Distributed Sparse Multicategory Discriminant Analysis

Hengchao Chen
University of Toronto

Qiang Sun
University of Toronto

Abstract

This paper proposes a convex formulation for sparse multicategory linear discriminant analysis and then extend it to the distributed setting when data are stored across multiple sites. The key observation is that for the purpose of classification it suffices to recover the discriminant subspace which is invariant to orthogonal transformations. Theoretically, we establish statistical properties ensuring that the distributed sparse multicategory linear discriminant analysis performs as good as the centralized version after a few rounds of communications. Numerical studies lend strong support to our methodology and theory.

1 INTRODUCTION

Classification aims to assign data points to the correct classes. One popular multicategory classification method is the Fisher's linear discriminant analysis (LDA), which has appealing performances in many applications (Hand, 2006; Michie et al., 1994). Similar to the principal component analysis (PCA), a well-known unsupervised dimensionality reduction technique, LDA can also serve as an efficient supervised dimensionality reduction tool. However, when the number of features is larger than the number of observations, which prevails in modern datasets, LDA performs poorly due to the diverging spectra (Bickel and Levina, 2008) and the noise accumulation (Fan and Fan, 2008). In the moderately high dimensional setting, the naive Bayes, which assumes conditional independence between features, is shown to outperform the Fisher's rule (Bickel and Levina, 2004), while in the ultra-high dimensionsal setting, any classifier using all the predictors will be no better than random

guessing due to the noise accumulation, even if the true covariance matrix is an identity matrix (Fan and Fan, 2008). Moreover, as pointed out by Witten and Tibshirani (2011), the Fisher's LDA classifier is not interpretable when the discriminant vectors have no particular structure.

In recent years, many high-dimensional extensions of LDA have been proposed. For binary classification problems, the linear programming discriminant (Cai and Liu, 2011), the regularized optimal affine discriminant (Fan et al., 2012), and the direct sparse discriminant analysis (Mai et al., 2012) are three popular sparse linear discriminant analysis (SLDA) methods. All these three methods assume some sparsity assumptions on the discriminant directions. Another SLDA method is the thresholding linear discriminant analysis (Shao et al., 2011), which assumes sparsity conditions on the common covariance matrix and the mean difference vector. It is non-trivial to extend these binary SLDA methods to the multicategory case. For multicategory classification problems, Qiao et al. (2009) proposed to solve a penalized least squares problem with the lasso penalty, Clemmensen et al. (2011) proposed the sparse optimal scoring, and Witten and Tibshirani (2011) proposed the ℓ_1 penalized LDA. All of these methods involve solving a non-convex optimization problem and thus can be computationally expensive. Moreover, it is unclear how to extend these methods to a distributed setting when data are possibly stored in multiple sites. Recently, Mai et al. (2019) proposed to directly estimate the Bayes rule via a composite-type loss function. Safo and Ahn (2016) proposed to estimate the discriminant vectors by applying a basis approach to the methods studied in Cai and Liu (2011) and Shao et al. (2011). Gaynanova et al. (2016) proposed to extract the sparse discriminant vectors simultaneously through a convex programming problem. The underlying idea of these three methods are the same: for the purpose of classification, it suffices to accurately estimate the discriminant subspace, i.e., the subspace spanned by the discriminant vectors. In this paper, we will exploit this observation and propose an $\ell_{1,1}$ regularized multicategory linear discriminant annalysis method, which can also esti-

mate the sparse discriminant vectors simultaneously. It is worth mentioning that we choose the $\ell_{1,1}$ penalty instead of the $\ell_{2,1}$ penalty in Mai et al. (2019) and Gaynanova et al. (2016), so that we can easily extend the multicategory LDA method to a distributed setting and establish the corresponding statistical properties.

In addition to challenges posed by the high dimensionality of many datasets, with the rapid developments of science and technology, we have seen more and more datasets that are often scattered across distant servers, possibly due to the limitation of storage resources. The difficulties of fusing or aggregating these datasets due to the communication cost and privacy concerns have inspired many works in communication-efficient statistical learning (Lee et al., 2017; Wang et al., 2017; Fan et al., 2019; Jordan et al., 2019). One popular distributed estimation framework is the one-shot divide-and-conquer algorithm, in which one first computes the local estimators and then obtains the distributed estimator as an average of the local ones (Lee et al., 2017; Fan et al., 2019). These averaging-based approaches suffer from at least three drawbacks: non-diminishing bias when local estimators are biased, allowing only a limited number of local machines (much smaller than $\sqrt{N}$, where N is the total sample size), and poor performance in nonlinear problems (Jordan et al., 2019). To overcome these issues, Jordan et al. (2019) and Wang et al. (2017) studied a multi-round framework for distributed learning, which is referred to as the communication-efficient surrogate likelihood (CSL) framework by Jordan et al. (2019).

Although distributed classification has attracted some attention (Kokiopoulou and Frossard, 2010; Wang et al., 2019; Lian and Fan, 2018), few research has been done on the distributed sparse discriminant analysis. For binary classification, Tian and Gu (2017) proposed a divide-and-conquer sparse LDA method, in which they averaged the debiased local estimators of the discriminant direction and then sparsified the aggregated estimator. Moreover, they showed that their distributed sparse LDA method can achieve the same performance as the centralized one which used all samples. However, it is not straightforward to generalize their approach to the multicategory case. To handle the multicategory case, in this paper, we propose an $\ell_{1,1}$ regularized multicategory SLDA formulation and then extend it to the distributed setting via the CSL framework. Theoretically, we show that the distributed multicategory SLDA (dmSLDA) method performs as good as the centralized SLDA after a few rounds of communications. Moreover, we conduct numerical experiments to further support our methodology and theory.

The rest of this paper proceeds as follows. In Section

2, we review the multicategory sparse discriminant analysis and propose an $\ell_{1,1}$ regularized multicategory SLDA method. We then extend it to the distributed setting and propose the dmSLDA method in Section 3. Section 4 establishes the $\ell_{2,2}$ and $\ell_{1,1}$ estimation error bounds of dmSLDA under certain conditions and explain how the proposed distributed estimator achieves the same performance as the centralized version. To further back up the methodology and theory, we conduct several numerical experiments in Section 5. In Section 6, we conclude this paper with some discussions.

Notation. We summarize some notations that will be used throughout this paper. By convention, we use regular letters for scalars and bold letters for both vectors and matrices. We employ $[M]$ to represent the set $\{1, \ldots, M\}$ for any positive integer M and $[(a, b)]$ to abbreviate the set $\{(1, 1), (1, 2), \ldots, (a, b)\}$ for any positive integers a and b. For a vector $\boldsymbol{u} = (u_1, \ldots, u_d)^\top \in \mathbb{R}^d$, we denote its ℓ_q norm by $\|\boldsymbol{u}\|_q = (\sum_{i=1}^d |u_i|^q)^{1/q}$, its ℓ_0 norm by $\|\boldsymbol{u}\|_0 = \sum_{i=1}^d 1_{u_i \neq 0}$, and its ℓ_∞ norm by $\|\boldsymbol{u}\|_\infty = \max_{1 \leq i \leq d} |u_i|$. Moreover, for any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we use $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\top \boldsymbol{v}$ to denote the Euclidean inner product of $\boldsymbol{u}$ and $\boldsymbol{v}$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we define the $\ell_{p,q}$ norm of $\boldsymbol{A}$ as $\|\boldsymbol{A}\|_{p,q} = \| (\|\boldsymbol{A}_{1\cdot}\|_p, \ldots, \|\boldsymbol{A}_{m\cdot}\|_p) \|_q$, where $\boldsymbol{A}_{i\cdot}$ is the ith row of $\boldsymbol{A}$. For any two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$, we denote $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \text{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, we use $\text{span}(\boldsymbol{A})$ to denote the subspace spanned by the columns of $\boldsymbol{A}$. For two sequences of real numbers $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$, we write $a_n \lesssim b_n$ if $a_n \leq C b_n$ for some constant $C > 0$ independent of n.

2 MULTICATEGORY SLDA

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be a collection of N i.i.d. samples with $y_i \in [K]$ drawn from a discrete distribution $\mathbb{P}(y_i = k) = \pi_k$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ drawn from a distribution satisfying $\mathbb{E}[\boldsymbol{x}_i | y_i = k] = \boldsymbol{\mu}_k$ and $\text{Var}(\boldsymbol{x}_i | y_i = k) = \boldsymbol{\Sigma}$, where K is the number of classes, π_k is the prior probability of class k such that $\pi_k \geq 0$ and $\sum_{k=1}^K \pi_k = 1$, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ are the class-conditional mean and covariance matrix of $\boldsymbol{x}_i$ in class k, respectively. The K centroids lie in an affine space of dimension $r_0 \leq K - 1$, which can be much smaller than d (Hastie et al., 2009). To exploit the underlying low dimensional structure, Fisher proposed to solve a set of discriminant vectors $\{\boldsymbol{v}_k^*\}_{k=1}^q \subset \mathbb{R}^d$ by maximizing the following sequence of Rayleigh quotients under orthogonal constraints,

$$\boldsymbol{v}_k^* = \operatorname*{argmax}_{\boldsymbol{v}_k \in \mathbb{R}^d} \frac{\boldsymbol{v}_k^\top \boldsymbol{B} \boldsymbol{v}_k}{\boldsymbol{v}_k^\top \boldsymbol{\Sigma} \boldsymbol{v}_k}, \tag{2.1}$$
$$\text{s.t.} \quad \boldsymbol{v}_k^{*\top} \boldsymbol{\Sigma} \boldsymbol{v}_j^* = 0, \quad \forall 1 \leq j \leq k \leq q,$$

