (\boldsymbol{a}_j^\top \boldsymbol{x})(\boldsymbol{a}_l^\top \boldsymbol{x})] \\
&\leq \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I})}[\sum_{k=1}^{K} T_{j,k,l}^2] \cdot \mathbb{E}[\sum_{k=1}^{K}(\langle \boldsymbol{x}, \boldsymbol{w}_k - \boldsymbol{w}_k^*\rangle (\boldsymbol{a}_j^\top \boldsymbol{x})(\boldsymbol{a}_l^\top \boldsymbol{x}))^2]} \\
&\leq \sqrt{\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I})}[\sum_{k=1}^{K} T_{j,k,l}^2]}\sqrt{\sum_{k=1}^{K}\sqrt{\mathbb{E}((\boldsymbol{w}_k - \boldsymbol{w}_k^*)^\top \boldsymbol{x})^4} \cdot \sqrt{\mathbb{E}[(\boldsymbol{a}_j^\top \boldsymbol{x})^4(\boldsymbol{a}_l^\top \boldsymbol{x})^4]}} \\
&\leq C_8\sqrt{\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I})}[\sum_{k=1}^{K} T_{j,k,l}^2]}\sqrt{\sum_{k=1}^{K}||\boldsymbol{w}_k - \boldsymbol{w}_k^*||_2^2 \cdot ||\boldsymbol{a}_j||_2^2 \cdot ||\boldsymbol{a}_l||_2^2} \\
&\quad \cdot \Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^4 \sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^8\Big)^{\frac{1}{4}}
\end{aligned} \tag{105}$$

for some constant $C_8 > 0$. All the three inequalities of (105) are derived from Cauchy-Schwarz inequality. Note that we have

$$\begin{aligned}
\Big|\frac{-2}{K^3 H^3(\boldsymbol{W})}(\phi'(\boldsymbol{w}_j^\top \boldsymbol{x}))^2\phi'(\boldsymbol{w}_k^\top \boldsymbol{x})\Big| &\leq \frac{2\phi^2(\boldsymbol{w}_j^\top \boldsymbol{x})(1-\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))^2\phi(\boldsymbol{w}_k^\top \boldsymbol{x})(1-\phi(\boldsymbol{w}_k^\top \boldsymbol{x}))}{K^3\frac{1}{K^3}\phi^2(\boldsymbol{w}_j^\top \boldsymbol{x})\phi(\boldsymbol{w}_k^\top \boldsymbol{x})} \\
&= 2(1-\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))^2(1-\phi(\boldsymbol{w}_k^\top \boldsymbol{x})) \leq 2
\end{aligned} \tag{106}$$

$$\Big|\frac{-2}{K^3 H^3(\boldsymbol{W})}\phi'(\boldsymbol{w}_j^\top \boldsymbol{x})\phi'(\boldsymbol{w}_l^\top \boldsymbol{x})\phi'(\boldsymbol{w}_k^\top \boldsymbol{x})\Big| \leq 2 \tag{107}$$

$$\begin{aligned}
&\Big|\frac{3}{K^2 H^2(\boldsymbol{W})}\phi''(\boldsymbol{w}_j^\top \boldsymbol{x})\phi'(\boldsymbol{w}_k^\top \boldsymbol{x})\Big| \\
\leq &\Big|\frac{3\phi(\boldsymbol{w}_j^\top \boldsymbol{x})(1-\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))(1-2\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))\phi(\boldsymbol{w}_k^\top \boldsymbol{x})(1-\phi(\boldsymbol{w}_k^\top \boldsymbol{x}))}{K^2\frac{1}{K^2}\phi(\boldsymbol{w}_j^\top \boldsymbol{x})\phi(\boldsymbol{w}_k^\top \boldsymbol{x})}\Big| \\
= &\Big|3(1-\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))(1-2\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))(1-\phi(\boldsymbol{w}_k^\top \boldsymbol{x}))\Big| \leq 3
\end{aligned} \tag{108}$$

$$\Big|\frac{\phi'''(\boldsymbol{w}_j^\top \boldsymbol{x})}{KH(\boldsymbol{W})}\Big| \leq \Big|\frac{\phi(\boldsymbol{w}_j^\top \boldsymbol{x})(1-\phi(\boldsymbol{w}_j^\top \boldsymbol{x}))(1-6\phi(\boldsymbol{w}_j^\top \boldsymbol{x})+6\phi^2(\boldsymbol{w}_j^\top \boldsymbol{x}))}{K\frac{1}{K}\phi(\boldsymbol{w}_j^\top \boldsymbol{x})}\Big| \leq 1 \tag{109}$$

Therefore, by combining (103), (104) and (106) to (109), we have

$$|T_{j,l,k}| \leq C_7 \quad \Rightarrow \quad T_{j,l,k}^2 \leq C_7^2, \forall j,l,k \in [K], \tag{110}$$

for some constants $C_7 > 0$. By (97), (98), (105), (110) and the Cauchy-Schwarz's Inequality, we have

$$\begin{aligned}
&\|\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^*)\| \\
\leq& C_8\sqrt{C_7^2 K}||\boldsymbol{W}-\boldsymbol{W}^*||_F\Big(\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^8\Big)^{\frac{1}{4}} \\
&\cdot \max_{||\boldsymbol{a}||=1}\sum_{j=1}^K\sum_{l=1}^K ||\boldsymbol{a}_j||_2||\boldsymbol{a}_l||_2 \\
\leq& C_8\sqrt{C_7^2 K}\cdot||\boldsymbol{W}-\boldsymbol{W}^*||_F\cdot\Big(\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^8\Big)^{\frac{1}{4}}\cdot\Big(\sum_{j=1}^K ||\boldsymbol{a}_j||\Big)^2 \\
\leq& C_8\sqrt{C_7^2 K^3}\cdot||\boldsymbol{W}-\boldsymbol{W}^*||_F\cdot\Big(\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^8\Big)^{\frac{1}{4}}
\end{aligned} \tag{111}$$

Hence, we have

$$||\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W}^*)|| \leq C_5 K^{\frac{3}{2}}\Big(\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^8\Big)^{\frac{1}{4}}||\boldsymbol{W}-\boldsymbol{W}^*||_F \tag{112}$$

for some constant $C_5 > 0$.

C.6 Proof of Lemma 7

From (Fu et al. (2020)), we know

$$\nabla^2 f(\boldsymbol{W}^*) \succeq \min_{||\boldsymbol{a}||=1}\frac{4}{K^2}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\Big(\sum_{j=1}^K \phi'(\boldsymbol{w}_j^{*\top}\boldsymbol{x})(\boldsymbol{a}_j^\top\boldsymbol{x})\Big)^2\Big]\cdot\boldsymbol{I}_{dK} \tag{113}$$

with $\boldsymbol{a} = (\boldsymbol{a}_1^\top, \cdots, \boldsymbol{a}_K^\top)^\top \in \mathbb{R}^{dK}$. And

$$\nabla^2 f(\boldsymbol{W}^*) \preceq \Big(\max_{||\boldsymbol{a}||=1}\boldsymbol{a}^\top\nabla^2 f(\boldsymbol{W}^*)\boldsymbol{a}\Big)\cdot\boldsymbol{I}_{dK} \preceq C_4\cdot\max_{||\boldsymbol{a}||=1}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\sum_{j=1}^K(\boldsymbol{a}_j^\top\boldsymbol{x})^2\Big]\cdot\boldsymbol{I}_{dK} \tag{114}$$

for some constant $C_4 > 0$. By applying Property 3, we can derive the upper bound in (114) as

$$C_4\cdot\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\sum_{j=1}^K(\boldsymbol{a}_j^\top\boldsymbol{x})^2\Big]\cdot\boldsymbol{I}_{dK} \preceq C_4\cdot\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\cdot\boldsymbol{I}_{dK} \tag{115}$$

To find a lower bound for (113), we can first transfer the expectation of the Gaussian Mixture Model to the weight sum of the expectations over general Gaussian distributions.

$$\begin{aligned}
&\min_{||\boldsymbol{a}||=1}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\Big(\sum_{j=1}^K \phi'(\boldsymbol{w}_j^{*\top}\boldsymbol{x})(\boldsymbol{a}_j^\top\boldsymbol{x})\Big)^2\Big] \\
=&\min_{||\boldsymbol{a}||=1}\sum_{l=1}^L \lambda_l\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\Big(\sum_{j=1}^K \phi'(\boldsymbol{w}_j^{*\top}\boldsymbol{x})(\boldsymbol{a}_j^\top\boldsymbol{x})\Big)^2\Big]
\end{aligned} \tag{116}$$

