This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3374593

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

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

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

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

I. Introduction

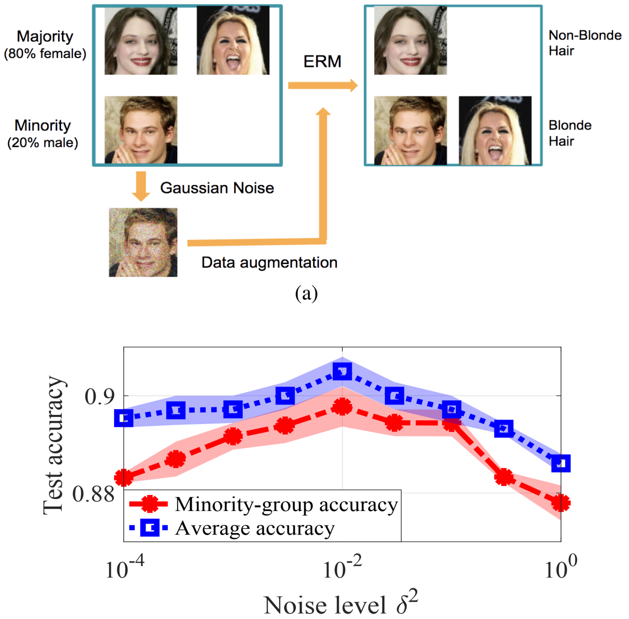
Training neural networks with empirical risk minimization (ERM) is a common practice to reduce the average loss of a machine learning task evaluated on a dataset. However, recent findings [1], [2], [3], [4], [5] have shown empirical evidence about a critical challenge of ERM, known as *group imbalance*, where a well-trained model that has high average accuracy may have significant errors on the minority group that infrequently appears in the data. Moreover, the group attributes that determine the majority and minority groups are usually hidden and unknown during the training. The training set can be augmented by data augmentation methods [6] with varying performance, such as cropping and rotation [7], noise injection [8], and generative adversarial network (GAN)-based methods [9].

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

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

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

The authors Hongkang Li and Dr. Meng Wang are with the Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute. Email: {lih35, wangm7}@rpi.edu. The author Dr. Shuai Zhang is with New Jersey Institute of Technology. Email: SZ457@njit.edu. The authors Yihua Zhang and Dr. Sijia Liu are with the Dept. of Computer Science and Engineering, Michigan State University. Email: {zhan1908, liusiji5}@msu.edu. The author Dr. Pin-Yu Chen is with IBM Research. Email: pin-yu.chen@ibm.com.



(a)

(b)

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

rate, and the average and group-level test error. The training algorithm is the gradient descent following a tensor initialization and converges linearly. Our key results include

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

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

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

II. Background and Related Work

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

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

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

Theoretical characterization of learning performance from other input distributions for one-hidden-layer neural networks. [60] analyzes the training loss with a single Gaussian with an arbitrary co-variance. [61] quantifies the SGD evolution trained on the Gaussian mixture model. When the hidden layer only contains one neuron, [62] analyzes rotationally invariant distributions. With an infinite number of neurons and an infinite input dimension, [63] analyzes the generalization error based on the mean-field analysis for distributions like Gaussian Mixture with the same mean. [64] considers inputs with low-dimensional structures. No sample complexity is provided in all these works.

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

III. Problem Formulation and Algorithm

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

Data Model. Let $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$ denote the input feature and label, respectively. We consider an unbalanced dataset that consists of L ($L \geq 2$) groups of data, where the feature $\boldsymbol{x}$ in the group l ($l \in [L]$) is drawn from a multi-variate Gaussian distribution with mean $\boldsymbol{\mu}_l \in \mathbb{R}^d$, and covariance $\boldsymbol{\Sigma}_l \in \mathbb{R}^{d\times d}$. Specifically, $\boldsymbol{x}$ follows the Gaussian mixture model (GMM) [65], [66], [67], [68], denoted as $\boldsymbol{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$[1]. $\lambda_l \in (0,1)$ is the probability of sampling from distribution-l and represents the expected fraction of group-l data. $\sum_{l=1}^L \lambda_l = 1$. Group l is defined as a minority group if λ_l is less than $1/L$. We use $\Psi = \{\lambda_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \forall l\}$ to denote all parameters of the mixture model[2]. We consider binary classification with label y generated by a ground-truth neural network with unknown weights $\boldsymbol{W}^* = [\boldsymbol{w}_1^*, ..., \boldsymbol{w}_K^*] \in \mathbb{R}^{d\times K}$ and sigmoid activation[3]. function $\phi(x) = \frac{1}{1+\exp(-x)}$, where[4]

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

Learning model. Learning is performed over a neural network that has the same architecture as in (1), which is a one-hidden-layer fully connected neural network[5] with its weights denoted by $\boldsymbol{W} \in \mathbb{R}^{d\times K}$. Given n training samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ where $\boldsymbol{x}_i$ follows the GMM model, and y_i is from (1), we aim to find the model weights via solving the empirical risk minimization (ERM), where $f_n(\boldsymbol{W})$ is the empirical risk,