where $\boldsymbol{B} = \mathrm{Var}\left(\mathbb{E}[\boldsymbol{x}|y]\right)$ is the between-class covariance and $\boldsymbol{\Sigma} = \mathbb{E}[\mathrm{Var}\left(\boldsymbol{x}|y\right)]$ is the within-class covariance. In other words, Fisher's proposal seeks a low dimensional projection of samples that maximizes the between-class variation relative to the within-class variation. Classification can then be performed on the projected data $\boldsymbol{V}_q^\top \boldsymbol{x} \in \mathbb{R}^q$, where $\boldsymbol{V}_q = (\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_q^*) \in \mathbb{R}^{d \times q}$. In the classical setting where $N \gg d$, the unknown between-class covariance $\boldsymbol{B}$ and within-class covariance $\boldsymbol{\Sigma}$ in (2.1) are substituted in practice by their corresponding sample versions $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Sigma}}$ respectively, where

$$\widehat{\boldsymbol{B}} = \sum_{k=1}^{K} \widehat{\pi}_k (\widehat{\boldsymbol{\mu}}_k - \widehat{\overline{\boldsymbol{\mu}}})(\widehat{\boldsymbol{\mu}}_k - \widehat{\overline{\boldsymbol{\mu}}})^\top,$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{y_i=k} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^\top, \quad (2.2)$$

in which $\widehat{\boldsymbol{\mu}}_k = \frac{1}{b_k} \sum_{y_i=k} \boldsymbol{x}_i$ is the sample mean of class k, $b_k = \sum_{i=1}^{N} 1_{y_i=k}$ is the sample size of class k, $\widehat{\pi}_k = b_k/N$ is the sample proportion of class k, and $\widehat{\overline{\boldsymbol{\mu}}} = \sum_{k=1}^{K} \widehat{\pi}_k \widehat{\boldsymbol{\mu}}_k = \frac{1}{N} \sum_i \boldsymbol{x}_i$ is the overall sample mean.

In high dimensions when $d > N$, the Fisher's approach (2.1) suffers from the singularity of $\widehat{\boldsymbol{\Sigma}}$, the noise accumulation, and the lack of interpretability (Bickel and Levina, 2004; Fan and Fan, 2008; Witten and Tibshirani, 2011). To overcome these issues, Witten and Tibshirani (2011) and Clemmensen et al. (2011) proposed sparse discriminant analysis, which requires solving a sequence of non-convex optimization problems and thus can be computationally intractable. On the other hand, Mai et al. (2019), Safo and Ahn (2016) and Gaynanova et al. (2016) proposed to compute any set of basis vectors of the discriminant subspace, i.e., $\mathcal{V}_q := \mathrm{span}(\boldsymbol{V}_q)$ and then sparsified these basis vectors. Though not directly reported, the core of these ideas is the following proposition.

Proposition 2.1. *Suppose $\boldsymbol{V}, \boldsymbol{W} \in \mathbb{R}^{d \times q}$ span the same subspace, then the LDA classification rule based on $\boldsymbol{V}^\top \boldsymbol{x}$ is equivalent to the LDA classification rule based on $\boldsymbol{W}^\top \boldsymbol{x}$.*

Proposition 2.1 is a generalization of Proposition 4 in Gaynanova et al. (2016), implying that for the purpose of classification, it suffices to find any set of basis vectors of the discriminant subspace.

Proposition 2.2. *Assume that the within-class covariance $\boldsymbol{\Sigma}$ is nonsingular and define $\boldsymbol{W}^* \in \mathbb{R}^{d \times (K-1)}$ as the solution to the following optimization problem*

$$\boldsymbol{W}^* = \operatorname*{argmin}_{\boldsymbol{W} \in \mathbb{R}^{d \times (K-1)}} \left\{ \frac{1}{2} \langle \boldsymbol{W}, \boldsymbol{\Sigma W} \rangle - \langle \boldsymbol{W}, \boldsymbol{U} \rangle \right\}, \quad (2.3)$$

where $\boldsymbol{U} = (\boldsymbol{\mu}_1 - \bar{\boldsymbol{\mu}}, \ldots, \boldsymbol{\mu}_{K-1} - \bar{\boldsymbol{\mu}}) \in \mathbb{R}^{d \times (K-1)}$ and $\bar{\boldsymbol{\mu}} = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$. Then we have $\mathcal{W} := \mathrm{span}(\boldsymbol{W}^*) = \mathcal{V}_{\mathrm{rank}(\boldsymbol{B})}$, where $\boldsymbol{B}$ is the between-class covariance.

Proposition 2.2 gives an explicit choice of the basis vectors of the discriminant subspace, simpler than that in Mai et al. (2019) or Gaynanova et al. (2016). Since $\boldsymbol{\Sigma}$ and $\boldsymbol{U}$ in (2.3) are unknown in practice, we will substitute them by their corresponding sample versions $\widehat{\boldsymbol{\Sigma}}$ given by (2.2) and $\widehat{\boldsymbol{U}} = (\widehat{\boldsymbol{\mu}}_1 - \widehat{\overline{\boldsymbol{\mu}}}, \ldots, \widehat{\boldsymbol{\mu}}_{K-1} - \widehat{\overline{\boldsymbol{\mu}}})$, respectively. In high dimensions, we assume element-wise sparsity on $\boldsymbol{W}^*$ and propose the following $\ell_{1,1}$ regularized estimator

$$\widehat{\boldsymbol{W}} = \operatorname*{argmin}_{\boldsymbol{W} \in \mathbb{R}^{d \times (K-1)}} \left\{ \frac{1}{2} \langle \boldsymbol{W}, \widehat{\boldsymbol{\Sigma}} \boldsymbol{W} \rangle - \langle \boldsymbol{W}, \widehat{\boldsymbol{U}} \rangle \right.$$
$$\left. + \lambda \|\boldsymbol{W}\|_{1,1} \right\}, \quad (2.4)$$

where $\lambda > 0$ is a tuning parameter. We choose the $\ell_{1,1}$ penalty (or the element-wise sparsity) rather than the $\ell_{2,1}$ penalty (or the row sparsity) used in Mai et al. (2019) and Gaynanova et al. (2016) because only with the $\ell_{1,1}$ penalty shall we extend the multicategory SLDA method to the distributed setting easily. Moreover, the assumption of element-wise sparsity is a weaker assumption than the row sparsity. Finally, to perform classification, one can apply LDA to the projected data $\widehat{\boldsymbol{W}}^\top \boldsymbol{x} \in \mathbb{R}^{K-1}$.

Remark 2.3. *The optimization problem (2.4) and similarly (3.1) and (3.2) in the next section can be solved efficiently using the fast iterative shrinkage-thresholding algorithm (FISTA) proposed by Beck and Teboulle (2009).*

3 DISTRIBUTED MULTICATEGORY SLDA

In this section, we extend the proposed $\ell_{1,1}$ regularized multicategory SLDA method to the distributed setting. Suppose there are total of M machines and on machine $m \in [M]$, n samples[1] $\{(\boldsymbol{x}_i^m, y_i^m)\}_{i=1}^n$ are drawn independently from the distribution specified in Section 2. Moreover, samples on different machines are assumed to be independent. To recover the discriminant subspace, the most straightforward yet unrealistic approach is to aggregate data from all machines together and then apply the $\ell_{1,1}$ regularized multicategory SLDA method based on the full sample. Another approach is to gather the local statis-

---

[1]Here we assume for simplicity that the sample sizes across different machines are equal. In general, our framework can allow different sample sizes on each machine as long as the sample sizes are approximately balanced; that is the proportion of any single class can not vanish to zero.

tics $\widehat{\boldsymbol{\Sigma}}^m$ and $\widehat{\boldsymbol{U}}^m$ from all machines and then estimate $\boldsymbol{W}^*$ using (2.4) with the aggregated statistics $\widehat{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\boldsymbol{\Sigma}}^m$ and $\widehat{\boldsymbol{U}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\boldsymbol{U}}^m$. Although these two centralized methods usually give the best estimation and prediction performance, they are communication expensive, with communication cost $\mathcal{O}(ndM)$ and $\mathcal{O}(d^2M)$, respectively. This motivates us to design a communication-efficient algorithm that achieves the same performance as the centralized SLDA.

We apply the communication-efficient surrogate likelihood framework (Jordan et al., 2019) to the multicategory SLDA method (2.4) and obtain the distributed multicategory SLDA method (dmSLDA). The full algorithm is collected in Algorithm 1. Specifically, we

---

Algorithm 1: Distributed multicategory SLDA (dmSLDA).