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

$$\boldsymbol{a}_i = \boldsymbol{U}\boldsymbol{b}_i + \boldsymbol{U}_\perp \boldsymbol{c}_i \tag{117}$$

where $\boldsymbol{U}_\perp \in \mathbb{R}^{d\times(d-K)}$ denotes the complement of $\boldsymbol{U}$. We also have $\boldsymbol{U}_\perp^\top \boldsymbol{\mu}_l = 0$ by (9). Plugging (117) into RHS of (116), and then we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\Big(\sum_{i=1}^K \boldsymbol{a}_i^\top \boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\Big)^2\Big] \\
=&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\Big(\sum_{i=1}^K (\boldsymbol{U}\boldsymbol{b}_i + \boldsymbol{U}_\perp\boldsymbol{c}_i)^\top \boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\Big)^2\Big] = A + B + C
\end{aligned} \tag{118}$$

$$A = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\Big(\sum_{i=1}^K \boldsymbol{b}_i^\top \boldsymbol{U}^\top \boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\Big)^2\Big] \tag{119}$$

$$\begin{aligned}
C &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[2\Big(\sum_{i=1}^K \boldsymbol{c}_i^\top \boldsymbol{U}_\perp^\top \boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\Big)\cdot\Big(\sum_{i=1}^K \boldsymbol{b}_i^\top \boldsymbol{U}^\top \boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\Big)\Big] \\
&= \sum_{i=1}^K\sum_{j=1}^K \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[2\boldsymbol{c}_i^\top \boldsymbol{U}_\perp^\top \boldsymbol{x}\Big]\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\boldsymbol{b}_i^\top \boldsymbol{U}^\top \boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\phi'(\boldsymbol{w}_j^{*\top}\boldsymbol{x})\Big] \\
&= \sum_{i=1}^K\sum_{j=1}^K \Big[2\boldsymbol{c}_i^\top \boldsymbol{U}_\perp^\top \boldsymbol{\mu}_l\Big]\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\boldsymbol{b}_i^\top \boldsymbol{U}^\top \boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\phi'(\boldsymbol{w}_j^{*\top}\boldsymbol{x})\Big] = 0
\end{aligned} \tag{120}$$

where the last step is by $\boldsymbol{U}_\perp^\top \boldsymbol{\mu}_l = 0$ by (9).

$$\begin{aligned}
B =&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[(\sum_{i=1}^K \boldsymbol{c}_i^\top \boldsymbol{U}_\perp^\top \boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}))^2\Big] \\
=&\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[(\boldsymbol{t}^\top\boldsymbol{s})^2] \qquad \text{by defining } \boldsymbol{t} = \sum_{i=1}^k \phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\boldsymbol{c}_i \in \mathbb{R}^{d-K} \text{ and } \boldsymbol{s} = \boldsymbol{U}_\perp^\top\boldsymbol{x} \\
=&\sum_{i=1}^K \mathbb{E}[t_i^2 s_i^2] + \sum_{i\neq j}\mathbb{E}[t_i t_j s_i s_j] \\
=&\sum_{i=1}^K \mathbb{E}[t_i^2]\sigma_l^2 + \Big(\sum_{i=1}^K \mathbb{E}[t_i^2](\boldsymbol{U}_\perp^\top\boldsymbol{\mu}_l)_i^2 + \sum_{i\neq j}\mathbb{E}[t_i t_j](\boldsymbol{U}_\perp^\top\boldsymbol{\mu}_l)_i\cdot(\boldsymbol{U}_\perp^\top\boldsymbol{\mu}_l)_j\Big) \\
=&\mathbb{E}[\sum_{i=1}^{d-K} t_i^2\sigma_l^2] + \mathbb{E}[(\boldsymbol{t}^\top\boldsymbol{U}_\perp^\top\boldsymbol{\mu}_l)^2] = \mathbb{E}[\sum_{i=1}^{d-K} t_i^2\sigma_l^2]
\end{aligned} \tag{121}$$

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

$$\mathbb{E}[s_i s_j] = \begin{cases} (\boldsymbol{U}_\perp^\top\boldsymbol{\mu}_l)_i\cdot(\boldsymbol{U}_\perp^\top\boldsymbol{\mu}_l)_j, & \text{if } i\neq j \\ (\boldsymbol{U}_\perp^\top\boldsymbol{\mu}_l)_i^2 + \sigma_l^2, & \text{if } i = j \end{cases} \tag{122}$$

Since $\Big(\sum_{i=1}^k \boldsymbol{p}_i^\top\boldsymbol{x}\cdot\phi'(\sigma\cdot x_i)\Big)^2$ is an even function, so from Property 2 we have

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[(\sum_{i=1}^k \boldsymbol{p}_i^\top\boldsymbol{x}\cdot\phi'(\sigma\cdot x_i))^2\Big] = \mathbb{E}_{\boldsymbol{x}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)+\frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[(\sum_{i=1}^k \boldsymbol{p}_i^\top\boldsymbol{x}\cdot\phi'(\sigma\cdot x_i))^2\Big] \tag{123}$$

Combining Lemma 4 and Property 2, we next follow the derivation for the standard Gaussian distribution in Page 36 of (Zhong et al., 2017b) and generalize the result to a Gaussian distribution

with an arbitrary mean and variance as follows.

$$
\begin{aligned}
A &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\Big(\sum_{i=1}^k \boldsymbol{b}_i^\top \boldsymbol{U}^\top\boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x})\Big)^2\Big] \\
&= \int (2\pi\sigma_l^2)^{-\frac{K}{2}}\Big[\Big(\sum_{i=1}^k \boldsymbol{b}_i^\top \boldsymbol{z}\cdot\phi'(\boldsymbol{v}_i^\top\boldsymbol{z})\Big)^2\Big]\exp\big(-\frac{||\boldsymbol{z}-\boldsymbol{U}^\top\boldsymbol{\mu}_l||^2}{2\sigma_l^2}\big)d\boldsymbol{z} \\
&= \int (2\pi\sigma_l^2)^{-\frac{K}{2}}\Big[\Big(\sum_{i=1}^k \boldsymbol{b}_i^\top \boldsymbol{V}^{\dagger\top}\boldsymbol{s}\cdot\phi'(s_i)\Big)^2\Big]\exp\big(-\frac{||\boldsymbol{V}^{\dagger\top}\boldsymbol{s}-\boldsymbol{U}^\top\boldsymbol{\mu}_l||^2}{2\sigma_l^2}\big)\Big|\det(\boldsymbol{V}^\dagger)\Big|d\boldsymbol{s} \\
&\geq \int (2\pi\sigma_l^2)^{-\frac{K}{2}}\Big[\Big(\sum_{i=1}^k \boldsymbol{b}_i^\top \boldsymbol{V}^{\dagger\top}\boldsymbol{s}\cdot\phi'(s_i)\Big)^2\Big]\exp\big(-\frac{||\boldsymbol{s}-\boldsymbol{V}^\top\boldsymbol{U}^\top\boldsymbol{\mu}_l||^2}{2\sigma_l^2\delta_K^2(\boldsymbol{W}^*)}\big)\Big|\det(\boldsymbol{V}^\dagger)\Big|d\boldsymbol{s} \\
&= \int (2\pi\sigma_l^2)^{-\frac{K}{2}}\Big[\Big(\sum_{i=1}^k \boldsymbol{b}_i^\top \boldsymbol{V}^{\dagger\top}(\delta_K(\boldsymbol{W}^*)\sigma_l)\boldsymbol{g}\cdot\phi'(\delta_K(\boldsymbol{W}^*)\sigma_l\cdot g_i)\Big)^2\Big] \\
&\quad\cdot\exp\big(-\frac{||\boldsymbol{g}-\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)}||^2}{2}\big)\Big|\det(\boldsymbol{V}^\dagger)\Big|\sigma_l^K(\boldsymbol{W}^*)\delta_K^K(\boldsymbol{W}^*)d\boldsymbol{g} \\
&= \frac{\sigma_l^2}{\eta}\mathbb{E}_{\boldsymbol{g}}\Big[(\sum_{i=1}^K(\boldsymbol{b}_i^\top\boldsymbol{V}^{\dagger\top}\delta_K(\boldsymbol{W}^*))\boldsymbol{g}\cdot\phi'(\sigma_l\delta_K(\boldsymbol{W}^*)\cdot g_i))^2\Big] \\
&\geq \frac{\sigma_l^2}{\kappa^2\eta}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))||\boldsymbol{b}||^2.
\end{aligned}
\tag{124}
$$

The second step is by letting $\boldsymbol{z}=\boldsymbol{U}^\top\boldsymbol{x}$. The third step is by letting $\boldsymbol{s}=\boldsymbol{V}^\top\boldsymbol{z}$. The last to second step follows from $\boldsymbol{g}=\frac{\boldsymbol{s}}{\sigma_l\delta_K(\boldsymbol{W}^*)}$, where $\boldsymbol{g}\sim\mathcal{N}(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\boldsymbol{I}_K)$ and the last inequality is by Lemma 4. Similarly, we extend the derivation in Page 37 of (Zhong et al., 2017b) for the standard Gaussian distribution to a general Gaussian distribution as follows.

$$
B=\sigma_l^2\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[||\boldsymbol{t}||^2]\geq\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))||\boldsymbol{c}||^2 \tag{125}
$$

Combining (118) - (121), (124) and (125), we have

$$
\min_{||\boldsymbol{a}||=1}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[(\sum_{i=1}^k\boldsymbol{a}_i^\top\boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}))^2\Big]\geq\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*)). \tag{126}
$$

For the Gaussian Mixture Model $\boldsymbol{x}\sim\sum_{l=1}^L\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)$, we have

$$
\min_{||\boldsymbol{a}||=1}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[(\sum_{i=1}^k\boldsymbol{a}_i^\top\boldsymbol{x}\cdot\phi'(\boldsymbol{w}_i^{*\top}\boldsymbol{x}))^2\Big]\geq\sum_{l=1}^L\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*)) \tag{127}
$$

Therefore,

$$
\frac{4}{K^2}\sum_{l=1}^L\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_k(\boldsymbol{W}^*))\cdot\boldsymbol{I}_{dK}\preceq\nabla^2 f(\boldsymbol{W}^*)\preceq C_4\cdot\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\cdot\boldsymbol{I}_{dK} \tag{128}
$$