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

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

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

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

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

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

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

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

Training Algorithm. Our algorithm starts from an initialization $\boldsymbol{W}_0 \in \mathbb{R}^{d\times K}$ computed based on the tensor initialization method (Subroutine 1 in in Appendix) and then updates the iterates $\boldsymbol{W}_t$ using gradient descent with the step size[6] η_0. The computational complexity of tensor initialization is $O(Knd)$. The per-iteration complexity of the gradient step is $O(Knd)$. We defer the details of Algorithm 1 in Appendix.

Algorithm 1 Our ERM learning algorithm

```
Input:  Training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, the step size $\eta_0 = O\Big(\big(\sum_{l=1}^L \lambda_l(\|\tilde{\boldsymbol{\mu}}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\big)^{-1}\Big)$, the total number of iterations $T$
Initialization: $\boldsymbol{W}_0 \leftarrow$ Tensor initialization method via Subroutine 1
Gradient Descent: for $t = 0, 1, \cdots, T-1$
    $\begin{aligned}\boldsymbol{W}_{t+1} &= \boldsymbol{W}_t - \eta_0 \cdot \frac{1}{n}\sum_{i=1}^n (\nabla l(\boldsymbol{W}, \boldsymbol{x}_i, y_i) + \nu_i) \\ &= \boldsymbol{W}_t - \eta_0\Big(\nabla f_n(\boldsymbol{W}) + \frac{1}{n}\sum_{i=1}^n \nu_i\Big)\end{aligned} \quad (6)$
Output:  $\boldsymbol{W}_T$
```

IV. Main Theoretical Results

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

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

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

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

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

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

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

the same order[7]. Let $\delta_i(\boldsymbol{W}^*)$ denote the i-th largest singular value of $\boldsymbol{W}^*$. Let $\kappa = \frac{\delta_1(\boldsymbol{W}^*)}{\delta_K(\boldsymbol{W}^*)}$, and define $\eta = \prod_{i=1}^K (\delta_i(\boldsymbol{W}^*)/\delta_K(\boldsymbol{W}^*))$.

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \bar{f} \le & \mathcal{E}(\Psi) \cdot poly(\kappa, \eta, \tau, \delta_1(\boldsymbol{W}^*)) \\ & \cdot \Theta\Big(K^{\frac{5}{2}}(1+\xi) \cdot \sqrt{d \log n/n}\Big) \end{aligned} \tag{11}$$

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

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

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

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

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

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

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

A. Theoretical Insights

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

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

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

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

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

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

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

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

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

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

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

B. Proof Idea and Technical Novelty

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

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

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

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

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

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

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

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

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

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

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

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

Third, we investigate and provide the order of the bound

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

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

V. Numerical Experiments

A. Experiments on Synthetic datasets

We first verify the theoretical bounds in Theorem 1 on synthetic data. Each entry of $\boldsymbol{W}^* \in \mathbb{R}^{d\times K}$ is generated from $\mathcal{N}(0,1)$. The training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ is generated using the GMM model and (1). If not otherwise specified, $L = 2$, $d = 5$, and $K = 3$[9]. To reduce the computational time, we randomly initialize near $\boldsymbol{W}^*$ instead of computing the tensor initialization[10].

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

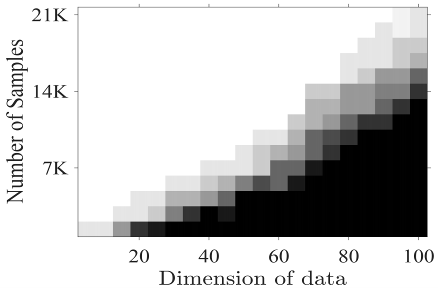


Fig. 2: The sample complexity when the feature dimension changes

We next study the impact on the sample complexity of the GMM model. In Figure 3 (a), $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}$, and let $\boldsymbol{\mu}_1 = \mu \cdot \boldsymbol{1}$, $\boldsymbol{\mu}_2 = -\boldsymbol{1}$. $\|\boldsymbol{\mu}_1\|$ varies from 0 to 5. Figure 3(a) shows that when the mean increases, the sample complexity increases. In Figure 3 (b), we fix $\boldsymbol{\mu}_1 = \boldsymbol{1}$, $\boldsymbol{\mu}_2 = -\boldsymbol{1}$, and let $\boldsymbol{\Sigma}_1 = \sigma^2\boldsymbol{I}$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{I}$. σ varies from 10^{-1} to 10^1. The sample complexity increases both when $\|\boldsymbol{\Sigma}_1\|$ increases and when $\|\boldsymbol{\Sigma}_1\|$ approaches zero. All results match predictions in Corollary 1.