Input: $\{(\boldsymbol{x}_i^m, y_i^m)\}_{i=1}^{n_m}$, T
Output: $\widehat{\boldsymbol{W}}$
1 compute $\widehat{\boldsymbol{\Sigma}}^m$ and $\widehat{\boldsymbol{U}}^m$ for all m;
2 compute $\widehat{\boldsymbol{W}}_0$ as the solution to (3.1);
3 **for** t in $1:T$ **do**
4   send $\widehat{\boldsymbol{W}}_{t-1}$ to local machines;
5   compute $\nabla\mathcal{L}_m(\widehat{\boldsymbol{W}}_{t-1})$ on local machines;
6   send $\nabla\mathcal{L}_m(\widehat{\boldsymbol{W}}_{t-1})$ back to the master machine;
7   compute $\widehat{\boldsymbol{W}}_t$ as the solution to (3.2);
8 **end**
9 return $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{W}}_{t^*}$ with $t^* = \mathrm{argmin}_{0 \le t \le T}\, \widehat{e}_t$;

---

first compute the local estimators $\widehat{\boldsymbol{\Sigma}}^m$ and $\widehat{\boldsymbol{U}}^m$. Then we compute the initial estimator $\widehat{\boldsymbol{W}}_0$ on the master machine, which is set to be the first local machine without loss of generality, as the solution to the following $\ell_{1,1}$ regularized optimization problem,

$$\widehat{\boldsymbol{W}}_0 = \underset{\boldsymbol{W} \in \mathbb{R}^{d \times (K-1)}}{\mathrm{argmin}} \left\{ \mathcal{L}_1(\boldsymbol{W}) + \lambda_0 \|\boldsymbol{W}\|_{1,1} \right\}, \quad (3.1)$$

where $\lambda_0 > 0$ is a tuning parameter and $\mathcal{L}_m(\boldsymbol{W})$ is the loss function on machine m given by

$$\mathcal{L}_m(\boldsymbol{W}) = \frac{1}{2}\langle \boldsymbol{W}, \widehat{\boldsymbol{\Sigma}}^m \boldsymbol{W}\rangle - \langle \boldsymbol{W}, \widehat{\boldsymbol{U}}^m\rangle.$$

At the tth iteration, we send the $(t-1)$th estimator $\widehat{\boldsymbol{W}}_{t-1}$ to local machines, compute the gradients $\nabla\mathcal{L}_m(\widehat{\boldsymbol{W}}_{t-1})$ of the mth loss function $\mathcal{L}_m(\cdot)$ at $\widehat{\boldsymbol{W}}_{t-1}$, and then send these gradients back to the master machine. Then on the master machine, we compute the tth estimator by minimizing the following shifted $\ell_{1,1}$

regularized objective function,

$$\widehat{\boldsymbol{W}}_t = \underset{\boldsymbol{W} \in \mathbb{R}^{d \times (K-1)}}{\mathrm{argmin}} \left\{ \mathcal{L}_1(\boldsymbol{W}) + \lambda_t \|\boldsymbol{W}\|_{1,1} \right.$$
$$\left. + \langle \frac{1}{M}\sum_{m=1}^{M} \nabla\mathcal{L}_m(\widehat{\boldsymbol{W}}_{t-1}) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{W}}_{t-1}), \boldsymbol{W}\rangle \right\}, \quad (3.2)$$

where $\lambda_t > 0$ is a tuning parameter. After repeating the procedure for T times, the dmSLDA algorithm returns $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{W}}_{t^*}$ with $t^* = \mathrm{argmin}_{0 \le t \le T}\, \widehat{e}_t$, where $\widehat{e}_t = \sum_{m=2}^{M} \mathcal{L}_m(\widehat{\boldsymbol{W}}_t)$. In other words, we choose the best estimator in $\{\widehat{\boldsymbol{W}}_t\}_{t=0}^{T}$ in the sense that it minimizes the validation loss. Finally, one can perform classification based on the projected data $\widehat{\boldsymbol{W}}^\top \boldsymbol{x} \in \mathbb{R}^{K-1}$.

Remark 3.1. *The communication cost of the dmSLDA method is $\mathcal{O}(dKM)$ since it only communicates the estimators $\widehat{\boldsymbol{W}}_t$ and the gradients $\nabla\mathcal{L}_m(\widehat{\boldsymbol{W}}_{t-1})$. This is significantly smaller than that of the centralized methods, i.e., $\mathcal{O}(ndM)$ (communicate all raw data $\{(\boldsymbol{x}_i^m, y_i^m)\}$) or $\mathcal{O}(d^2M)$ (communicate local estimators $\widehat{\boldsymbol{\Sigma}}^m$ and $\widehat{\boldsymbol{U}}^m$), when $K \ll d, n$.*

Remark 3.2. *In numerical experiments, we propose to choose λ_t from a candidate set Λ_t by minimizing the validation loss $\widehat{e}(\widehat{\boldsymbol{W}}_t) := \sum_{m=2}^{M} \mathcal{L}_m(\widehat{\boldsymbol{W}}_t)$, i.e., the loss on all other machines.*

Remark 3.3. *Since in general a few iterations suffice for the procedure to match the same accuracy of the centralized estimator (Jordan et al., 2019), we choose $T = 3$ or 5 in practice.*

4 STATISTICAL ANALYSIS

In this section, we establish statistical properties ensuring that dmSLDA can achieve the performance of the centralized SLDA after a few rounds of communication. We prove this in two steps: first, we will establish a recursive estimation error bound, i.e., we upper bound the estimation error $\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|$ in terms of the error in the previous round $\|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|$; second, by applying this recursive estimation error bound iteratively, we can obtain an estimation error bound indicating that dmSLDA matches the performance of the centralized SLDA after a few rounds.

To begin with, we make a few assumptions. For the sake of simplicity, we consider a balanced setting defined below[2]. In the balanced setting, the prior probability and the subsample size of each class are equal

---

[2]If the balance assumption does not hold and some class has significantly small prior probability and subsample size, one may refer to literature related for imbalanced data, such as Krawczyk (2016) and the references therein.

across different classes and machines. Our framework does allow approximately balanced settings when the sample size proportions of any single class and any single machine do not vanish.

Assumption 4.1 (Balanced). *We assume all classes have equal prior probability, i.e., $\pi_k = \frac{1}{K}$. Moreover, the subsample size of class k on machine m is assumed to be constant, denoted by b. Thus, the subsample size on a single machine is $n = Kb$ and the full sample size is $N = Mn$.*

Furthermore, we assume that the conditional distribution of $\boldsymbol{x}$ given $y = k$ is sub-Gaussian with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}$.

Assumption 4.2 (sub-Gaussian). *We assume that given $y = k$, the transformed variable $\boldsymbol{z}^k = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}_k)$ follows a sub-Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{I}_d$. In particular, there exists some $\sigma^2 > 0$ such that the following inequality holds,*

$$\mathbb{E}[e^{\lambda \langle \boldsymbol{v}, \boldsymbol{z}^k \rangle} \mid y = k] \leq e^{\frac{\lambda^2 \sigma^2}{2}}, \forall \lambda \in \mathbb{R}, \boldsymbol{v} \in \mathbb{R}^d, \|\boldsymbol{v}\|_2 = 1.$$

Moreover, we assume a specific restricted eigenvalue (RE) condition on the population covariance matrix $\boldsymbol{\Sigma}$, which is a common assumption in high-dimensional statistics (Cai and Liu, 2011; Lee et al., 2017; Tian and Gu, 2017; Wang et al., 2017). First, given a subset $\mathcal{A} \subset [d]$, we define the cone $\mathcal{C}(\mathcal{A}, 3) \subset \mathbb{R}^d$ as

$$\mathcal{C}(\mathcal{A}, 3) = \left\{ \boldsymbol{\zeta} \in \mathbb{R}^d \mid \|\boldsymbol{\zeta}_{\mathcal{A}^c}\|_1 \leq 3\|\boldsymbol{\zeta}_{\mathcal{A}}\|_1 \right\},$$

where $\boldsymbol{\zeta}_{\mathcal{A}} \in \mathbb{R}^d$ is $\boldsymbol{\zeta}$ with nonzeros restricted on the set $\mathcal{A}$, i.e., $\boldsymbol{\zeta}_{\mathcal{A},i} = \boldsymbol{\zeta}_i$, if $i \in \mathcal{A}$, $\boldsymbol{\zeta}_{\mathcal{A},i} = 0$ otherwise; and $\boldsymbol{\zeta}_{\mathcal{A}^c}$ is analogously defined. Then we assume $\boldsymbol{\Sigma}$ satisfies the $RE(\kappa_0, s)$ condition for some constant $\kappa_0 > 0$ and positive integer s, which is defined below.

Assumption 4.3 (Restricted Eigenvalue). *We assume that the covariance matrix $\boldsymbol{\Sigma}$ satisfies the $RE(\kappa_0, s)$ condition, i.e., there exists some constant $\kappa_0 > 0$ such that*

$$\boldsymbol{\zeta}^\top \boldsymbol{\Sigma} \boldsymbol{\zeta} \geq \kappa_0 \|\boldsymbol{\zeta}\|_2^2, \ \forall \boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A}, 3), \ |\mathcal{A}| \leq s.$$

Assumption 4.2 and 4.3 allow us to establish concentration inequalities for $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{U}}$ as well as an RE condition on the sample covariance $\widehat{\boldsymbol{\Sigma}}^1$, which are stated in Lemmas A.1 and A.4. We are ready to present our main theoretical results. We need to define the shifted loss function $\widetilde{\mathcal{L}}_1(\boldsymbol{W}, \widehat{\boldsymbol{W}}_t)$ as follows,

$$\widetilde{\mathcal{L}}_1(\boldsymbol{W}, \widehat{\boldsymbol{W}}_t) = \mathcal{L}_1(\boldsymbol{W}) + \langle \frac{1}{M} \sum_{m=1}^M \nabla \mathcal{L}_m(\widehat{\boldsymbol{W}}_t) - \nabla \mathcal{L}_1(\widehat{\boldsymbol{W}}_t), \boldsymbol{W} \rangle. \quad (4.1)$$

We establish in Lemma 4.4 an upper bound on the $\ell_{\infty,\infty}$ norm $\|\nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)\|_{\infty,\infty}$ of the gradient of the shifted loss function at $\boldsymbol{W}^*$, which suggests an appropriate choice of the tuning parameter λ_{t+1}.