From (77) in Lemma 6, since that we have the condition $\|\boldsymbol{W}-\boldsymbol{W}^*\|_F\leq r$ and (54), we can obtain

$$
\begin{aligned}
&||\nabla^2 f(\boldsymbol{W})-\nabla^2 f(\boldsymbol{W}^*)|| \\
\leq& C_5K^{\frac{3}{2}}\Big(\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^8\Big)^{\frac{1}{4}}||\boldsymbol{W}-\boldsymbol{W}^*||_F \\
\leq& \frac{4\epsilon_0}{K^2}\sum_{l=1}^L\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*)),
\end{aligned}
\tag{129}
$$

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

$$\begin{aligned} ||\nabla^2 f(\boldsymbol{W})|| &\geq ||\nabla^2 f(\boldsymbol{W}^*)|| - ||\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W})|| \\ &\geq \frac{4(1-\epsilon_0)}{K^2} \sum_{l=1}^{L} \lambda_l \frac{\sigma_l^2}{\eta\kappa^2} \rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l \delta_K(\boldsymbol{W}^*)}, \sigma_l \delta_K(\boldsymbol{W}^*)) \end{aligned} \tag{130}$$

$$\begin{aligned} ||\nabla^2 f(\boldsymbol{W})|| &\leq ||\nabla^2 f(\boldsymbol{W}^*)|| + ||\nabla^2 f(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W})|| \\ &\leq C_4 \cdot \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2 + \frac{4}{K^2} \sum_{l=1}^{L} \lambda_l \frac{\sigma_l^2}{\eta\kappa^2} \rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l \delta_K(\boldsymbol{W}^*)}, \sigma_l \delta_K(\boldsymbol{W}^*)) \\ &\lesssim C_4 \cdot \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}\|_\infty + \sigma_l)^2 \end{aligned} \tag{131}$$

The last inequality of (131) holds since $C_4 \cdot \sum_{l=1} \lambda_l (\|\boldsymbol{\mu}\|_\infty + \sigma_l)^2 = \Omega(\sigma_{\max}^2)$, $\frac{4}{K^2} \sum_{l=1}^{L} \lambda_l \frac{\sigma_l^2}{\eta\kappa^2} \rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l \delta_K(\boldsymbol{W}^*)}, \sigma_l \delta_K(\boldsymbol{W}^*)) = O(\frac{\sigma_{\max}^2}{K^2})$ and $O(\sigma_{\max}^2) \geq \Omega(\frac{\sigma_{\max}^2}{K^2})$. Combining (130) and (131), we have

$$\frac{4(1-\epsilon_0)}{K^2} \sum_{l=1}^{L} \lambda_l \frac{\sigma_l^2}{\eta\kappa^2} \rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l \delta_K(\boldsymbol{W}^*)}, \sigma_l \delta_K(\boldsymbol{W}^*)) \cdot \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{W}) \preceq C_4 \cdot \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}\|_\infty + \sigma_l)^2 \cdot \boldsymbol{I} \tag{132}$$

C.7 Proof of Lemma 8

Let N_ϵ be the ϵ-covering number of the Euclidean ball $\mathbb{B}(\boldsymbol{W}^*, r)$. It is known that $\log N_\epsilon \leq dK \log(\frac{3r}{\epsilon})$ from (Vershynin, 2010). Let $\mathcal{W}_\epsilon = \{\boldsymbol{W}_1, ..., \boldsymbol{W}_{N_\epsilon}\}$ be the ϵ-cover set with N_ϵ elements. For any $\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^*, r)$, let $j(\boldsymbol{W}) = \arg\min_{j \in [N_\epsilon]} ||\boldsymbol{W} - \boldsymbol{W}_{j(\boldsymbol{W})}||_F \leq \epsilon$ for all $\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^*, r)$.

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

$$\begin{aligned} &\|\nabla^2 f_n(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W})\| \\ \leq &\frac{1}{n} || \sum_{i=1}^{n} [\nabla^2 \ell(\boldsymbol{W}; \boldsymbol{x}_i) - \nabla^2 \ell(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i)] || \\ &+ || \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} [\nabla^2 \ell(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i)] || \\ &+ || \mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} [\nabla^2 \ell(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i)] - \mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} [\nabla^2 \ell(\boldsymbol{W}; \boldsymbol{x}_i)] || \end{aligned} \tag{133}$$

Hence, we have

$$\mathbb{P}\Big(\sup_{\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^*, r)} ||\nabla^2 f_n(\boldsymbol{W}) - \nabla^2 f(\boldsymbol{W})|| \geq t \Big) \leq \mathbb{P}(A_t) + \mathbb{P}(B_t) + \mathbb{P}(C_t) \tag{134}$$

where A_t, B_t and C_t are defined as

$$A_t = \{ \sup_{\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^*, r)} \frac{1}{n} || \sum_{i=1}^{n} [\nabla^2 \ell(\boldsymbol{W}; \boldsymbol{x}_i) - \nabla^2 \ell(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i)] || \geq \frac{t}{3} \} \tag{135}$$

$$B_t = \{ \sup_{\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^*, r)} || \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} [\nabla^2 \ell(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i)] || \geq \frac{t}{3} \} \tag{136}$$

$$\begin{aligned} C_t = \{ &\sup_{\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^*, r)} || \mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} [\nabla^2 \ell(\boldsymbol{W}_{j(\boldsymbol{W})}; \boldsymbol{x}_i)] \\ &- \mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} [\nabla^2 \ell(\boldsymbol{W}; \boldsymbol{x}_i)] || \geq \frac{t}{3} \} \end{aligned} \tag{137}$$

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 (Fu et al., 2020), we obtain

$$\begin{aligned}&\Big|\Big|\frac{1}{n}\sum_{i=1}^{n}\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)-\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)]\Big|\Big|\\ \leq& 2\sup_{\boldsymbol{v}\in\boldsymbol{V}_{\frac{1}{4}}}\Big|\Big\langle\boldsymbol{v},(\frac{1}{n}\sum_{i=1}^{n}\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)-\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)])\boldsymbol{v}\Big\rangle\Big|\end{aligned}\tag{138}$$

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

$$\begin{aligned}\mathbb{P}(B_t)\leq&\mathbb{P}\Big(\sup_{\boldsymbol{W}\in\mathcal{W}_\epsilon,\boldsymbol{v}\in\boldsymbol{V}_{\frac{1}{4}}}\Big|\frac{1}{n}\sum_{i=1}^{n}G_i\Big|\geq\frac{t}{6}\Big)\\ \leq&\exp(dK(\log\frac{3r}{\epsilon}+\log 12))\sup_{\boldsymbol{W}\in\mathcal{W}_\epsilon,\boldsymbol{v}\in\boldsymbol{V}_{\frac{1}{4}}}\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n}G_i|\geq\frac{t}{6})\end{aligned}\tag{139}$$

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

$$\begin{aligned}|G_i|=&\Big|\sum_{j=1}^{K}\sum_{l=1}^{K}\Big[\xi_{j,l}\boldsymbol{u}_j^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{u}_l-\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I})}(\xi_{j,l}\boldsymbol{u}_j^\top\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{u}_l)\Big]\Big|\\ \leq& C_9\cdot\Big[\sum_{j=1}^{K}(\boldsymbol{u}_j^\top\boldsymbol{x})^2+\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}(\boldsymbol{u}_j^\top\boldsymbol{x})^2\Big]\end{aligned}\tag{140}$$

for some $C_9>0$. The first step of (140) is by (45). The last step is by (102) 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}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)} \\
&\quad \cdot \Big[(\sum_{j=1}^{K}(\boldsymbol{u}_j^\top \boldsymbol{x})^2)^l\Big]\Big(\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}(\boldsymbol{u}_j^\top \boldsymbol{x})^2\Big)^{p-l} \\
&= \sum_{l=1}^{p} \binom{p}{l} C_9 \cdot \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\sum_{l_1+\cdots+l_K=l}\frac{l!}{\prod_{j=1}^{K} l_j!}\prod_{j=1}^{K}(\boldsymbol{u}_j^\top \boldsymbol{x})^{2l_j}\Big] \\
&\quad \cdot \Big(\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}(\boldsymbol{u}_j^\top \boldsymbol{x})^2\Big)^{p-l} \\
&= \sum_{l=1}^{p} \binom{p}{l} C_9 \cdot \Big[\sum_{l_1+\cdots+l_K=l}\frac{l!}{\prod_{j=1}^{K} l_j!}\prod_{j=1}^{K}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}(\boldsymbol{u}_j^\top \boldsymbol{x})^{2l_j}\Big] \\
&\quad \cdot \Big(\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}(\boldsymbol{u}_j^\top \boldsymbol{x})^2\Big)^{p-l} \\
&= C_9 \cdot \sum_{l=1}^{p} \binom{p}{l}\Big(\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}(\boldsymbol{u}_j^\top \boldsymbol{x})^2\Big)^{l} \\
&\quad \cdot \Big(\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}(\boldsymbol{u}_j^\top \boldsymbol{x})^2\Big)^{p-l} \\
&= C_9 \cdot \Big(\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}(\boldsymbol{u}_j^\top \boldsymbol{x})^2\Big)^{p} \\
&\leq C_9 \cdot \Big(\sum_{j=1}^{K} 1!!||\boldsymbol{u}_j||^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\Big)^{p} \\
&\leq C_9 \cdot \Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\Big)^{p}
\end{aligned}
\tag{141}
$$

where the second to last inequality results from Property 3. The last inequality is because $\boldsymbol{v} \in \boldsymbol{V}_{\frac{1}{4}}$, $\sum_{j=1}^{K}||u_j||^2 = ||\boldsymbol{v}||^2 \leq 1$.