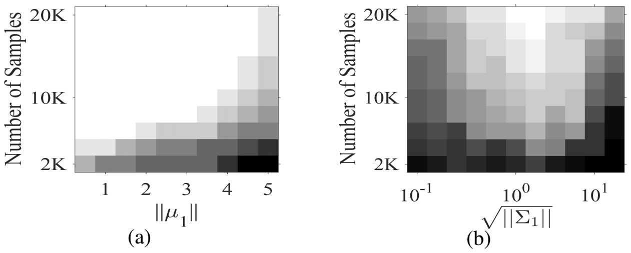


Fig. 3: The sample complexity (a) when one mean changes, (b) when one co-variance changes.

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

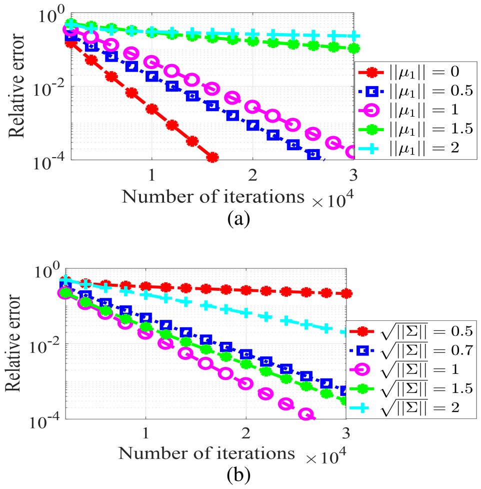
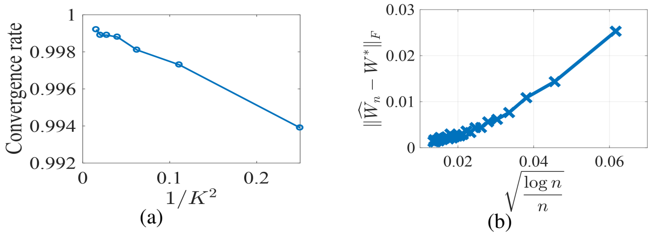


Fig. 4: (a) The convergence rate with different $\boldsymbol{\mu}_1$. (b) The convergence rate with different $\boldsymbol{\Sigma}$. (c) Convergence rate when the number of neurons K changes.

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

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

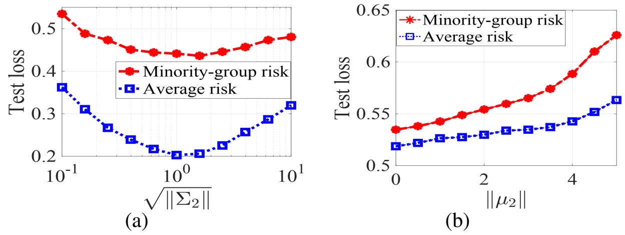
Average and group-level generalization performance. The distance between $\widehat{\boldsymbol{W}}_n$ returned by Algorithm 1 and $\boldsymbol{W}^*$ is measured by $||\widehat{\boldsymbol{W}}_n - \boldsymbol{W}^*||_F$. n ranges from 2×10^3 to 6×10^4. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = 9\boldsymbol{I}$, $\boldsymbol{\mu}_1 = \mathbf{1}$, $\boldsymbol{\mu}_2 = -\mathbf{1}$. Each point in Figure 5 (b) is averaged over 20 experiments of different $\boldsymbol{W}^*$ and training set. The error is indeed linear in $\sqrt{\log(n)/n}$, as predicted by (8).



(a) (b)

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

We evaluate the impact of one mean/co-variance of the minority group on the generalization. $n = 2 \times 10^4$. Let $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, $\boldsymbol{\mu}_1 = 2 \cdot \mathbf{1}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{I}$. First, we let $\boldsymbol{\mu}_2 = (\mu_2 - 2) \cdot \mathbf{1}$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{I}$. Figure 6 (b) shows that both the average risk and the group-2 risk increase as μ_2 increases, consistent with (P5). Then we set $\boldsymbol{\mu}_2 = -2 \cdot \mathbf{1}$, $\boldsymbol{\Sigma}_2 = \sigma_2^2 \cdot \boldsymbol{I}$. Figure 6 (a) indicates that both the average and the group-2 risk will first decrease and then increase as $\|\Sigma\|_2$ increases, consistent with (P3).