Lemma 4.4. *Suppose Assumptions 4.1 and 4.2 hold and $\sup_j \boldsymbol{\Sigma}_{jj} \leq c_0$ for some constant c_0, then there exist some universal contants c_1 and c_2 such that*

$$\|\nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)\|_{\infty,\infty} \leq a_n + b_n \|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1},$$

holds for all t with probability at least $1 - 2\delta$, where

$$a_n = c_1 \sqrt{\frac{\log(dM/\delta)}{N}} \|\boldsymbol{W}^*\|_{1,1}$$
$$+ c_2 \sqrt{\frac{K \log(dMK/\delta)}{N}}, \quad (4.2)$$
$$b_n = 2c_1 \sqrt{\frac{\log(dM/\delta)}{n}}. \quad (4.3)$$

The $\ell_{1,1}$ regularization parameter λ_{t+1} is set as twice the upper bound in Lemma 4.4, i.e.,

$$\lambda_{t+1} = 2a_n + 2b_n \|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1}. \quad (4.4)$$

Although the right hand side of (4.4) involves the oracle matrix $\boldsymbol{W}^*$ and thus unknown, it however suggests how to tune the parameter in practice. For example, the parameter λ_{t+1} should decrease as the estimation error $\|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1}$ goes down during iterations. Next, with this specific choice of parameter λ_{t+1}, we establish in Theorem 4.5 a recursive estimation error bound connecting the estimation error $\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|$ to that of the previous iteration $\|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|$, where the norm is the $\ell_{1,1}$ or $\ell_{2,2}$ norm.

Theorem 4.5. *Suppose that Assumptions 4.1, 4.2, and 4.3 hold, $\sup_j \boldsymbol{\Sigma}_{jj} \leq c_0$ for some constant c_0, and the regularization parameter λ_{t+1} is set as in (4.4), then there exist some universal constants c_1, c_2, c_3 such that for $n \geq c_3 s \log(d/\delta)$, the recursive estimation error bounds*

$$\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2} \leq \frac{48 a_n \sqrt{s}}{\kappa_0}$$
$$+ \frac{48 b_n \sqrt{s}}{\kappa_0} \|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1},$$
$$\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1} \leq \frac{192 a_n s}{\kappa_0}$$
$$+ \frac{192 b_n s}{\kappa_0} \|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1},$$

hold for all t with probability at least $1 - 3\delta$, where a_n and b_n are given by (4.2) and (4.3), respectively, and $s = |\mathcal{S}|$ is the cardinality of the support $\mathcal{S}$ of $\boldsymbol{W}^*$.
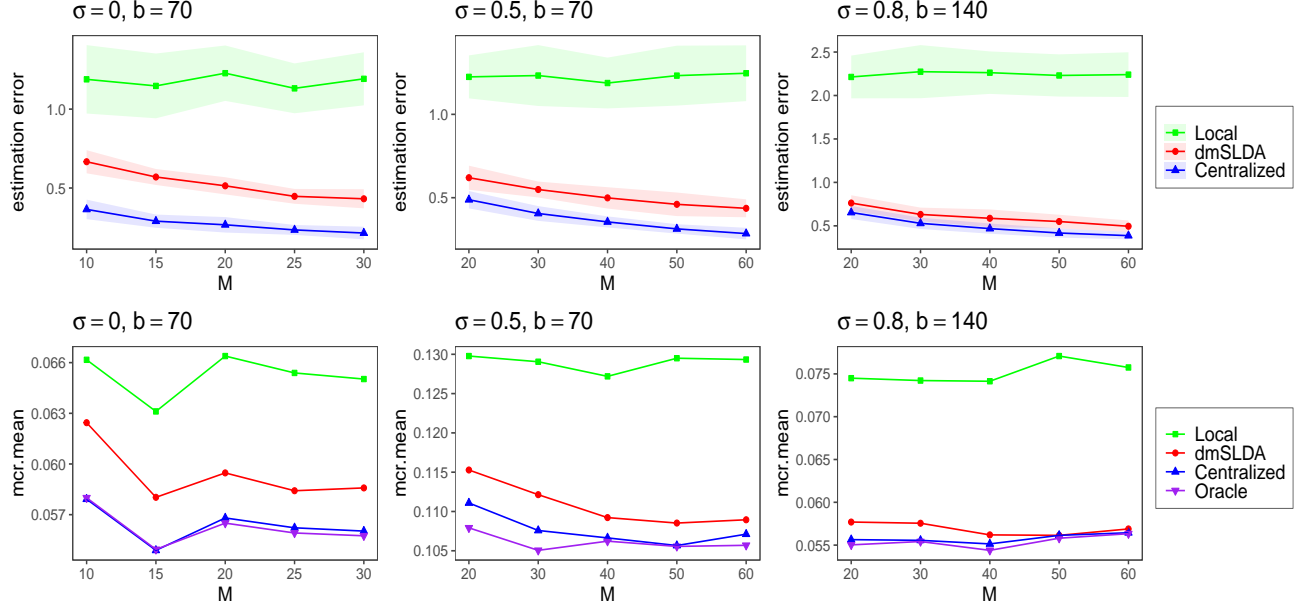
Figure 1: The top plots show the mean and the standard deviation of $\ell_{2,2}$ estimation error for dmSLDA, local SLDA, and centralized SLDA over 40 repetitions for different σ, b and M. The bottom plots show the averaged MCR on the testing data for all four methods over 40 repetitions.

Theorem 4.5 upper bounds the $(t+1)$th estimation error $\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1}$ by a linear function of the tth estimation error $\|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1}$. Thus by applying Theorem 4.5 iteratively, we can bound $\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|$ in terms of the size of the local estimation error $\|\widehat{\boldsymbol{W}}_0 - \boldsymbol{W}^*\|$.

Theorem 4.6. *Assume that the conditions in Theorem 4.5 hold and the regularization parameter λ_{t+1} is set as in (4.4), then there exist some universal constants c_1, c_2, c_3 such that for $n \geq c_3 s \log(d/\delta)$, the estimation error bounds*

$$\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1} \leq \frac{\mathfrak{a}(1 - \mathfrak{b}^{t+1})}{1 - \mathfrak{b}}$$
$$+ \mathfrak{b}^{t+1}\|\widehat{\boldsymbol{W}}_0 - \boldsymbol{W}^*\|_{1,1},$$

$$\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2} \leq \frac{\mathfrak{a}(1 - \mathfrak{b}^{t+1})}{4\sqrt{s}(1 - \mathfrak{b})}$$
$$+ \frac{\mathfrak{b}^{t+1}}{4\sqrt{s}}\|\widehat{\boldsymbol{W}}_0 - \boldsymbol{W}^*\|_{1,1},$$

hold with probablity at least $1 - 3\delta$, where $\mathfrak{a} = \frac{192 a_n s}{\kappa_0}$ and $\mathfrak{b} = \frac{192 b_n s}{\kappa_0}$.

Theorem 4.6 provides theoretical guarantees that dmSLDA performs as well as the centralized SLDA in terms of the estimation accuracy. To see that, we observe that the coefficient $\mathfrak{b} < 1$ when the local sample size n is sufficiently large and thus the terms involving $\mathfrak{b}^{t+1}$ are negligible after a few (logarithmic) iterations.

In other words, the estimation error for sufficiently large t satisfies

$$\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2} \lesssim \sqrt{\frac{s \log(dM)}{N}},$$

which is the same as that of the centralized method.

5 NUMERICAL STUDIES

In this section, we conduct numerical studies to futher back up our methodology and theory. The R code is released here.

5.1 dmSLDA

In this subsection, we conduct numerical simulations to illustrate the theoretical results in Section 4. Four methods are implemented and compared: the dmSLDA, the local SLDA ($\widehat{\boldsymbol{W}}_0$), the centralized SLDA ($\widehat{\boldsymbol{W}}$ based on $\widehat{\boldsymbol{\Sigma}} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\Sigma}}^m$ and $\widehat{\boldsymbol{U}} = \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{U}}^m$), and the oracle ($\boldsymbol{W}^*$). For the dm-SLDA, we set $T = 3$ as the maximum rounds of communication. In the experiments, we fix the number of classes $K = 3$, the dimension $d = 400$, and the class-conditional means $\boldsymbol{\mu}_1 = (-2, -2, -2, 0, \ldots, 0)^\top$, $\boldsymbol{\mu}_2 = (0, 0, 0, 2, 2, 2, 0, \ldots, 0)^\top$, and $\boldsymbol{\mu}_3 = (0, \ldots, 0)^\top$. The class-conditional covariance $\boldsymbol{\Sigma}$ is set as $\boldsymbol{\Sigma}_{ij} = \sigma^{|i-j|}$ for all $1 \leq i \leq j \leq d$, where $\sigma \in \{0, 0.5, 0.8\}$ gives three different settings and $0^0 = 1$ by convention. We

set the subsample size $b = 70$ of each class on one machine when $\sigma \in \{0, 0.5\}$ and set $b = 140$ when $\sigma = 0.8$.[3] Let the number of machines M vary from 20 to 60 by an increment of 10. For each setting, i.e., $(\sigma, b) \in \{(0, 70), (0.5, 70), (0.8, 140)\}$ and $M \in \{20, 30, \ldots, 60\}$, we repeat the following procedure 40 times. First, we generate the training data $\{\boldsymbol{x}_{kj}^m\}_{j=1}^b$ of class k on machine m and the testing data $\{\boldsymbol{x}_{kj}\}_{j=1}^{300}$ of class k from $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ independently. Once the data are generated, we estimate the transformation matrix $\boldsymbol{W}^*$ based on the training data and then compute both the $\ell_{2,2}$ estimation error $\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_{2,2}$ and the misclassification rate (MCR) based on the testing data. The results are displayed in Figure 1.