$$
\begin{aligned}
\mathbb{E}[\exp(\theta G_i)] &= 1 + \theta\mathbb{E}[G_i] + \sum_{p=2}^{\infty}\frac{\theta^p\mathbb{E}[|G_i|^p]}{p!} \\
&\leq 1 + \sum_{p=2}^{\infty}\frac{|e\theta|^p}{p^p}C_9 \cdot \Big(\sum_{l=1}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\Big)^{p} \\
&\leq 1 + C_9 \cdot |e\theta|^2\Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\Big)^{2}
\end{aligned}
\tag{142}
$$

where the first inequality holds from $p! \geq (\frac{p}{e})^p$ and (141), and the third line holds provided that

$$
\max_{p\geq 2}\Big\{\frac{\frac{|e\theta|^{(p+1)}}{(p+1)^{(p+1)}} \cdot \Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\Big)^{p+1}}{\frac{|e\theta|^p}{p^p} \cdot \Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\Big)^{p}}\Big\} \leq \frac{1}{2}
\tag{143}
$$

Note that the quantity inside the maximization in (143) achieves its maximum when $p = 2$, because it is monotonously decreasing. Therefore, (143) holds if $\theta \leq \frac{27}{4e}\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2$. Then

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}G_i \geq \frac{t}{6}\Big) &= \mathbb{P}\Big(\exp(\theta\sum_{i=1}^{n}G_i) \geq \exp(\frac{n\theta t}{6})\Big) \leq e^{-\frac{n\theta t}{6}}\prod_{i=1}^{n}\mathbb{E}[\exp(\theta G_i)] \\ &\leq \exp(C_{10}\theta^2 n\Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2\Big)^2 - \frac{n\theta t}{6})\end{aligned} \tag{144}$$

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

$$\begin{aligned}\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}G_i \geq \frac{t}{6}) \leq \exp\Big(\max\{&-\frac{C_{10}nt^2}{144\Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2\Big)^2}, \\ &- C_{11}n\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2 \cdot t\}\Big)\end{aligned} \tag{145}$$

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

$$\begin{aligned}\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n}G_i| \geq \frac{t}{6}) \leq 2\exp\Big(\max\{&-\frac{C_{10}nt^2}{144\Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2\Big)^2}, \\ &- C_{11}n\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2 \cdot t\}\Big)\end{aligned} \tag{146}$$

Thus, as long as

$$t \geq C_6 \cdot \max\{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2\sqrt{\frac{dK\log\frac{36r}{\epsilon} + \log\frac{4}{\delta}}{n}}, \frac{dK\log\frac{36r}{\epsilon} + \log\frac{4}{\delta}}{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2 n}\} \tag{147}$$

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

$$\begin{aligned}&\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)}||\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla^2\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla^2\ell(\boldsymbol{W};\boldsymbol{x})]|| \\ \leq &\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)}\frac{||\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla^2\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla^2\ell(\boldsymbol{W};\boldsymbol{x})]||}{||\boldsymbol{W} - \boldsymbol{W}_{j(\boldsymbol{W})}||_F} \\ &\cdot \sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)}||\boldsymbol{W} - \boldsymbol{W}_{j(\boldsymbol{W})}||_F \\ \leq &C_{12}\cdot d^{\frac{3}{2}}K^{\frac{5}{2}}\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^4}\cdot\epsilon\end{aligned} \tag{148}$$

Therefore, C_t holds if

$$t \geq C_{12}\cdot d^{\frac{3}{2}}K^{\frac{5}{2}}\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^4}\cdot\epsilon \tag{149}$$

We can bound the A_t as below.

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)}\frac{1}{n}||\sum_{i=1}^{n}[\nabla^2\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i)-\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)]||\geq\frac{t}{3}\Big)\\
&\leq\frac{3}{t}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)}\frac{1}{n}||\sum_{i=1}^{n}[\nabla^2\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i)-\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)]||\Big]\\
&=\frac{3}{t}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}\Big[\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)}||\nabla^2\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i)-\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)||\Big]\\
&\leq\frac{3}{t}\mathbb{E}\Big[\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)}\frac{||\nabla^2\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i)-\nabla^2\ell(\boldsymbol{W};\boldsymbol{x}_i)||}{||\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}||_F}\Big]\cdot\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)}||\boldsymbol{W}-\boldsymbol{W}_{j(\boldsymbol{W})}||_F\\
&\leq\frac{C_{12}\cdot d^{\frac{3}{2}}K^{\frac{5}{2}}\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4}\cdot\epsilon}{t}
\end{aligned}\tag{150}$$

Thus, taking

$$t\geq\frac{C_{12}\cdot d^{\frac{3}{2}}K^{\frac{5}{2}}\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4}\cdot\epsilon}{\delta}\tag{151}$$

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

3) **Final step**
Let $\epsilon=\frac{\delta}{C_{12}\cdot d^{\frac{3}{2}}K^{\frac{5}{2}}\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4}\cdot ndK}$ and $\delta=d^{-10}$, then from (147) and (151) we need

$$\begin{aligned}
t>\max\{&\frac{1}{ndK},\ C_6\cdot\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\\
&\cdot\sqrt{\frac{dK\log(36rnd^{\frac{25}{2}}K^{\frac{7}{2}}\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4})+\log\frac{4}{\delta}}{n}},\\
&\frac{dK\log(36rnd^{\frac{25}{2}}K^{\frac{7}{2}}\cdot\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^4})+\log\frac{4}{\delta}}{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2n}\}
\end{aligned}\tag{152}$$

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

$$\mathbb{P}(\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)}||\nabla^2f_n(\boldsymbol{W})-\nabla^2f(\boldsymbol{W})||\geq C_6\cdot\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\sqrt{\frac{dK\log n}{n}})\leq d^{-10}\tag{153}$$

D Proof of Lemma 2

We first present a lemma used in proving Lemma 2 in Section D.1 and then prove Lemma 2 in Section D.2.

D.1 A useful lemma used in the proof

Lemma 9 *If r is defined in (54) for $\epsilon_0\in(0,\frac{1}{4})$, then with probability at least $1-d^{-10}$, we have*

$$\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r_{FCN})}||\nabla f_n(\boldsymbol{W})-\nabla f(\boldsymbol{W})||\leq C_{13}\cdot\sqrt{K\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}\|_\infty+\sigma_l)^2}\sqrt{\frac{d\log n}{n}}\tag{154}$$

for some constant $C_{13}>0$.

Proof:
Similar to the idea of the proof of Lemma 8, we adopt an ϵ-covering net of the ball $\mathbb{B}(\boldsymbol{W}^*, r)$ 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 $\nabla f_n(\boldsymbol{W})$ and $\nabla f(\boldsymbol{W})$ into three parts, similar to (133). (155) to (157) can be derived in a similar way as (135) to (137), 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.

$$A_t' = \{\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)} \frac{1}{n}||\sum_{i=1}^{n}[\nabla\ell(\boldsymbol{W};\boldsymbol{x}_i) - \nabla\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i)]|| \geq \frac{t}{3}\} \tag{155}$$

$$B_t' = \{\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)} ||\frac{1}{n}\sum_{i=1}^{n}\nabla\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i)]|| \geq \frac{t}{3}\} \tag{156}$$

$$\begin{aligned} C_t' =&\{\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*,r)} ||\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i)] \\ &- \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla\ell(\boldsymbol{W};\boldsymbol{x}_i)]|| \geq \frac{t}{3}\} \end{aligned} \tag{157}$$

(a) Upper bound of $\mathbb{P}(B_t')$. Applying Lemma 3 in (Mei et al., 2018), we have

$$\begin{aligned} &||\frac{1}{n}\sum_{i=1}^{n}\nabla\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i)]|| \\ \leq& 2\sup_{\boldsymbol{v}\in V_{\frac{1}{2}}}\left|\left\langle\frac{1}{n}\sum_{i=1}^{n}\nabla\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla\ell(\boldsymbol{W}_{j(\boldsymbol{W})};\boldsymbol{x}_i)], \boldsymbol{v}\right\rangle\right| \end{aligned} \tag{158}$$

Define $G_i' = \left\langle \boldsymbol{v}, (\nabla\ell(\boldsymbol{W},\boldsymbol{x}_i) - \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\nabla\ell(\boldsymbol{W},\boldsymbol{x}_i)])\right\rangle$. Here $\boldsymbol{v}\in\mathbb{R}^d$. To compute $\nabla\ell(\boldsymbol{W},\boldsymbol{x}_i)$, we require the derivation in Property 5. Then we can have an upper bound of $\zeta(\boldsymbol{W})$ in (44).

$$\zeta(\boldsymbol{W}) = \begin{cases} \left|-\frac{1}{K}\frac{1}{H(\boldsymbol{W})}\phi'(\boldsymbol{w}_j^\top\boldsymbol{x})\right| \leq \frac{\phi(\boldsymbol{w}_j^\top\boldsymbol{x})(1-\phi(\boldsymbol{w}_j^\top\boldsymbol{x}))}{K\cdot\frac{1}{K}\phi(\boldsymbol{w}_j^\top\boldsymbol{x})} \leq 1, & y=1 \\ \left|\frac{1}{K}\frac{1}{1-H(\boldsymbol{W})}\phi'(\boldsymbol{w}_j^\top\boldsymbol{x})\right| \leq \frac{\phi(\boldsymbol{w}_j^\top\boldsymbol{x})(1-\phi(\boldsymbol{w}_j^\top\boldsymbol{x}))}{K\cdot\frac{1}{K}(1-\phi(\boldsymbol{w}_j^\top\boldsymbol{x}))} \leq 1, & y=0 \end{cases} \tag{159}$$

Then we have an upper bound of G_i'.