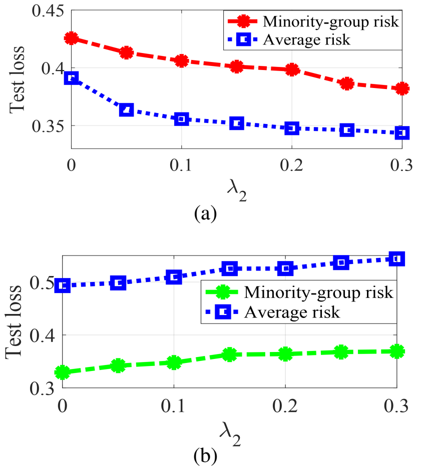
(a) (b)

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

Next, we study the impact of increasing the fraction of the minority group. $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$. Let group 2 be the minority group. In Figure 7 (a), $\boldsymbol{\Sigma}_1 = 10 \cdot \boldsymbol{I}$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{I}$, the minority group has a smaller level of co-variance. Then when λ_2 increases from 0 to 0.5, both the average and group-2 risk decease. In Figure 7 (b), $\boldsymbol{\Sigma}_1 = \boldsymbol{I}$ and $\boldsymbol{\Sigma}_2 = 10 \cdot \boldsymbol{I}$, and the minority group has a higher-level of co-variance. Then when λ_2 increases from 0 to 0.3, both the average and group-2 risk increase. As predicted by insight (P4), increasing λ_2 does not necessarily improve the generalization of group 2.

B. Image classification on dataset CelebA

We choose the attribute “blonde hair” as the binary classification label. ResNet 9 [73] is selected to be the learning model here because it was applied in many simple computer vision tasks [74], [75]. To study the impact of co-variance, we pick 4000 female (majority) and 1000 male (minority) images and implement Gaussian data augmentation to create additional 300 images for the male group. Specifically, we select 300 out of 1000 male images and add i.i.d. noise drawn from $\mathcal{N}(0, \delta^2)$ to every entry. The test set includes 500 male and 500 female images. Figure 1 shows that when δ^2 increases, i.e., when the co-variance of the minority group increases, both the minority-group and average test accuracy increase first and then decrease, coinciding with our insight (P3).



(a)

(b)

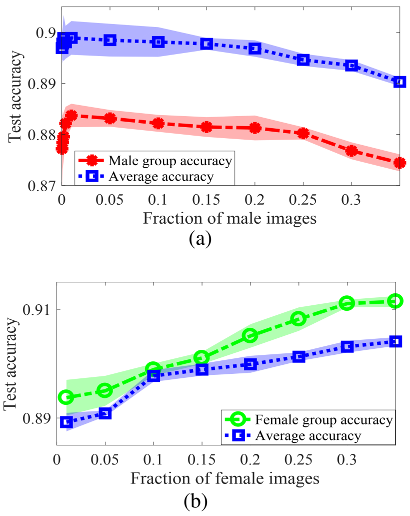
Fig. 7: The test loss (cross entropy loss) of synthetic data with different λ_2 values. (a) Group 2 has a smaller level of co-variance. (b) Group 2 has a larger level of co-variance.

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

VI. Conclusions and Future Directions

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

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



(a)

(b)

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

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

APPENDIX

A. Definitions

Definition 1. *(ρ-function). Let $\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{u},\boldsymbol{I}_d)\in\mathbb{R}^d$. 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 $\boldsymbol{z}$ and $\boldsymbol{u}$, respectively. Define $\rho(\boldsymbol{u},\sigma)$ as*

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

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

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

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

B. Proof of Lemma 1

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

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

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

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

$$\Omega\Big(\frac{1-2\epsilon_0}{K^2}\sum_{l=1}^L\lambda_l\frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2}\rho(\frac{{\boldsymbol{W}^*}^\top\boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}},\\ \delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})\Big)\cdot\boldsymbol{I}_{dK} \tag{16}$$

$$\preceq\nabla^2 f_n(\boldsymbol{W})\preceq C_2\sum_{l=1}^L\lambda_l(||\tilde{\boldsymbol{\mu}}_l||_\infty+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\cdot\boldsymbol{I}_{dK}$$

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

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

Lemma 4.

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

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

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

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

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

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

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

$$\frac{4(1-\epsilon_0)}{K^2}\sum_{l=1}^L\lambda_l\frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}},\delta_K(\boldsymbol{W}^*)$$
$$\cdot\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})\cdot\boldsymbol{I}_{dK}\preceq\nabla^2\bar{f}(\boldsymbol{W})\preceq\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\boldsymbol{\Sigma}_l^{\frac{1}{2}})^2\cdot\boldsymbol{I}_{dK} \tag{20}$$

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

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

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

We next show the proof of Lemma 1.