It can be seen from Figure 1 that in all settings, the $\ell_{2,2}$ estimation error of the dmSLDA is close to that of the centralized SLDA, which is much smaller than that of the local SLDA, as implied by Theorem 4.6. Moreover, the averaged MCR of the centralized SLDA is almost equal to that of the oracle SLDA, to which the averaged MCR of the dmSLDA is comparable. They all outperform the local SLDA in terms of the averaged MCR. The rationale behind this phenomenon is that a more accurate estimator of $\boldsymbol{W}^*$ leads to a better linear discriminant classifier.

5.2 Binary Classification

In this subsection, we compare the dmSLDA method with other distributed classifiers in the literature. To the best of our knowledge, dmSLDA is the first distributed multicategorical classifier for high dimensional data. Thus, we restrict our experiments to the binary setting. Specifically, we compare dmSLDA with the Centralized SLDA (Cen.SLDA) and two other distributed binary classifiers, the divide-and-conquer SLDA (dSLDA) method proposed by Tian and Gu (2017) and the divide-and-conquer ℓ_1 SVM (dSVM) method proposed by Lian and Fan (2018). We adopt the following binary balanced setting: $K = 2$, $d = 200$, $b = 200$, $\boldsymbol{\Sigma}_{ij} = 0.8^{|i-j|}$, $\boldsymbol{\mu}_1 = (\text{rep}(0.5, 10), 0, \ldots)$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$ and $M \in \{5, 10, 15, 20, 30\}$. For each setting, we generate the training data $\{\boldsymbol{x}_{kj}^m\}_{j=1}^b$ of class k on machine m and the testing data $\{\boldsymbol{x}_{kj}\}_{j=1}^{300}$ of class k from $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ independently and then implement all four methods. The whole procedure is repeated 40 times and the averaged misclassification rates (MCR) are reported in Figure 2. It turns out that dmSLDA, dSLDA and Cen.SLDA are comparable and all these three methods outperform dSVM.
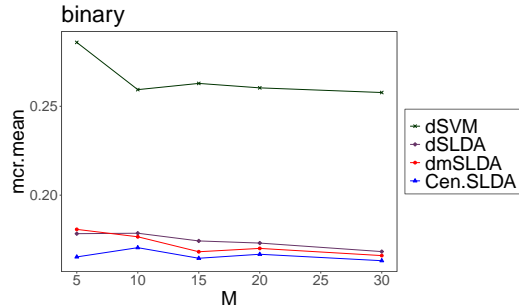

Figure 2: Comparsion between dSVM, dSLDA, dmSLDA, and Cen.SLDA. The averaged misclassification rate is computed over 40 repetitions.

6 DISCUSSION

In this paper, we propose an $\ell_{1,1}$ regularized multicategory SLDA method and extend it to the distributed setting when data are stored across multiple sites. We establish statistical properties ensuring that the distributed sparse multicategory LDA performs as good as the centralized sparse LDA after a few rounds of communications. Here we assume the data points follow sub-Gaussian distributions. It is possible to extend the current framework to deal with heavy-tailed data or data with adversarial contamination. We leave this to future work.

References

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989 – 1010.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association*, 106(496):1566–1577.

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605–2637.

---

[3]When σ increases, κ_0 decreases and thus we need a larger subsample size.

Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771.

Fan, J., Wang, D., Wang, K., and Zhu, Z. (2019). Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031.

Gaynanova, I., Booth, J. G., and Wells, M. T. (2016). Simultaneous sparse estimation of canonical vectors in the p ≫ n setting. *Journal of the American Statistical Association*, 111(514):696–706.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical science*, 21(1):1–14.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, Berlin, 2 edition.

Jordan, M. I., Lee, J. D., and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681.

Kokiopoulou, E. and Frossard, P. (2010). Distributed classification of multiple observation sets by consensus. *IEEE Transactions on Signal Processing*, 59(1):104–114.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. (2017). Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144.

Lian, H. and Fan, Z. (2018). Divide-and-conquer for debiased l1-norm support vector machine in ultra-high dimensions. *Journal of Machine Learning Research*, 18:1–26.

Mai, Q., Yang, Y., and Zou, H. (2019). Multiclass sparse discriminant analysis. *Statistica Sinica*, 29(1):97–111.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42.

Michie, D., Spiegelhalter, D. J., Taylor, C. C., and Campbell, J., editors (1994). *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, Chichester.

Pilanci, M. and Wainwright, M. J. (2015). Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115.

Qiao, Z., Zhou, L., and Huang, J. Z. (2009). Sparse linear discriminant analysis with applications to high dimensional low sample size data. *International Journal of Applied Mathematics*, 39(1):48–60.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE transactions on information theory*, 59(6):3434–3447.

Safo, S. E. and Ahn, J. (2016). General sparse multiclass linear discriminant analysis. *Computational Statistics & Data Analysis*, 99:81–90.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics*, 39(2):1241–1265.

Tian, L. and Gu, Q. (2017). Communication-efficient distributed sparse linear discriminant analysis. In *Artificial Intelligence and Statistics*, volume 54, pages 1178–1187. PMLR.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, J., Kolar, M., Srebro, N., and Zhang, T. (2017). Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, volume 70, pages 3636–3645. PMLR.

Wang, X., Yang, Z., Chen, X., and Liu, W. (2019). Distributed inference for linear support vector machine. *Journal of Machine Learning Research*, 20:1–41.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

Supplementary Material:
Distributed Sparse Multicategory Discriminant Analysis

A PROOF OF MAIN RESULTS

A.1 Proof of Proposition 2.1

Proof of Proposition 2.1. Since $\boldsymbol{V}, \boldsymbol{W} \in \mathbb{R}^{d \times q}$ span the same subspace, there exists an invertible matrix $\boldsymbol{R} \in \mathbb{R}^{q \times q}$ such that $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{R}$. The remaining proof follows the same argument of Proposition 4 in Gaynanova et al. (2016), while we note that the generalization of the orthogonal matrix $\boldsymbol{R}$ to the general invertible matrix $\boldsymbol{R}$ does not affect the proof. □

A.2 Proof of Proposition 2.2

Proof of Proposition 2.2. Since $\boldsymbol{\Sigma}$ is invertible, the optimization problem (2.3) has a closed form solution $\boldsymbol{W}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{U}$. The core of the proof lies in the decomposition of the between-class covariance matrix $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{P}\boldsymbol{U}^\top$ for some symmetric positive definite matrix $\boldsymbol{P} \in \mathbb{R}^{(K-1) \times (K-1)}$. By definition,

$$
\begin{aligned}
\boldsymbol{B} &= \sum_{k=1}^{K} \pi_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top \\
&= \sum_{k=1}^{K-1} \pi_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top + \frac{1}{\pi_K} \Big(\sum_{k=1}^{K-1} \pi_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}) \Big) \Big(\sum_{k=1}^{K-1} \pi_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}) \Big)^\top,
\end{aligned}
$$

where the second equality holds since $\pi_K (\boldsymbol{\mu}_K - \bar{\boldsymbol{\mu}}) = -\sum_{k=1}^{K-1} \pi_k (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}))$. Then we can write $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{P}\boldsymbol{U}^\top$, with $\boldsymbol{P} \in \mathbb{R}^{(K-1) \times (K-1)}$ given by

$$
\boldsymbol{P} = \text{diag}(\pi_1, \ldots, \pi_{K-1}) + \frac{1}{\pi_K} (\pi_1, \ldots, \pi_{K-1})^\top (\pi_1, \ldots, \pi_{K-1}).
$$

Note that $\boldsymbol{P}$ is a symmetric positive definite matrix. Then we study the relationship between $\boldsymbol{W}^*$ and discriminant vectors $\{\boldsymbol{v}_k^*\}_{k=1}^q$ solved by Fisher's reduced rank LDA problem (2.1), where $q = \text{rank}(\boldsymbol{B})$. Let $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}\boldsymbol{P}\boldsymbol{U}^\top\boldsymbol{\Sigma}^{-1/2}$, then the Fisher's reduced rank LDA problem can be transformed into a generalized eigenvalue problem

$$
\begin{aligned}
\widetilde{\boldsymbol{v}}_k^* = \underset{\widetilde{\boldsymbol{v}}_k \in \mathbb{R}^d}{\arg\max} \ & \frac{\widetilde{\boldsymbol{v}}_k^\top \boldsymbol{A} \widetilde{\boldsymbol{v}}_k}{\widetilde{\boldsymbol{v}}_k^\top \widetilde{\boldsymbol{v}}_k}, \\
\text{s.t.} \quad & \widetilde{\boldsymbol{v}}_k^\top \widetilde{\boldsymbol{v}}_j = 0, \quad \forall 1 \le j \le k \le q,
\end{aligned} \tag{A.1}
$$

where $\widetilde{\boldsymbol{v}}_k^* = \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k^*$. Since $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{B}) = q$, the matrix $\boldsymbol{A}$ can be orthogonally diagonalized as

$$
\boldsymbol{A} = \boldsymbol{Q}\text{diag}(\lambda_1, \ldots, \lambda_q, 0, \ldots, 0)\boldsymbol{Q}^\top,
$$

where $\boldsymbol{Q} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d) \in \mathbb{R}^{d \times d}$ is an orthogonal matrix and $\lambda_1 \ge \cdots \ge \lambda_q > 0$. Then the solution to the problem (A.1) is $(\widetilde{\boldsymbol{v}}_1^*, \ldots, \widetilde{\boldsymbol{v}}_q^*) = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q)$ and thus the solution to the Fisher's reduced rank problem (2.1) is