$$\begin{aligned} |G_i'| &= \left|\zeta_{j,l}\boldsymbol{v}^\top\boldsymbol{x} - \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\zeta\boldsymbol{v}^\top\boldsymbol{x}]\right| \\ &\leq |\boldsymbol{v}^\top\boldsymbol{x}| + \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[|\boldsymbol{v}^\top\boldsymbol{x}|] \end{aligned} \tag{160}$$

Following the idea of (141) and (142), and by $\boldsymbol{v}\in V_{\frac{1}{2}}$, we have

$$\mathbb{E}[|G_i'|^p] \leq \Big(\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2\Big)^{\frac{p}{2}} \tag{161}$$

$$\mathbb{E}[\exp(\theta G_i')] \leq 1 + |e\theta^2|\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2 \tag{162}$$

where (162) holds if $\theta \leq \frac{27}{4e}\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2}$. Following the derivation of (139) and (144) to (147), we have

$$\begin{aligned} &\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n}G_i'| \geq \frac{t}{6}) \\ \leq& 2\exp\Big(\max\Big\{-\frac{C_{14}nt^2}{144\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2}, -C_{15}n\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2\cdot t}\Big\}\Big) \end{aligned} \tag{163}$$

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

$$t \geq C_{13} \cdot \max\{\sqrt{\sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2} \sqrt{\frac{dK \log \frac{18r}{\epsilon} + \log \frac{4}{\delta}}{n}}, \frac{dK \log \frac{18r}{\epsilon} + \log \frac{4}{\delta}}{\sqrt{\sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2} \cdot n}\} \tag{164}$$

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

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} \Big[\sup_{\boldsymbol{W} \neq \boldsymbol{W}' \in \mathbb{B}(\boldsymbol{W}^*, r)} \frac{||\nabla \ell(\boldsymbol{W}, \boldsymbol{x}) - \nabla \ell(\boldsymbol{W}', \boldsymbol{x})||}{||\boldsymbol{W} - \boldsymbol{W}'||_F} \Big] \\
\leq &\mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} \Big[\sup_{\boldsymbol{W} \neq \boldsymbol{W}' \in \mathbb{B}(\boldsymbol{W}^*, r)} \frac{|\zeta(\boldsymbol{W}) - \zeta(\boldsymbol{W}')| \cdot ||\boldsymbol{x}||}{||\boldsymbol{W} - \boldsymbol{W}'||_F} \Big] \\
\leq &\mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} \Big[\sup_{\boldsymbol{W} \neq \boldsymbol{W}' \in \mathbb{B}(\boldsymbol{W}^*, r)} \frac{\max_{1 \leq j,l \leq K} \{|\xi_{j,l}(\boldsymbol{W}'')|\} \cdot ||\boldsymbol{x}||^2 \sqrt{K} ||\boldsymbol{W} - \boldsymbol{W}'||_F}{||\boldsymbol{W} - \boldsymbol{W}'||_F} \Big] \\
\leq &\mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^{L} \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)} \Big[\sup_{\boldsymbol{W} \neq \boldsymbol{W}' \in \mathbb{B}(\boldsymbol{W}^*, r)} \frac{C_9 \cdot ||\boldsymbol{x}||^2 \sqrt{K} ||\boldsymbol{W} - \boldsymbol{W}'||_F}{||\boldsymbol{W} - \boldsymbol{W}'||_F} \Big] \\
\leq &C_9 \cdot 3\sqrt{K} d \cdot \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2
\end{aligned} \tag{165}$$

The first inequality is by (44). The second inequality is by the Mean Value Theorem. The third step is by (102). The last inequality is by Property 4. Therefore, following the steps in part (2) of Lemma 8, we can conclude that C_t' holds if

$$t \geq 3C_9 \cdot \sqrt{K} d \cdot \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2 \cdot \epsilon \tag{166}$$

Moreover, from (151) in Lemma 8 we have that

$$t \geq \frac{18C_9 \cdot \sqrt{K} d \cdot \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2 \cdot \epsilon}{\delta} \tag{167}$$

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

$$\mathbb{P}(\sup_{\boldsymbol{W} \in \mathbb{B}(\boldsymbol{W}^*, r)} ||\nabla f_n(\boldsymbol{W}) - \nabla f(\boldsymbol{W})|| \geq C_{13} \cdot \sqrt{K \sum_{l=1}^{L} \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \sigma_l)^2} \sqrt{\frac{d \log n}{n}}) \leq d^{-10} \tag{168}$$

D.2 Proof of Lemma 2

Following the proof of Theorem 2 in [Fu et al. (2020)], first we have the Taylor's expansion of $f_n(\widehat{\boldsymbol{W}}_n)$

$$\begin{aligned}
f_n(\widehat{\boldsymbol{W}}_n) = &f_n(\boldsymbol{W}^*) + \Big\langle \nabla f_n(\boldsymbol{W}^*), \text{vec}(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*) \Big\rangle \\
&+ \frac{1}{2} \text{vec}(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*) \nabla^2 f_n(\boldsymbol{W}') \text{vec}(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*)
\end{aligned} \tag{169}$$

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

$$\frac{1}{2} \text{vec}(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*) \nabla^2 f_n(\boldsymbol{W}') \text{vec}(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*) \leq \Big| \nabla f_n(\boldsymbol{W}^*)^\top \text{vec}(\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*) \Big| \tag{170}$$

From Lemma 7 and Lemma 9, we have

$$\begin{aligned}&\frac{4}{K^2}\sum_{l=1}^{L}\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))||\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^*||_F^2\\ \leq&\frac{1}{2}\text{vec}(\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^*)\nabla^2 f_n(\boldsymbol{W}')\text{vec}(\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^*)\end{aligned}\tag{171}$$

and

$$\begin{aligned}&\left|\nabla f_n(\boldsymbol{W}^*)^\top\text{vec}(\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^*)\right|\\ \leq&\|\nabla f_n(\boldsymbol{W}^*)\|\cdot\|\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^*\|_F\\ \leq&(\|\nabla f_n(\boldsymbol{W}^*)-\nabla f(\boldsymbol{W}^*)\|+\|\nabla f(\boldsymbol{W}^*)\|)\cdot\|\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^*\|_F\\ \leq&O\Big(\sqrt{K\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2}\sqrt{\frac{d\log n}{n}}\Big)||\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^*||_F\end{aligned}\tag{172}$$

The second to last step of (172) comes from the triangle inequality and the last step follows from the fact $\nabla f(\boldsymbol{W}^*)=0$. Combining (170), (171) and (172), we have

$$||\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^*||_F\leq O\Big(\frac{K^{\frac{5}{2}}\sqrt{\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2}}{\sum_{l=1}^{L}\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))}\sqrt{\frac{d\log n}{n}}\Big)\tag{173}$$

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

$$\boldsymbol{W}_{t+1}-\widehat{\boldsymbol{W}}_n=\boldsymbol{W}_t-\eta_0\nabla f_n(\boldsymbol{W}_t)-(\widehat{\boldsymbol{W}}_n-\eta_0\nabla f_n(\widehat{\boldsymbol{W}}_n))=\Big(\boldsymbol{I}-\eta_0\int_0^1\nabla^2 f_n(\boldsymbol{W}(\gamma))\Big)(\boldsymbol{W}_t-\widehat{\boldsymbol{W}}_n)\tag{174}$$

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

$$H_{\min}\cdot\boldsymbol{I}\preceq\nabla^2 f_n(\boldsymbol{W}(\gamma))\leq H_{\max}\cdot\boldsymbol{I}\tag{175}$$

where $H_{\min}=\Omega\Big(\frac{1}{K^2}\sum_{l=1}^{L}\lambda_l\frac{\sigma_l^2}{\eta\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\sigma_l\delta_K(\boldsymbol{W}^*)},\sigma_l\delta_K(\boldsymbol{W}^*))\Big)$, $H_{\max}=C_4\cdot\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2$. Therefore,

$$||\boldsymbol{W}_{t+1}-\widehat{\boldsymbol{W}}_n||=||\boldsymbol{I}-\eta_0\int_0^1\nabla^2 f_n(\boldsymbol{W}(\gamma))||\cdot||\boldsymbol{W}_t-\widehat{\boldsymbol{W}}_n||\leq(1-\eta_0 H_{\min})||\boldsymbol{W}_t-\widehat{\boldsymbol{W}}_n||_F\tag{176}$$

By setting $\eta_0=\frac{1}{H_{\max}}=\Omega\Big(\big(C_4\sum_{l=1}^{L}\lambda_l(\|\boldsymbol{\mu}_l\|_\infty+\sigma_l)^2\big)^{-2}\Big)$, we obtain

$$||\widehat{\boldsymbol{W}}_{t+1}-\widehat{\boldsymbol{W}}_n||_F\leq(1-\frac{H_{\min}}{H_{\max}})||\boldsymbol{W}_t-\widehat{\boldsymbol{W}}_n||_F\tag{177}$$

Therefore, Algorithm 1 converges linearly to the local minimizer.

E Proof of Lemma 3

We need Lemma 10 to Lemma 14, which are stated in Section E.1, for the proof of Lemma 3. Section E.2 summarizes the proof of Lemma 3. The proofs of Lemma 10 to Lemma 12 are provided in Section E.3 to Section E.5. Lemma 13 and Lemma 14 are cited from (Zhong et al., 2017b). Although (Zhong et al., 2017b) 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.