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

$$\nabla^2 f_n(\boldsymbol{W})\succeq\nabla^2\bar{f}(\boldsymbol{W})-||\nabla^2\bar{f}(\boldsymbol{W})-\nabla^2 f_n(\boldsymbol{W})||\cdot\boldsymbol{I}$$
$$\succeq\Omega\Big(\frac{(1-\epsilon_0)}{K^2}\sum_{l=1}^L\lambda_l\frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}},$$
$$\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})\Big)\cdot\boldsymbol{I}$$
$$-O\Big(C_6\cdot\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\sqrt{\frac{dK\log n}{n}}\Big)\cdot\boldsymbol{I} \tag{22}$$

As long as the sample complexity is set to satisfy

$$\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\cdot\sqrt{\frac{dK\log n}{n}}\leq\frac{\epsilon_0}{K^2}\sum_{l=1}^L\lambda_l\frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2}$$
$$\cdot\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}},\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})\cdot\boldsymbol{I} \tag{23}$$

i.e.,

$$n\gtrsim\epsilon_0^{-2}\cdot(\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2)^2$$
$$\cdot\Big(\sum_{l=1}^L\lambda_l\frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \tag{24}$$
$$\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})\cdot\boldsymbol{I}\Big)^{-2}dK^5\log^2 d$$

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

$$\nabla^2 f_n(\boldsymbol{W})\succeq\Omega\Big(\frac{1-2\epsilon_0}{K^2}\sum_{l=1}^L\lambda_l\frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2}$$
$$\cdot\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}},\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})\Big)\cdot\boldsymbol{I} \tag{25}$$

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

$$||\nabla^2 f_n(\boldsymbol{W})||\leq||\nabla^2\bar{f}(\boldsymbol{W})||+||\nabla^2 f_n(\boldsymbol{W})-\nabla^2\bar{f}(\boldsymbol{W})||$$
$$\lesssim\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2$$
$$+\sum_{1=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\sqrt{\frac{dK\log n}{n}}$$
$$\lesssim\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \tag{26}$$

Combining (25) and (26), we have

$$\Omega\Big(\frac{1-2\epsilon_0}{K^2}\sum_{l=1}^L\lambda_l\frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}},$$
$$\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})\Big)\cdot\boldsymbol{I}\preceq\nabla^2 f_n(\boldsymbol{W}) \tag{27}$$
$$\preceq\sum_{l=1}^L\lambda_l(||\tilde{\boldsymbol{\mu}}_l||_\infty+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\cdot\boldsymbol{I}$$

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

C. Proof of Lemma 2

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

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

$$||\widehat{\boldsymbol{W}}_n-\boldsymbol{W}^*\boldsymbol{P}||_F$$
$$\lesssim\frac{K^{\frac{5}{2}}\sqrt{\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}(1+\xi)\cdot\sqrt{d\log n/n}}{\sum_{l=1}^L\lambda_l\frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}},\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})} \tag{28}$$

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

$$||\boldsymbol{W}_t-\widehat{\boldsymbol{W}}_n||_F\leq||\boldsymbol{W}_0-\widehat{\boldsymbol{W}}_n||_F\cdot\Big(1-$$
$$\Omega(\frac{\sum_{l=1}^L\frac{\lambda_l\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2}\rho(\frac{\boldsymbol{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}},\delta_K(\boldsymbol{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}})}{K^2\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2})\Big)^t \tag{29}$$

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

1) A useful lemma:

Lemma 9. *If r is defined in (139) for $\epsilon_0\in(0,\frac{1}{4})$, then with probability at least $1-d^{-10}$, we have*[12]

$$\sup_{\boldsymbol{W}\in\mathbb{B}(\boldsymbol{W}^*\boldsymbol{P},r)}||\nabla\tilde{f}_n(\boldsymbol{W})-\nabla\tilde{f}(\boldsymbol{W})||$$
$$\lesssim\sqrt{K\sum_{l=1}^L\lambda_l(\|\boldsymbol{\mu}_l\|+\|\boldsymbol{\Sigma}_l\|)^2}\sqrt{\frac{d\log n}{n}}(1+\xi) \tag{30}$$

[12] $\nabla\tilde{f}_n(\boldsymbol{W})$ is defined as $\frac{1}{n}\sum_{i=1}^n(\nabla l(\boldsymbol{W},\boldsymbol{x}_i,y_i)+\nu_i)$ in algorithm 1

, where $\boldsymbol{P}$ is a permutation matrix.

We next show the proof of Lemma 2.

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

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

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

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

From Lemma 7 and Lemma 9, we have

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

and

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

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

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

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

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

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

$$H_{\min} \cdot \boldsymbol{I} \preceq \nabla^2 f_n(\boldsymbol{W}(\gamma)) \leq H_{\max} \cdot \boldsymbol{I} \tag{37}$$

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

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

By setting $\eta_0 = \frac{1}{H_{\max}} = O\Big(\frac{1}{\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|)^2}\Big)$, we obtain

$$||\widehat{\boldsymbol{W}}_{t+1} - \widehat{\boldsymbol{W}}_n||_F \leq (1 - \frac{H_{\min}}{H_{\max}})||\boldsymbol{W}_t - \widehat{\boldsymbol{W}}_n||_F + \frac{\eta_0}{n}\sum_{i=1}^n \|\nu_i\|_F \tag{39}$$

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

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

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

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

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

D. *Proof of Lemma 3*

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

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

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

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

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

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

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

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

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

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

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

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

Lemma 11. *Let $\boldsymbol{U}\in\mathbb{E}^{d\times K}$ 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 $\boldsymbol{R}_3=\boldsymbol{Q}_3(\widehat{\boldsymbol{U}},\widehat{\boldsymbol{U}},\widehat{\boldsymbol{U}})$, where $\boldsymbol{Q}_3$ is defined in Definition 3. 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 $\sum_{l=1}^{L}\lambda_l\mathcal{N}(\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)$. Then with a probability at least $1-n^{-\Omega(\delta^4(\boldsymbol{W}^*))}$, we have*

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

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

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

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

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

which implies

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

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

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

which implies

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

if the mixed Gaussian distribution is symmetric.