$\boldsymbol{V}_q = (\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_q^*) = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q)$. As a result, we can express the linear subspace $\mathcal{V}_q := \text{span}(\boldsymbol{V}_q)$ as

$$
\begin{aligned}
\mathcal{V}_q &= \text{span}(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_q)) \\
&\overset{(i)}{=} \text{span}(\boldsymbol{\Sigma}^{-1/2}(\sqrt{\lambda_1}\boldsymbol{a}_1, \ldots, \sqrt{\lambda_q}\boldsymbol{a}_q)) \\
&\overset{(ii)}{=} \text{span}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}) \\
&\overset{(iii)}{=} \text{span}(\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}\boldsymbol{P}^{1/2})) \\
&\overset{(iv)}{=} \text{span}(\boldsymbol{\Sigma}^{-1}\boldsymbol{U}) = \text{span}(\boldsymbol{W}^*),
\end{aligned}
$$

where (i) holds since $\lambda_1 \geq \cdots \geq \lambda_q > 0$, (ii) holds since $\text{span}(\sqrt{\lambda_1}\boldsymbol{a}_1, \ldots, \sqrt{\lambda_q}\boldsymbol{a}_q) = \text{span}(\boldsymbol{A})$, (iii) holds since $\text{span}(\boldsymbol{A}) = \text{span}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}\boldsymbol{P}^{1/2})$, and (iv) holds since $\boldsymbol{P}^{1/2}$ is invertible. $\qquad\square$

A.3 Proof of Lemma 4.4

Proof of Lemma 4.4. Taking derivative of the shifted loss function $\widetilde{\mathcal{L}}_1(\boldsymbol{W}, \widehat{\boldsymbol{W}}_t)$ with respect to the first entity at $\boldsymbol{W}^*$, we get

$$
\begin{aligned}
\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t) &= \nabla\mathcal{L}_1(\boldsymbol{W}^*) + \frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\widehat{\boldsymbol{W}}_t) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{W}}_t) \\
&= \frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{W}^*) + \nabla\mathcal{L}_1(\boldsymbol{W}^*) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{W}}_1) + \frac{1}{M}\sum_{m=1}^{M}(\nabla\mathcal{L}_m(\widehat{\boldsymbol{W}}_t) - \nabla\mathcal{L}_m(\boldsymbol{W}^*)),
\end{aligned}
$$

where the second equality is obtained by adding and subtracting a term $\frac{1}{M}\nabla\mathcal{L}_m(\boldsymbol{W}^*)$. By definition of $\mathcal{L}_m(\cdot)$, we have

$$
\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t) = \frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{W}^*) + (\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})(\boldsymbol{W}^* - \widehat{\boldsymbol{W}}_t) + \frac{1}{M}\sum_{m=1}^{M}(\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma})(\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*).
$$

Thus, by the triangle inequality, we have

$$
\begin{aligned}
\|\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)\|_{\infty,\infty} \leq & \|\frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{W}^*)\|_{\infty,\infty} \\
&+ \left(\|\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma}\|_{\infty,\infty} + \frac{1}{M}\sum_{m=1}^{M}\|\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|_{\infty,\infty}\right)\|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1}. \qquad\text{(A.2)}
\end{aligned}
$$

Then we proceed to finish the proof by establishing upper bounds on the $\ell_{\infty,\infty}$ norm of $\frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{W}^*)$ and $\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$. Fisrt, we establish in Lemma A.1 upper bounds on the $\ell_{\infty,\infty}$ norm of the error matrix $\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$. Then in Lemma A.2, we give upper bounds on the $\ell_{\infty,\infty}$ norm of the gradient term $\frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{W}^*)$.

Lemma A.1. *Suppose the assumptions in Lemma 4.4 hold, then there exist some universal constants c_1 and c_2 such that the following bounds*

$$
\|\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|_{\infty,\infty} \leq c_1\sqrt{\frac{\log(dM/\delta)}{n}}, \quad \forall m \in [M], \tag{A.3}
$$

$$
\|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|_{\infty,\infty} \leq c_1\sqrt{\frac{\log(dM/\delta)}{N}}, \tag{A.4}
$$

$$
\|\widehat{\boldsymbol{U}}^m - \boldsymbol{U}\|_{\infty,\infty} \leq c_2\sqrt{\frac{K\log(dMK/\delta)}{n}}, \quad \forall m \in [M], \tag{A.5}
$$

$$
\|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{U}}^m - \boldsymbol{U}\|_{\infty,\infty} \leq c_2\sqrt{\frac{K\log(dMK/\delta)}{N}}, \tag{A.6}
$$

hold with probability at least $1 - 2\delta$.

Lemma A.2. *Assume that the covariance matrix Σ is nonsingular and the upper bounds* (A.3), (A.4), (A.5), *and* (A.6) *in Lemma A.1 hold, then we have for all $m \in [M]$,*

$$\|\nabla\mathcal{L}_m(\boldsymbol{W}^*)\|_{\infty,\infty} \leq c_1\sqrt{\frac{\log(dM/\delta)}{n}}\|\boldsymbol{W}^*\|_{1,1} + c_2\sqrt{\frac{K\log(dMK/\delta)}{n}}, \tag{A.7}$$

$$\|\frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{W}^*)\|_{\infty,\infty} \leq c_1\sqrt{\frac{\log(dM/\delta)}{N}}\|\boldsymbol{W}^*\|_{1,1} + c_2\sqrt{\frac{K\log(dMK/\delta)}{N}}, \tag{A.8}$$

where the constants c_1, c_2 are defined in Lemma A.1.

By Lemma A.1 and A.2, we have with probability at least $1 - 2\delta$ that inequalities (A.3), (A.4), (A.5), (A.6), (A.7) and (A.8) hold for some universal constants c_1 and c_2. Substituting these inequalities into the inequality (A.2), we have

$$\|\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)\|_{\infty,\infty} \leq c_1\sqrt{\frac{\log(dM/\delta)}{N}}\|\boldsymbol{W}^*\|_{1,1} + c_2\sqrt{\frac{K\log(dMK/\delta)}{N}}$$
$$+ 2c_1\sqrt{\frac{\log(dM/\delta)}{n}}\|\boldsymbol{W}^* - \widehat{\boldsymbol{W}}_t\|_{1,1},$$

with probability at least $1 - 2\delta$, which concludes the proof. $\qquad\square$

A.4   Proof of Theorem 4.5

Before giving a proof of Theorem 4.5, we first state two technical lemmas that will be used. Let us begin with the definition of a series of subsets in $\mathbb{R}^{d\times(K-1)}$. For any integer s, we define the subset $\mathcal{H}(s) \subset \mathbb{R}^{d\times(K-1)}$ by

$$\mathcal{H}(s) = \left\{\boldsymbol{\Delta} \in \mathbb{R}^{d\times(K-1)} \big| \boldsymbol{\Delta}_{\cdot k} \in \mathcal{C}(\mathcal{A}, 3), \text{ for some } |\mathcal{A}| \leq s, \forall 1 \leq k \leq K-1\right\},$$

where $\boldsymbol{\Delta}_{\cdot k}$ is the kth column of $\boldsymbol{\Delta}$. Furthermore, given an index set $\mathcal{S} \subset [(d, K-1)]$, we define the cone $\widetilde{\mathcal{C}}(\mathcal{S}, 3)$ in $\mathbb{R}^{d\times(K-1)}$ analogous to $\mathcal{C}(\mathcal{A}, 3)$ in $\mathbb{R}^d$ by

$$\widetilde{\mathcal{C}}(\mathcal{S}, 3) = \left\{\boldsymbol{\Delta} \in \mathbb{R}^{d\times(K-1)} \big| \|\boldsymbol{\Delta}_{\mathcal{S}^c}\|_{1,1} \leq 3\|\boldsymbol{\Delta}_{\mathcal{S}}\|_{1,1}\right\},$$

where $\boldsymbol{\Delta}_{\mathcal{S}} \in \mathbb{R}^{d\times(K-1)}$ is the restriction of $\boldsymbol{\Delta}$ on the set $\mathcal{S}$ and $\boldsymbol{\Delta}_{\mathcal{S}^c}$ is analogously defined. It is shown in Lemma A.3 that the tth error $\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*$ is in the subset $\mathcal{H}(s) \cap \widetilde{\mathcal{C}}(\mathcal{S}, 3)$ under certain conditions, where $s = |\mathcal{S}|$ and $\mathcal{S} = \text{supp}(\boldsymbol{W}^*)$ is the support of $\boldsymbol{W}^*$. In addition, we establish in Lemma A.4 a strong restricted convexity property of $\mathcal{L}_1(\cdot)$, which proves to be useful.