E.1 USEFUL LEMMAS IN THE PROOF

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

$$||\boldsymbol{P}_2-\widehat{\boldsymbol{P}}_2||\lesssim\sqrt{\frac{d\log n}{n}}\cdot\delta_1^2\cdot\tau^6\sqrt{D_2(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})D_4(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})} \tag{178}$$

Lemma 11 *Let $\boldsymbol{U}\in\mathbb{E}^{d\times K}$ be the orthogonal column span of $\boldsymbol{W}^*$. Let $\boldsymbol{\alpha}$ be a fixed unit vector and $\widehat{\boldsymbol{U}}\in\mathbb{R}^{d\times K}$ denote an orthogonal matrix satisfying $||\boldsymbol{U}\boldsymbol{U}^\top-\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top||\leq\frac{1}{4}$. Define $\boldsymbol{R}_3=\boldsymbol{M}_3(\widehat{\boldsymbol{U}},\widehat{\boldsymbol{U}},\widehat{\boldsymbol{U}})$, where $\boldsymbol{M}_3$ is defined in Definition 1. Let $\widehat{\boldsymbol{R}}_3$ be the empirical version of $\boldsymbol{R}_3$ using data set S, where each sample of S is i.i.d. sampled from the mixed Gaussian distribution $\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I})$. Then with probability at least $1-n^{-\Omega(\delta^4)}$, we have*

$$||\widehat{\boldsymbol{R}}_3-\boldsymbol{R}_3||\lesssim\delta_1^2\cdot\left(\tau^6\sqrt{D_6(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})}\right)\cdot\sqrt{\frac{\log n}{n}} \tag{179}$$

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

$$||\widehat{\boldsymbol{M}}_1-\boldsymbol{M}_1||\lesssim\left(\tau^2\sqrt{D_2(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})}\right)\cdot\sqrt{\frac{d\log n}{n}} \tag{180}$$

Lemma 13 *((Zhong et al., 2017b), Lemma E.6) Let $\boldsymbol{P}_2$ be defined in Definition 1 and $\widehat{\boldsymbol{P}}_2$ be its empirical version. Let $\boldsymbol{U}\in\mathbb{R}^{d\times K}$ be the column span of $\boldsymbol{W}^*$. Assume $||\boldsymbol{P}_2-\widehat{\boldsymbol{P}}_2||\leq\frac{\delta_K(\boldsymbol{P}_2)}{10}$. Then after $T=O(\log(\frac{1}{\epsilon}))$ iterations, the output of the Tensor Initialization Method 3, $\widehat{\boldsymbol{U}}$ will satisfy*

$$||\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top-\boldsymbol{U}\boldsymbol{U}^\top||\lesssim\frac{||\widehat{\boldsymbol{P}}_2-\boldsymbol{P}_2||}{\delta_K(\boldsymbol{P}_2)}+\epsilon \tag{181}$$

which implies

$$||(\boldsymbol{I}-\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top)\boldsymbol{w}_i^*||\lesssim(\frac{||\boldsymbol{P}_2-\widehat{\boldsymbol{P}}_2||}{\delta_K(\boldsymbol{P}_2)}+\epsilon)||\boldsymbol{w}_i^*|| \tag{182}$$

Lemma 14 *((Zhong et al., 2017b), Lemma E.13) Let $\boldsymbol{U}\in\mathbb{R}^{d\times K}$ be the orthogonal column span of $\boldsymbol{W}^*$. Let $\widehat{\boldsymbol{U}}\in\mathbb{R}^{d\times K}$ be an orthogonal matrix such that $||\boldsymbol{U}\boldsymbol{U}^\top-\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top||\lesssim\gamma_1\lesssim\frac{1}{\kappa^2\sqrt{K}}$. For each $i\in[K]$, let $\widehat{\boldsymbol{v}}_i$ denote the vector satisfying $||\widehat{\boldsymbol{v}}_i-\widehat{\boldsymbol{U}}^\top\bar{\boldsymbol{w}}_i{}^*||\leq\gamma_2\lesssim\frac{1}{\kappa^2\sqrt{K}}$. Let $\boldsymbol{M}_1$ be defined in Lemma 12 and $\widehat{\boldsymbol{M}}_1$ be its empirical version. If $||\boldsymbol{M}_1-\widehat{\boldsymbol{M}}_1||\leq\gamma_3||\boldsymbol{M}_1||\lesssim\frac{1}{4}||\boldsymbol{M}_1||$, then we have*

$$\left|||\boldsymbol{w}_i^*||-\widehat{\alpha}_i\right|\leq(\kappa^4K^{\frac{3}{2}}(\gamma_1+\gamma_2)+\kappa^2K^{\frac{1}{2}}\gamma_3)||\boldsymbol{w}_i^*|| \tag{183}$$

E.2 PROOF OF LEMMA 3

$$\begin{aligned}
||\boldsymbol{w}_j^*-\widehat{\alpha}_j\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j|| &= \left|\left|\boldsymbol{w}_j^*-||\boldsymbol{w}_j^*||\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j+||\boldsymbol{w}_j^*||\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j-\widehat{\alpha}_j\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j\right|\right| \\
&\leq \left|\left|\boldsymbol{w}_j^*-||\boldsymbol{w}_j^*||\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j\right|\right|+\left|\left||\boldsymbol{w}_j^*||\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j-\widehat{\alpha}_j\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j\right|\right| \\
&\leq ||\boldsymbol{w}_j^*||\left|\left|\bar{\boldsymbol{w}}_j{}^*-\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j\right|\right|+\left|\left||\boldsymbol{w}_j^*||-\widehat{\alpha}_j\right|\right||\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j|| \\
&\leq ||\boldsymbol{w}_j^*||\left|\left|\bar{\boldsymbol{w}}_j{}^*-\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top\bar{\boldsymbol{w}}_j^*+\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top\bar{\boldsymbol{w}}_j{}^*-\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j\right|\right|+\left|\left||\boldsymbol{w}_j^*||-\widehat{\alpha}_j\right|\right||\widehat{\boldsymbol{U}}\widehat{\boldsymbol{v}}_j|| \\
&\leq \delta_1(\boldsymbol{W}^*)\left(\left|\left|\bar{\boldsymbol{w}}_j{}^*-\widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top\bar{\boldsymbol{w}}_j{}^*\right|\right|+\left|\left|\widehat{\boldsymbol{U}}^\top\bar{\boldsymbol{w}}_j{}^*-\widehat{\boldsymbol{v}}_j\right|\right|\right)+\left|\left||\boldsymbol{w}_j^*||-\widehat{\alpha}_j\right|\right|
\end{aligned} \tag{184}$$

By Lemma 10, Lemma 13 and $\delta_K(\boldsymbol{P}_2) \lesssim \delta_K^2$, we have

$$\begin{aligned}\Big|\Big|\bar{\boldsymbol{w}}_j{}^* - \widehat{\boldsymbol{U}}\widehat{\boldsymbol{U}}^\top \bar{\boldsymbol{w}}_j{}^*\Big|\Big| &\lesssim \frac{||\boldsymbol{P}_2 - \widehat{\boldsymbol{P}}_2||}{\delta_K(\boldsymbol{P}_2)} \lesssim \sqrt{\frac{d\log n}{n}} \cdot \frac{\delta_1^2}{\delta_K^2} \cdot \tau^6 \sqrt{D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})} \\ &= \sqrt{\frac{d\log n}{n}} \cdot \kappa^2 \cdot \tau^6 \sqrt{D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\end{aligned} \tag{185}$$

Moreover, we have

$$\Big|\Big|\widehat{\boldsymbol{U}}^\top \bar{\boldsymbol{w}}_j{}^* - \widehat{\boldsymbol{v}}_j\Big|\Big| \leq \frac{K^{\frac{3}{2}}}{\delta_K^2(\boldsymbol{W}^*)} ||\boldsymbol{R}_3 - \widehat{\boldsymbol{R}}_3|| \lesssim \kappa^2 \cdot \big(\tau^6 \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\big) \cdot \sqrt{\frac{K^3 \log n}{n}} \tag{186}$$

in which the first step is by Theorem 3 in [Kuleshov et al. (2015)] and the second step is by Lemma 11. By Lemma 14, we have

$$\Big|\Big|\,||\boldsymbol{w}_j^*|| - \widehat{\alpha}_j\Big|\Big| \leq (\kappa^4 K^{\frac{3}{2}}(\gamma_1 + \gamma_2) + \kappa^2 K^{\frac{1}{2}} \gamma_3)||\boldsymbol{W}^*|| \tag{187}$$

Therefore, taking the union bound of failure probabilities in Lemmas 10, 11 and 12 and by $D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) \leq D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})$ from Property 6, we have that if the sample size $n \geq \kappa^8 K^4 \tau^{12} D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) \cdot d \log^2 d$, then the output $\boldsymbol{W}_0 \in \mathbb{R}^{d \times K}$ satisfies

$$||\boldsymbol{W}_0 - \boldsymbol{W}^*|| \lesssim \kappa^6 K^3 \cdot \tau^6 \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})} \sqrt{\frac{d\log n}{n}} ||\boldsymbol{W}^*|| \tag{188}$$