Lemma 14. *([26], 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 $\boldsymbol{Q}_1$ be defined in Lemma 12 and $\widehat{\boldsymbol{Q}}_1$ be its empirical version. If $||\boldsymbol{Q}_1-\widehat{\boldsymbol{Q}}_1||\leq\gamma_3||\boldsymbol{Q}_1||\lesssim\frac{1}{4}||\boldsymbol{Q}_1||$, then we have*

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

We next show the proof of Lemma 3.

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

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

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

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

if the mixed Gaussian distribution is not symmetric, and

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

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

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

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

$$\Big|\Big|\,||\boldsymbol{w}_j^*|| - \widehat{\alpha}_j\Big|\Big| \le (\kappa^4 K^{\frac{3}{2}}(\gamma_1+\gamma_2) + \kappa^2 K^{\frac{1}{2}}\gamma_3)||\boldsymbol{W}^*|| \quad (57)$$

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

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

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

E. Extension to Multi-Classification

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

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

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

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

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

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

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

One can observe from Figure 9 that when the noise level δ^2 increases, i.e., when the co-variance of the minority group increases, both the minority-group and average test accuracy increase first and then decrease, coinciding with our insight (P3). In Figure 10 (a) and (b), we can see opposite trends if we increase the fraction of the minority group in the training data, with the male being the minority or the female being the minority. Figures 9 and 10 are consistent with our findings in Figures 1 and 8, respectively.

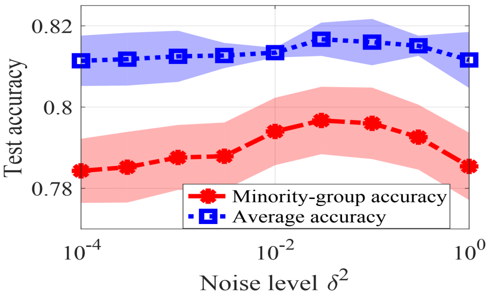
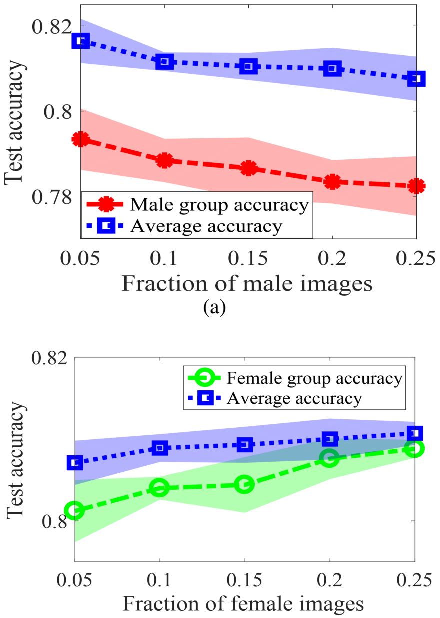


Fig. 9: Test accuracy against the augmented noise level for 4-classification.



(a)

(b)

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

F. Discussion about Gaussian Mixture Model (GMM)

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

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

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

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

G. Discussion about $\sigma_{\min}$ and τ

In this section, we show that the assumption that $\sigma_{\min}$ is not very close to zero, or equivalently, $\tau = \Theta(1)$, is mild. Even when the real data have singular values very close to zero, they can be approximated by low-rank data without hurting the performance by only keeping a few significant singular values and setting the small ones to zero. Thus, every practical dataset can be approximated by a dataset with $\tau = \Theta(1)$ while maintaining the same performance. We verify this by an experiment of binary classification on CelebA [31]. After training with a ResNet-9, the output feature of each testing image is 256-dimensional. One can find that the singular value of the covariance matrix of features is close to 0 except for the top singular values. The feature matrix reconstructed with top singular values can achieve comparable testing accuracy as using all singular values, as shown in Table II. One can observe that the feature matrix reconstruction with top-5 singular values, which is 2% of all the singular vectors, leads to a test accuracy already close to that using all singular vectors, and the performance gap is smaller than 4.5%. We can compute that $\tau = 4.6155 = \Theta(1)$ for the feature matrix reconstructed by top 5 singular values.

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

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

References

[1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[2] T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3428–3448.