Lemma A.3. *Assume that the upper bound in Lemma 4.4 holds, λ_{t+1} is chosen by* (4.4), *and the cardinality $|\mathcal{S}| = s$, where $\mathcal{S} = \text{supp}(\boldsymbol{W}^*)$ is the support of $\boldsymbol{W}^*$. Then for any t, the $(t+1)$th error $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*$ is in the subset $\mathcal{H}(s) \cap \widetilde{\mathcal{C}}(\mathcal{S}, 3)$.*

Lemma A.4 (Strong Restricted Convexity)**.** *Suppose the conditions in Theorem 4.5 hold, then there exists some constant c_3 such that for $n \geq c_3 s \log(d/\delta)$, the empirical loss function $\mathcal{L}_1$ satisfies the strong restricted convexity property, i.e.,*

$$\mathcal{L}_1(\boldsymbol{W}^* + \boldsymbol{\Delta}) - \mathcal{L}_1(\boldsymbol{W}^*) - \langle\nabla\mathcal{L}_1(\boldsymbol{W}^*), \boldsymbol{\Delta}\rangle \geq \frac{\kappa_0}{4}\|\boldsymbol{\Delta}\|_{2,2}^2, \ \forall\boldsymbol{\Delta} \in \mathcal{H}(s), \tag{A.9}$$

with probability at least $1 - \delta$.

Remark A.5. *By definition of $\mathcal{L}_1(\cdot)$, the strong restricted convexity is reduced to the following inequality,*

$$\langle\boldsymbol{\Delta}, \widehat{\boldsymbol{\Sigma}}^1\boldsymbol{\Delta}\rangle \geq \frac{\kappa_0}{2}\|\boldsymbol{\Delta}\|_{2,2}^2, \ \forall\boldsymbol{\Delta} \in \mathcal{H}(s).$$

As shown in the proof, we prove this inequality by establishing the restricted eigenvalue $RE(\kappa_0/2, s)$ property on the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}^1$. When we are finishing the paper, we find that the restricted eigenvalue property has been established in Rudelson and Zhou (2013). They proved the RE property via a reduction principle while in this paper we prove the RE property directly by applying Proposition 1 in Pilanci and Wainwright (2015).

Proof of Theorem 4.5. By definition of $\widetilde{\mathcal{L}}_1(\cdot, \cdot)$, we have

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{W}}_{t+1}, \widehat{\boldsymbol{W}}_t) - \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)$$
$$= \mathcal{L}_1(\widehat{\boldsymbol{W}}_{t+1}) - \mathcal{L}_1(\boldsymbol{W}^*) + \left\langle \frac{1}{M} \sum_{m=1}^{M} \nabla \mathcal{L}_m(\widehat{\boldsymbol{W}}_t) - \nabla \mathcal{L}_1(\widehat{\boldsymbol{W}}_t), \widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \right\rangle. \tag{A.10}$$

Since Assumption 4.1 and 4.2 hold and $\sup \boldsymbol{\Sigma}_{jj} \leq c_0$, the upper bound in Lemma 4.4 holds for some universal constants c_1 and c_2 with probability at least $1 - 2\delta$. Since λ_{t+1} is chosen by (4.4), the $(t+1)$th error $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \mathcal{H}(s) \cap \widetilde{\mathcal{C}}(\mathcal{S}, 3)$ by Lemma A.3, where $s = |\mathcal{S}|$ and $\mathcal{S} = \text{supp}(\boldsymbol{W}^*)$ is the support of $\boldsymbol{W}^*$. In addition, since Assumption 4.3 also holds, the empirical loss function $\mathcal{L}_1$ satisfies the strong restricted convexity property (A.9) when $n \geq c_3 s \log(d/\delta)$ for some universal constant c_3 with probability at least $1 - \delta$ by Lemma A.4. In the remaining part of analysis, we first assume the conclusions in Lemma 4.4, A.3 and A.4 hold and we will go back to the high-probability language in the end.

Combining the fact that $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \mathcal{H}(s)$ and the strong restricted convexity property (A.9) of $\mathcal{L}_1$, we obtain the following inequality,

$$\mathcal{L}_1(\widehat{\boldsymbol{W}}_{t+1}) - \mathcal{L}_1(\boldsymbol{W}^*) \geq \left\langle \nabla \mathcal{L}_1(\boldsymbol{W}^*), \widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \right\rangle + \frac{\kappa_0}{4} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}^2. \tag{A.11}$$

Substitute this inequality into the equation (A.10), we get

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{W}}_{t+1}, \widehat{\boldsymbol{W}}_t) - \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)$$
$$\geq \left\langle \nabla \mathcal{L}_1(\boldsymbol{W}^*) + \frac{1}{M} \sum_{m=1}^{M} \nabla \mathcal{L}_m(\widehat{\boldsymbol{W}}_t) - \nabla \mathcal{L}_1(\widehat{\boldsymbol{W}}_t), \widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \right\rangle + \frac{\kappa_0}{4} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}^2$$
$$= \left\langle \nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t), \widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \right\rangle + \frac{\kappa_0}{4} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}^2, \tag{A.12}$$

where the equality follows from the definition of $\nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)$. By the optimality (3.2) of $\widehat{\boldsymbol{W}}_{t+1}$, we can derive the following inequality,

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{W}}_{t+1}, \widehat{\boldsymbol{W}}_t) - \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t) + \lambda_{t+1} \|\widehat{\boldsymbol{W}}_{t+1}\|_{1,1} - \lambda_{t+1} \|\boldsymbol{W}^*\|_{1,1} \leq 0. \tag{A.13}$$

Combining the inequalities (A.12) and (A.13), the upper bound in Lemma 4.4, and the definition (4.4) of λ_{t+1}, we get

$$\lambda_{t+1} \|\boldsymbol{W}^*\|_{1,1} - \lambda_{t+1} \|\widehat{\boldsymbol{W}}_{t+1}\|_{1,1} \geq \widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{W}}_{t+1}, \widehat{\boldsymbol{W}}_t) - \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)$$
$$\geq \left\langle \nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t), \widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \right\rangle + \frac{\kappa_0}{4} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}^2$$
$$\geq -\|\nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)\|_{\infty,\infty} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1} + \frac{\kappa_0}{4} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}^2$$
$$\geq -\frac{\lambda_{t+1}}{2} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1} + \frac{\kappa_0}{4} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}^2, \tag{A.14}$$

where the third inequality follows from the Hölder's inequality. Combining this inequality with the fact that $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \widetilde{\mathcal{C}}(\mathcal{S}, 3)$, we obtain that

$$\frac{\kappa_0}{4} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}^2 \leq \lambda_{t+1} \|\boldsymbol{W}^*\|_{1,1} - \lambda_{t+1} \|\widehat{\boldsymbol{W}}_{t+1}\|_{1,1} + \frac{\lambda_{t+1}}{2} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1}$$
$$\leq \frac{3\lambda_{t+1}}{2} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1}$$
$$\leq 6\sqrt{s} \lambda_{t+1} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}, \tag{A.15}$$

where the last inequality holds since $\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1} \leq 4\sqrt{s} \|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}$ when $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \widetilde{\mathcal{C}}(\mathcal{S}, 3)$. Dividing $\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2}$ from both sides of the inequality (A.15), we have

$$\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2} \leq \frac{24\sqrt{s} \lambda_{t+1}}{\kappa_0}. \tag{A.16}$$

Again by the fact that $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \widetilde{\mathcal{C}}(\mathcal{S}, 3)$, we have

$$\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1} \leq 4\sqrt{s}\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2} \leq \frac{96s\lambda_{t+1}}{\kappa_0}. \tag{A.17}$$

Recall the definition (4.4) of λ_{t+1}, the inequality (A.16) and (A.17) are exactly what we want. Finally, let us conclude the proof by noting that this analysis holds for all t and thus the inequalities (A.16) and (A.17) hold for all t with probability $1 - 3\delta$ when $n \geq c_3 s \log(d/\delta)$. $\qquad\square$

A.5 Proof of Theorem 4.6

Proof of Theorem 4.6. Setting $\mathfrak{a} = \frac{192 s a_n}{\kappa_0}$ and $\mathfrak{b} = \frac{192 s b_n}{\kappa_0}$, we can rewrite Theorem 4.5 in the following iteration form,

$$\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1} \leq \mathfrak{a} + \mathfrak{b}\|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1},$$

Applying Theorem 4.5 iteratively, we can obtain the following upper bound on $\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1}$,

$$\begin{aligned}
\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{1,1} &\leq \mathfrak{a} + \mathfrak{b}\|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1} \\
&\leq \mathfrak{a} + \mathfrak{b}\left(\mathfrak{a} + \mathfrak{b}\|\widehat{\boldsymbol{W}}_{t-1} - \boldsymbol{W}^*\|_{1,1}\right) \\
&\leq \cdots \\
&\leq \frac{\mathfrak{a}(1 - \mathfrak{b}^{t+1})}{1 - \mathfrak{b}} + \mathfrak{b}^{t+1}\|\widehat{\boldsymbol{W}}_0 - \boldsymbol{W}^*\|_{1,1}.
\end{aligned} \tag{A.18}$$