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

E.3 Proof of Lemma 10

From Assumption 1, if the Gaussian Mixture Model is a symmetric probability distribution defined in (8), then $\boldsymbol{P}_2 = \boldsymbol{M}_3(\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\alpha})$. Therefore, by Definition 1, we have

$$\begin{aligned}||\widehat{\boldsymbol{M}}_3(\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\alpha}) - \boldsymbol{M}_3(\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\alpha})|| = \Big|\Big|\frac{1}{n}\sum_{i=1}^n \Big[y_i \cdot p(\boldsymbol{x})^{-1} \sum_{l=1}^L \lambda_l (2\pi\sigma_l)^{-\frac{d}{2}} \exp(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_l||^2}{2\sigma_l^2}) \\ \Big((\frac{\boldsymbol{x} - \boldsymbol{\mu}_l}{\sigma_l^2})^{\otimes 3} - (\frac{\boldsymbol{x} - \boldsymbol{\mu}_l}{\sigma_l^2}) \widetilde{\otimes} \sigma_l^{-2} \boldsymbol{I}\Big)\Big](\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\alpha}) \\ -\mathbb{E}\Big[y \cdot p(\boldsymbol{x})^{-1} \sum_{l=1}^L \lambda_l (2\pi\sigma_l)^{-\frac{d}{2}} \exp(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_l||^2}{2\sigma_l^2}) \\ \Big((\frac{\boldsymbol{x} - \boldsymbol{\mu}_l}{\sigma_l^2})^{\otimes 3} - (\frac{\boldsymbol{x} - \boldsymbol{\mu}_l}{\sigma_l^2}) \widetilde{\otimes} \sigma_l^{-2} \boldsymbol{I}\Big)\Big](\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\alpha})\Big|\Big|\end{aligned} \tag{189}$$

Following (Zhong et al., 2017b), $\widetilde{\otimes}$ is defined such that for any $\boldsymbol{v} \in \mathbb{R}^{d_1}$ and $\boldsymbol{Z} \in \mathbb{R}^{d_1 \times d_2}$,

$$\boldsymbol{v} \widetilde{\otimes} \boldsymbol{Z} = \sum_{i=1}^{d_2} (\boldsymbol{v} \otimes \boldsymbol{z}_i \otimes \boldsymbol{z}_i + \boldsymbol{z}_i \otimes \boldsymbol{v} \otimes \boldsymbol{z}_i + \boldsymbol{z}_i \otimes \boldsymbol{z}_i \otimes \boldsymbol{v}), \tag{190}$$

where $\boldsymbol{z}_i$ is the i-th column of $\boldsymbol{Z}$. By Definition 1, we have

$$\begin{aligned}&\Big|\Big|\Big[y \cdot p(\boldsymbol{x})^{-1} \sum_{l=1}^L \lambda_l (2\pi\sigma_l)^{-\frac{d}{2}} \exp(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_l||^2}{2\sigma_l^2}) \Big((\frac{\boldsymbol{x} - \boldsymbol{\mu}_l}{\sigma_l^2})^{\otimes 3} - (\frac{\boldsymbol{x} - \boldsymbol{\mu}_l}{\sigma_l^2}) \widetilde{\otimes} \sigma_l^{-2} \boldsymbol{I}\Big)\Big](\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{\alpha})\Big|\Big| \\ &\lesssim \Big|\Big|\frac{\sum_{l=1}^L \lambda_l (2\pi\sigma_l^2)^{-\frac{d}{2}} \exp(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_l||^2}{2\sigma_l^2}) \cdot (\frac{\boldsymbol{x} - \boldsymbol{\mu}_l}{\sigma_l^2})^{\otimes 2} (\boldsymbol{\alpha}^\top (\frac{\boldsymbol{x} - \boldsymbol{\mu}_l}{\sigma_l^2}))}{\sum_{l=1}^L \lambda_l (2\pi\sigma_l^2)^{-\frac{d}{2}} \exp(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_l||^2}{2\sigma_l^2})}\Big|\Big| \\ &\lesssim ||\sigma_{\min}^{-6}(\boldsymbol{x}_i^\top \boldsymbol{\alpha}) \boldsymbol{x}_i \boldsymbol{x}_i^\top||\end{aligned} \tag{191}$$

The first step of (191) is because $(\frac{\boldsymbol{x}-\boldsymbol{\mu}_l}{\sigma_l^2})^{\otimes 2}(\boldsymbol{\alpha}^\top(\frac{\boldsymbol{x}-\boldsymbol{\mu}_l}{\sigma_l^2}))$ 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 $(\frac{\boldsymbol{x}-\boldsymbol{\mu}_l}{\sigma_l^2})^{\otimes 2}(\boldsymbol{\alpha}^\top(\frac{\boldsymbol{x}-\boldsymbol{\mu}_l}{\sigma_l^2}))$ and $(\boldsymbol{x}_i^\top\boldsymbol{\alpha})\boldsymbol{x}_i\boldsymbol{x}_i^\top$ is its dominant term. Define $S_m(\boldsymbol{x}) = (-1)^m \frac{\nabla_{\boldsymbol{x}}^m p(\boldsymbol{x})}{p(\boldsymbol{x})}$, where $p(\boldsymbol{x})$ is the probability density function of the random variable $\boldsymbol{x}$. From Definition 1, we can verify that

$$\boldsymbol{M}_j = \mathbb{E}[y \cdot S_m(\boldsymbol{x})] \quad j \in \{1,2,3\} \tag{192}$$

Then define $Gp_i = \big\langle \boldsymbol{v}, ([y_i \cdot S_3(\boldsymbol{x}_i)](\boldsymbol{I}_d, \boldsymbol{I}_d, \boldsymbol{\alpha}) - \mathbb{E}\big[[y_i \cdot S_3(\boldsymbol{x}_i)](\boldsymbol{I}_d, \boldsymbol{I}_d, \boldsymbol{\alpha})\big]\boldsymbol{v})\big\rangle$, where $||\boldsymbol{v}|| = 1$, then $\mathbb{E}[Gp_i] = 0$. Similar to the proof of (140), (141) and (142) in Lemma 8, we have

$$|Gp_i|^p \lesssim \big|\sigma_{\min}^{-6}(\boldsymbol{x}_i^\top\boldsymbol{\alpha})(\boldsymbol{x}_i^\top\boldsymbol{v})^2 + \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2\boldsymbol{I}_d)}[\sigma_{\min}^{-6}(\boldsymbol{x}_i^\top\boldsymbol{\alpha})(\boldsymbol{x}_i^\top\boldsymbol{v})^2]\big|^p \tag{193}$$

$$\begin{aligned}\mathbb{E}[|Gp_i|^p] &\lesssim \big(\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2\boldsymbol{I}_d)}[\sigma_{\min}^{-6}(\boldsymbol{x}_i^\top\boldsymbol{\alpha})(\boldsymbol{x}_i^\top\boldsymbol{v})^2]\big)^p \\ &\leq \sigma_{\min}^{-6p}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2\boldsymbol{I}_d)}[(\boldsymbol{x}^\top\boldsymbol{\alpha})^2]^{\frac{p}{2}}\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2\boldsymbol{I}_d)}[(\boldsymbol{x}^\top\boldsymbol{v})^4]^{\frac{p}{2}} \\ &\leq \tau^{6p}\sqrt{D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}^p\end{aligned} \tag{194}$$

$$\begin{aligned}\mathbb{E}[\exp(\theta Gp_i)] &\lesssim 1 + \sum_{p=2}^{\infty}\frac{\theta^p\mathbb{E}[|Gp_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty}\frac{|e\theta|^p\tau^{6p}(D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}))^{\frac{p}{2}}}{p^p} \\ &\lesssim 1 + \theta^2\tau^{12}D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})\end{aligned} \tag{195}$$

Hence, similar to the derivation of (144), we have

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n Gp_i \geq t\Big) \leq \exp\Big(-n\theta t + C_{16}n\theta^2\big(\tau^6\sqrt{D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\big)^2\Big) \tag{196}$$

for some constant $C_{16} > 0$. Let $\theta = \frac{t}{2C_{16}\left(\tau^6\sqrt{D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\right)^2}$ and $t = \delta_1^2 \cdot \big(\tau^6\sqrt{D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\big) \cdot \sqrt{\frac{d\log n}{n}}$, then we have

$$||\widehat{\boldsymbol{M}}_3(\boldsymbol{I}_d, \boldsymbol{I}_d, \boldsymbol{\alpha}) - \boldsymbol{M}_3(\boldsymbol{I}_d, \boldsymbol{I}_d, \boldsymbol{\alpha})|| \leq \delta_1^2 \cdot \big(\tau^6\sqrt{D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\big) \cdot \sqrt{\frac{d\log n}{n}} \tag{197}$$

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

If the Gaussian Mixture Model is not a symmetric distribution which is defined in (8), then $\boldsymbol{P}_2 = \boldsymbol{M}_2$. We would have a similar result as follows.

$$||\widehat{\boldsymbol{M}}_2 - \boldsymbol{M}_2|| = \Big|\Big|\frac{1}{n}\sum_{i=1}^n[y_i \cdot S_2(\boldsymbol{x})] - \mathbb{E}[y \cdot S_2(\boldsymbol{x})]\Big|\Big| \tag{198}$$

$$||y_i \cdot S_2(\boldsymbol{x}_i)|| \lesssim ||\sigma_{\min}^{-4}\frac{1}{K}\sum_{j=1}^K \phi(\boldsymbol{w}_j^{*\top}\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^\top|| \tag{199}$$