[3] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, "An investigation of why overparameterization exacerbates spurious correlations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8346–8356.

[4] S. Sagawa*, P. W. Koh*, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=ryxGuJrFvS

[5] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, "Improving out-of-distribution robustness via selective augmentation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 407–25 437.

[6] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[8] F. J. Moreno-Barea, F. Strazzera, J. M. Jerez, D. Urda, and L. Franco, "Forward noise adjustment scheme for data augmentation," in *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2018, pp. 728–734.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," in *36th International Conference on Machine Learning, ICML 2019*. International Machine Learning Society (IMLS), 2019, pp. 477–502.

[11] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 242–252.

[12] Z. Allen-Zhu, Y. Li, and Y. Liang, "Learning and generalization in overparameterized neural networks, going beyond two layers," in *Advances in neural information processing systems*, 2019, pp. 6158–6169.

[13] Y. Cao and Q. Gu, "Generalization bounds of stochastic gradient descent for wide and deep neural networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 836–10 846.

[14] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in *Advances in Neural Information Processing Systems*, 2018, pp. 8157–8166.

[15] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in neural information processing systems*, 2018, pp. 8571–8580.

[16] H. Li, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Generalization guarantee of training graph convolutional networks with graph topology sampling," in *International Conference on Machine Learning*. PMLR, 2022, pp. 13 014–13 051.

[17] Z. Shi, J. Wei, and Y. Liang, "A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features," in *International Conference on Learning Representations*, 2021.

[18] S. Karp, E. Winston, Y. Li, and A. Singh, "Local signal adaptivity: Provable feature learning in neural networks beyond kernels," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 883–24 897, 2021.

[19] Z. Allen-Zhu and Y. Li, "Feature purification: How adversarial training performs robust deep learning," in *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2022, pp. 977–988.

[20] H. Li, M. Wang, S. Liu, and P.-Y. Chen, "A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity," in *The Eleventh International Conference on Learning Representations*, 2023.

[21] Z. Allen-Zhu and Y. Li, "Towards understanding ensemble, knowledge distillation and self-distillation in deep learning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=Uuf2q9TfXGA

[22] H. Li, M. Wang, T. Ma, S. Liu, Z. ZHANG, and P.-Y. Chen, "What improves the generalization of graph transformer? a theoretical dive into self-attention and positional encoding," in *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=BaxFC3z9R6

[23] Y. Chen, W. Huang, K. Zhou, Y. Bian, B. Han, and J. Cheng, "Understanding and improving feature learning for out-of-distribution generalization," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=eozEoAtjG8

[24] H. Li, M. Wang, S. Lu, H. Wan, X. Cui, and P.-Y. Chen, "Transformers as multi-task feature selectors: Generalization analysis of in-context learning," in *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=BMQ4i2RVbE

[25] Y. Deng, Y. Yang, B. Mirzasoleiman, and Q. Gu, "Robust learning with progressive data expansion against spurious correlation," *arXiv e-prints*, pp. arXiv–2306, 2023.

[26] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon, "Recovery guarantees for one-hidden-layer neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 4140–4149. [Online]. Available: https://arxiv.org/pdf/1706.03175.pdf

[27] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Fast learning of graph neural networks with guaranteed generalizability: One-hidden-layer case," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 268–11 277.

[28] H. Fu, Y. Chi, and Y. Liang, "Guaranteed recovery of one-hidden-layer neural networks via cross entropy," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3225–3235, 2020.

[29] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis," in *International Conference on Learning Representations*, 2021.

[30] S. Zhang, H. Li, M. Wang, M. Liu, P.-Y. Chen, S. Lu, S. Liu, K. Murugesan, and S. Chaudhury, "On the convergence and sample complexity analysis of deep q-networks with ϵ-greedy exploration," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=HWGWeaN76q

[31] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[33] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*. Springer, 1998, pp. 9–50.

[34] L. M. Koch, C. M. Schürch, A. Gretton, and P. Berens, "Hidden in plain sight: Subgroup shifts escape OOD detection," in *Medical Imaging with Deep Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=aZgiUNye2Cz

[35] J. Ma, J. Deng, and Q. Mei, "Subgroup generalization and fairness of graph neural networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[36] A. Biswas and S. Mukherjee, "Ensuring fairness under prior probability shifts," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 414–424.

[37] S. Giguere, B. Metevier, Y. Brun, P. S. Thomas, S. Niekum, and B. C. da Silva, "Fairness guarantees under demographic shift," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=wbPObLm6ueA

[38] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[39] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert systems with applications*, vol. 73, pp. 220–239, 2017.

[40] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[41] J. Byrd and Z. Lipton, "What is the effect of importance weighting in deep learning?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 872–881.

[42] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 219–226.

[43] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[44] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[45] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 680–697.

[46] S. S. Mullick, S. Datta, and S. Das, "Generative adversarial minority oversampling," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1695–1704.

[47] J. Kim, J. Jeong, and J. Shin, "M2m: Imbalanced classification via major-to-minor translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 896–13 905.

[48] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *European Conf. on Computer Vision (ECCV)*, 2020.

[49] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5212–5221.

[50] C. Fang, H. He, Q. Long, and W. J. Su, "Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training," *Proceedings of the National Academy of Sciences*, vol. 118, no. 43, p. e2103091118, 2021.

[51] L. Yang, H. Jiang, Q. Song, and J. Guo, "A survey on long-tailed visual recognition," *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1837–1872, 2022.

[52] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi, "The majority can help the minority: Context-rich minority oversampling for long-tailed classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6887–6896.

[53] S. S. Du, J. D. Lee, Y. Tian, A. Singh, and B. Poczos, "Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima," in *International Conference on Machine Learning*, 2018, pp. 1338–1347.

[54] R. Ge, J. D. Lee, and T. Ma, "Learning one-hidden-layer neural networks with landscape design," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=BkwHObbRZ

[55] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with relu activation," in *Advances in neural information processing systems*, 2017, pp. 597–607.

[56] I. Safran and O. Shamir, "Spurious local minima are common in two-layer relu neural networks," in *International Conference on Machine Learning*, 2018, pp. 4430–4438.

[57] X. Zhang, Y. Yu, L. Wang, and Q. Gu, "Learning one-hidden-layer relu networks via gradient descent," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1524–1534.

[58] S. Zhang, M. Wang, J. Xiong, S. Liu, and P.-Y. Chen, "Improved linear convergence of training cnns with generalizability guarantees: A one-hidden-layer case," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[59] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, "Why lottery ticket wins? a theoretical perspective of sample complexity on sparse neural networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

This article has been accepted for publication in IEEE Journal of Selected Topics in Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2024.3374593

[60] Y. Yoshida and M. Okada, "Data-dependence of plateau phenomenon in learning with neural network — statistical mechanical analysis," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019, pp. 1722–1730.

[61] F. Mignacco, F. Krzakala, P. Urbani, and L. Zdeborová, "Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9540–9550, 2020.

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

[63] S. Mei, A. Montanari, and P.-M. Nguyen, "A mean field view of the landscape of two-layer neural networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 33, pp. E7665–E7671, 2018.

[64] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, "When do neural networks outperform kernel methods?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 820–14 830, 2020.

[65] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[66] D. Hsu and S. M. Kakade, "Learning mixtures of spherical gaussians: moment methods and spectral decompositions," in *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, 2013, pp. 11–20.

[67] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of gaussians," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 93–102.

[68] O. Regev and A. Vijayaraghavan, "On learning mixtures of well-separated gaussians," in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2017, pp. 85–96.

[69] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM review*, vol. 26, no. 2, pp. 195–239, 1984.

[70] N. Ho and X. Nguyen, "Convergence rates of parameter estimation for some weakly identifiable finite mixtures," *Ann. Statist.*, vol. 44, no. 6, pp. 2726–2755, 12 2016. [Online]. Available: https://doi.org/10.1214/16-AOS1444

[71] R. Dwivedi, N. Ho, K. Khamaru, M. I. Jordan, M. J. Wainwright, and B. Yu, "Singularity, misspecification, and the convergence rate of em," *To appear, Annals of Statistics*, 2020.

[72] R. Dwivedi, N. Ho, K. Khamaru, M. Wainwright, M. Jordan, and B. Yu, "Sharp analysis of expectation-maximization for weakly identifiable models," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. Online: PMLR, 26–28 Aug 2020, pp. 1866–1876.

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[74] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer, "Shift: A zero flop, zero parameter alternative to spatial convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9127–9135.

[75] A. Dutta, E. H. Bergou, A. M. Abdelmoniem, C.-Y. Ho, A. N. Sahu, M. Canini, and P. Kalnis, "On the discrepancy between the theoretical analysis and practical implementations of compressed communication for distributed deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3817–3824.

[76] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[77] V. Kuleshov, A. Chaganty, and P. Liang, "Tensor factorization via matrix factorization," in *Artificial Intelligence and Statistics*, 2015, pp. 507–516.

[78] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5513–5522.

[79] D. Zoran and Y. Weiss, "Natural images, gaussian mixtures and dead leaves," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

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

[81] D. Zou, Y. Cao, D. Zhou, and Q. Gu, "Gradient descent optimizes over-parameterized deep relu networks," *Machine Learning*, vol. 109, no. 3, pp. 467–492, 2020.

[82] M. Janzamin, H. Sedghi, and A. Anandkumar, "Score function features for discriminative learning: Matrix and tensor framework," *arXiv preprint arXiv:1412.2863*, 2014.

[83] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for non-convex losses," *arXiv preprint arXiv:1607.06534*, 2016.