Then define $Gp_i' = \big\langle \boldsymbol{v}, ([y_i \cdot S_2(\boldsymbol{x}_i)] - \mathbb{E}\big[y_i \cdot S_2(\boldsymbol{x}_i)\big]\boldsymbol{v})\big\rangle$, where $||\boldsymbol{v}|| = 1$, then $\mathbb{E}[Gp_i'] = 0$. Similar to the proof of (140), (141) and (142) in Lemma 8, we have

$$|Gp_i'|^p \lesssim \big|\sigma_{\min}^{-4}(\boldsymbol{x}_i^\top\boldsymbol{v})^2 + \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2\boldsymbol{I}_d)}[\sigma_{\min}^{-4}(\boldsymbol{x}_i^\top\boldsymbol{v})^2]\big|^p \tag{200}$$

$$\mathbb{E}[|Gp_i'|^p] \lesssim \big(\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2\boldsymbol{I}_d)}[\sigma_{\min}^{-4}(\boldsymbol{x}_i^\top\boldsymbol{v})^2]\big)^p \leq \tau^{4p}D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})^p \tag{201}$$

$$\begin{aligned}\mathbb{E}[\exp(\theta Gp_i')] &\lesssim 1 + \sum_{p=2}^{\infty}\frac{\theta^p\mathbb{E}[|Gp_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty}\frac{|e\theta|^p\tau^{4p}D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})^p}{p^p} \\ &\lesssim 1 + \theta^2\tau^8 D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})^2\end{aligned} \tag{202}$$

Hence, similar to the derivation of (144), we have

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n Gp_i \geq t\Big) \leq \exp\Big(-n\theta t + C_{17}n\theta^2\big(\tau^4 D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})\big)^2\Big) \tag{203}$$

for some constant $C_{17} > 0$. Let $\theta = \frac{t}{2C_{17}\left(\tau^4 D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})\right)^2}$ and $t = \delta_1^2 \cdot \left(\tau^4 D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})\right) \cdot \sqrt{\frac{d \log n}{n}}$, then we have

$$||\widehat{\boldsymbol{M}}_2 - \boldsymbol{M}_2|| \lesssim \delta_1^2 \cdot \tau^4 D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) \cdot \sqrt{\frac{d \log n}{n}} \tag{204}$$

with probability at least $1 - 2n^{-\Omega(\delta_1^4 d)}$.
To sum up, from (197) and (204) we have

$$\begin{aligned} ||\boldsymbol{P}_2 - \widehat{\boldsymbol{P}}_2|| &\lesssim \sqrt{\frac{d \log n}{n}} \cdot \delta_1^2 \cdot \max\{\tau^4 D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})), \tau^6 \sqrt{D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\} \\ &\lesssim \sqrt{\frac{d \log n}{n}} \cdot \delta_1^2 \cdot \tau^6 \sqrt{D_2(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma}) D_4(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})} \end{aligned} \tag{205}$$

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

E.4 Proof of Lemma 11

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

$$\boldsymbol{T}_i(\boldsymbol{x}) = [\phi({\boldsymbol{w}_i^*}^\top \boldsymbol{x}) \cdot S_3(\boldsymbol{x})](\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{U}}, \widehat{\boldsymbol{U}}) \tag{206}$$

We flatten $\boldsymbol{T}_i(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^{K \times K \times K}$ along the first dimension to obtain function $\boldsymbol{B}_i(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^{K \times K^2}$. Similar to the derivation of the last step of Lemma E.8 in (Zhong et al., 2017b), we can obtain $\|\boldsymbol{T}_i(\boldsymbol{x})\| \leq \|\boldsymbol{B}_i(\boldsymbol{x})\|$. By (189), we have

$$||\boldsymbol{B}_i(\boldsymbol{x})|| \lesssim \sigma_{\min}^{-6} \frac{1}{K} \sum_{j=1}^K \phi({\boldsymbol{w}_j^*}^\top \boldsymbol{x}_i)(\widehat{\boldsymbol{U}}^\top \boldsymbol{x})^3 \tag{207}$$

Define $Gr_i = \langle \boldsymbol{v}, \boldsymbol{B}_i(\boldsymbol{x}_i)) - \mathbb{E}[\boldsymbol{B}_i(\boldsymbol{x}_i)]\boldsymbol{v})\rangle$, where $||\boldsymbol{v}|| = 1$, so $\mathbb{E}[Gr_i] = 0$. Similar to the proof of (140), (141) and (142) in Lemma 8, we have

$$|Gr_i|^p \lesssim \left|\sigma_{\min}^{-6}(\boldsymbol{v}^\top \widehat{\boldsymbol{U}}^\top \boldsymbol{x})^3 + \mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)}[\sigma_{\min}^{-6}(\boldsymbol{v}^\top \widehat{\boldsymbol{U}}^\top \boldsymbol{x})^3]\right|^p \tag{208}$$

$$\mathbb{E}[|Gr_i|^p] \lesssim \left(\mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \boldsymbol{I}_d)}[\sigma_{\min}^{-6}(\boldsymbol{v}^\top \widehat{\boldsymbol{U}}^\top \boldsymbol{x})^3]\right)^p \lesssim \tau^{6p} \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}^p \tag{209}$$

$$\begin{aligned} \mathbb{E}[\exp(\theta Gr_i)] &\lesssim 1 + \sum_{p=2}^\infty \frac{\theta^p \mathbb{E}[|Gr_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^\infty \frac{|e\theta|^p \tau^{6p} D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})^{\frac{p}{2}}}{p^p} \\ &\leq 1 + \theta^2 (\tau^{12} \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})})^2 \end{aligned} \tag{210}$$

Hence, similar to the derivation of (144), we have

$$\mathbb{P}\Big(\frac{1}{n} \sum_{i=1}^n Gr_i \geq t\Big) \leq \exp\Big(-n\theta t + C_{18}\theta^2 \big(\tau^6 \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\big)^2\Big) \tag{211}$$

for some constant $C_{18} > 0$. Let $\theta = \frac{t}{C_{18}\left(\tau^6 \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\right)^2}$ and $t = \delta_1^2 \cdot \left(\tau^6 \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\right) \cdot \sqrt{\frac{\log n}{n}}$, then we have

$$||\widehat{\boldsymbol{R}}_3 - \boldsymbol{R}_3|| \lesssim \delta_1^2 \cdot \left(\tau^6 \sqrt{D_6(\boldsymbol{\lambda}, \boldsymbol{M}, \boldsymbol{\sigma})}\right) \cdot \sqrt{\frac{\log n}{n}} \tag{212}$$

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

E.5 PROOF OF LEMMA 12

From the Definition 1, we have

$$||\widehat{\boldsymbol{M}}_1 - \boldsymbol{M}_1|| = \Big|\Big|\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot S_1(\boldsymbol{x})] - \mathbb{E}[y \cdot S_1(\boldsymbol{x})]\Big|\Big|. \tag{213}$$

Based on Definition 1,

$$\Big|\Big|[y_i \cdot S_1(\boldsymbol{x}_i)]\Big|\Big| \lesssim \Big|\Big|\frac{\sum_{l=1}^{L}\lambda_l(2\pi\sigma_l^2)^{-\frac{d}{2}}\exp(-\frac{||\boldsymbol{x}-\boldsymbol{\mu}_l||^2}{2\sigma_l^2})\cdot(\frac{\boldsymbol{x}-\boldsymbol{\mu}_l}{\sigma_l^2})}{\sum_{l=1}^{L}\lambda_l(2\pi\sigma_l^2)^{-\frac{d}{2}}\exp(-\frac{||\boldsymbol{x}-\boldsymbol{\mu}_l||^2}{2\sigma_l^2})}\Big|\Big| \lesssim \Big|\Big|\sigma_{\min}^{-2}\frac{1}{K}\sum_{j=1}^{K}\phi(\boldsymbol{w}_j^{*\top}\boldsymbol{x}_i)\boldsymbol{x}_i\Big|\Big| \tag{214}$$

Define $Gq_i = \big\langle \boldsymbol{v}, ([y_i \cdot S_1(\boldsymbol{x}_i)] - \mathbb{E}\big[[y_i \cdot S_1(\boldsymbol{x}_i)]\big]\boldsymbol{v})\big\rangle$, where $||\boldsymbol{v}|| = 1$, so $\mathbb{E}[Gq_i] = 0$. Similar to the proof of (140), (141) and (142) in Lemma 8, we have

$$|Gq_i|^p \lesssim \big|\sigma_{\min}^{-2}(\boldsymbol{x}_i^\top\boldsymbol{v}) + \mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\sigma_{\min}^{-2}(\boldsymbol{x}_i^\top\boldsymbol{v})]\big|^p \tag{215}$$

$$\mathbb{E}[|Gq_i|^p] \lesssim \big(\mathbb{E}_{\boldsymbol{x}\sim\sum_{l=1}^{L}\mathcal{N}(\boldsymbol{\mu}_l,\sigma_l^2\boldsymbol{I}_d)}[\sigma_{\min}^{-2}(\boldsymbol{x}_i^\top\boldsymbol{v})]\big)^p \leq \tau^{2p}\sqrt{D_2(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})}^p \tag{216}$$

$$\begin{aligned}\mathbb{E}[\exp(\theta Gq_i)] &\lesssim 1 + \sum_{p=2}^{\infty}\frac{\theta^p\mathbb{E}[|Gq_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty}\frac{|e\theta|^p\tau^{2p}D_2(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})^{\frac{p}{2}}}{p^p} \\ &\leq 1 + \theta^2(\tau^2\sqrt{D_2(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})})^2\end{aligned} \tag{217}$$

Hence, similar to the derivation of (144), we have

$$|\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}Gq_i \geq t\Big) \leq \exp\Big(-n\theta t + C_{19}\theta^2\big(\tau^2\sqrt{D_2(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})}\big)^2\Big) \tag{218}$$

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

$$||\widehat{\boldsymbol{M}}_1 - \boldsymbol{M}_1|| \lesssim \big(\tau^2\sqrt{D_2(\boldsymbol{\lambda},\boldsymbol{M},\boldsymbol{\sigma})}\big)\cdot\sqrt{\frac{d\log n}{n}} \tag{219}$$

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