Then by Theorem 4.5 again, we can upper bound $\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_{2,2}$ as follows,

$$\begin{aligned}
\|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|_{2,2} &\leq \frac{\mathfrak{a}}{4\sqrt{s}} + \frac{\mathfrak{b}}{4\sqrt{s}}\|\widehat{\boldsymbol{W}}_t - \boldsymbol{W}^*\|_{1,1} \\
&\leq \frac{\mathfrak{a}(1 - \mathfrak{b}^{t+1})}{4\sqrt{s}(1 - \mathfrak{b})} + \frac{\mathfrak{b}^{t+1}}{4\sqrt{s}}\|\widehat{\boldsymbol{W}}_0 - \boldsymbol{W}^*\|_{1,1},
\end{aligned}$$

where the second inequality follows by (A.18). $\qquad\square$

B PROOF OF AUXILIARY LEMMAS

B.1 Proof of Lemma A.1

Proof of Lemma A.1. First, let us upper bound the $\ell_{\infty,\infty}$ norm of $\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$ for a fixed m. By definition of $\widehat{\boldsymbol{\Sigma}}^m$ and Assumption 4.1, we have

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma} ={}& \frac{1}{Kb - K}\sum_{k=1}^{K}\sum_{y_i^m = k}\left\{(\boldsymbol{x}_i^m - \boldsymbol{\mu}_k)(\boldsymbol{x}_i^m - \boldsymbol{\mu}_k)^\top - \boldsymbol{\Sigma}\right\} \\
&- \frac{1}{Kb - K}\sum_{k=1}^{K}\left\{b(\bar{\boldsymbol{x}}^{mk} - \boldsymbol{\mu}_k)(\bar{\boldsymbol{x}}^{mk} - \boldsymbol{\mu}_k)^\top - \boldsymbol{\Sigma}\right\},
\end{aligned}$$

where $\bar{\boldsymbol{x}}^{mk} = \frac{1}{b}\sum_{y_i^m = k}\boldsymbol{x}_i^m$. For convenience, we introduce the following transformed random vectors,

$$\begin{aligned}
\widetilde{\boldsymbol{z}}_i^m &= \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x}_i^m - \boldsymbol{\mu}_k), \quad \text{if } y_i^m = k, \quad \forall i, \\
\widetilde{\boldsymbol{z}}^{mk} &= \sqrt{b}\boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{x}}^{mk} - \boldsymbol{\mu}_k), \quad \forall k.
\end{aligned}$$

Note that conditional on $\{y_i^m\}_{i=1}^n$, $\{\widetilde{\boldsymbol{z}}_i^m\}_{i=1}^n$ are i.i.d. zero mean σ-sub-Gaussian random vectors and $\{\widetilde{\boldsymbol{z}}^{mk}\}_{k=1}^K$ are i.i.d. zero mean σ-sub-Gaussian random vectors, though $\{\widetilde{\boldsymbol{z}}_i^m\}_{i=1}^n$ and $\{\widetilde{\boldsymbol{z}}^{mk}\}_{k=1}^K$ can be dependent. Using these transformed random vectors, we can rewrite the error $\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$ as

$$\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma} = \frac{Kb}{Kb - K}\boldsymbol{A} - \frac{K}{Kb - K}\boldsymbol{A}',$$

where

$$\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\Sigma}^{1/2}, \tag{B.1}$$

$$\boldsymbol{A}' = \boldsymbol{\Sigma}^{1/2} \left(\frac{1}{K} \sum_{k=1}^{K} \widetilde{\boldsymbol{z}}^{mk} \widetilde{\boldsymbol{z}}^{mk\top} - \boldsymbol{I}_d \right) \boldsymbol{\Sigma}^{1/2}. \tag{B.2}$$

By the triangle inequality, we have

$$\|\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|_{\infty,\infty} \leq \frac{Kb}{Kb - K} \|\boldsymbol{A}\|_{\infty,\infty} + \frac{K}{Kb - K} \|\boldsymbol{A}'\|_{\infty,\infty}. \tag{B.3}$$

In the remaining part of the proof, we will apply Proposition 1 in Pilanci and Wainwright (2015) to upper bound the $\ell_{\infty,\infty}$ norm of $\boldsymbol{A}$ and $\boldsymbol{A}'$ separately, which leads to an upper bound on the $\ell_{\infty,\infty}$ norm of $\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$ by (B.3). For reader's convenience, we present this proposition below.

Proposition B.1. *Let $\{\boldsymbol{s}_i\}_{i=1}^{n} \subset \mathbb{R}^d$ be i.i.d. samples from a zero-mean σ-sub-Gaussian distribution with $\mathrm{Cov}(\boldsymbol{s}_i) = \boldsymbol{I}_d$. Then there exist some universal constants c_1, c_2 such that for any subset $\mathcal{Y} \subset \mathbb{S}^{d-1}$, we have with probability at least $1 - e^{-\frac{c_2 n \delta^2}{\sigma^4}}$,*

$$\sup_{\boldsymbol{\eta} \in \mathcal{Y}} \left| \boldsymbol{\eta}^\top \left(\frac{\boldsymbol{S}^\top \boldsymbol{S}}{n} - \boldsymbol{I}_d \right) \boldsymbol{\eta} \right| \leq c_1 \frac{\mathbb{W}(\mathcal{Y})}{\sqrt{n}} + \delta,$$

where $\boldsymbol{S}^\top = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ and $\mathbb{W}(\mathcal{Y})$ is the Gaussian width of the subset $\mathcal{Y}$. Specifically, $\mathbb{W}(\mathcal{Y})$ is defined by

$$\mathbb{W}(\mathcal{Y}) = \mathbb{E}[\sup_{\boldsymbol{\eta} \in \mathcal{Y}} |\langle \boldsymbol{g}, \boldsymbol{\eta} \rangle|],$$

where the expectation is taken on $\boldsymbol{g} \in \mathbb{R}^d$, which is a standard normal random vector.

Let us first bound the $\ell_{\infty,\infty}$ norm of $\boldsymbol{A}$ by applying Proposition B.1 to $\{\boldsymbol{s}_i\}_{i=1}^{n} = \{\widetilde{\boldsymbol{z}}_i^m\}_{i=1}^{n}$ and $\mathcal{Y} = \mathcal{Y}_1 \subset \mathbb{S}^{d-1}$ defined by

$$\mathcal{Y}_1 = \left\{ \boldsymbol{\beta}_j = \frac{\boldsymbol{\Sigma}_{\cdot j}^{1/2}}{\|\boldsymbol{\Sigma}_{\cdot j}^{1/2}\|_2} \right\}_{1 \leq j \leq d} \bigcup \left\{ \boldsymbol{\gamma}_{jk} = \frac{\boldsymbol{\beta}_j + \boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_j + \boldsymbol{\beta}_k\|_2} \right\}_{1 \leq j < k \leq d},$$

where $\boldsymbol{\Sigma}_{\cdot j}^{1/2}$ is the jth column of the matrix $\boldsymbol{\Sigma}^{1/2}$. Note that the subset $\mathcal{Y}_1$ is a finite set with at most $\frac{d(d+1)}{2}$ elements. Since $\{\widetilde{\boldsymbol{z}}_i^m\}_{i=1}^{n}$ are i.i.d. samples from a zero mean σ-sub-Gaussian distribution with $\mathrm{Cov}(\widetilde{\boldsymbol{z}}_i^m) = \boldsymbol{I}_d$, we have with probability at least $1 - e^{-\frac{c_2 n \delta^2}{\sigma^4}}$,

$$\sup_{\boldsymbol{\eta} \in \mathcal{Y}_1} \left| \boldsymbol{\eta}^\top \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\eta} \right| \leq c_1 \frac{\mathbb{W}(\mathcal{Y}_1)}{\sqrt{n}} + \delta, \tag{B.4}$$

for some universal constants c_1 and c_2. By definition of $\boldsymbol{A}$, we have

$$\begin{aligned}
\|\boldsymbol{A}\|_{\infty,\infty} &= \sup_{1 \leq j \leq k \leq d} |\boldsymbol{A}_{jk}| = \sup_{1 \leq j \leq k \leq d} \left| \boldsymbol{\Sigma}_{\cdot j}^{1/2\top} \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\Sigma}_{\cdot k}^{1/2} \right| \\
&= \sup_{1 \leq j \leq k \leq d} \left| \boldsymbol{\beta}_j^\top \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\beta}_k \right| \cdot (\boldsymbol{\Sigma}_{jj} \boldsymbol{\Sigma}_{kk})^{1/2} \\
&\overset{(i)}{\leq} c_0 \cdot \sup_{1 \leq j \leq k \leq d} \left| \boldsymbol{\beta}_j^\top \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\beta}_k \right|,
\end{aligned}$$

where the inequality (i) follows from the condition $\sup_{1 \leq j \leq d} \boldsymbol{\Sigma}_{jj} \leq c_0$. By the polarization identity, we have

$$
\boldsymbol{\beta}_j^\top \left(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\beta}_k
$$

$$
= \frac{1}{2} (\boldsymbol{\beta}_j + \boldsymbol{\beta}_k)^\top \left(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) (\boldsymbol{\beta}_j + \boldsymbol{\beta}_k)
$$

$$
- \frac{1}{2} \boldsymbol{\beta}_j^\top \left(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\beta}_j - \frac{1}{2} \boldsymbol{\beta}_k^\top \left(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\beta}_k.
$$

Then by the triangle inequality, we have

$$
\|\boldsymbol{A}\|_{\infty,\infty} \leq c_0 \cdot \sup_{1 \leq j \leq k \leq d} \left| \boldsymbol{\beta}_j^\top \left(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\beta}_k \right|
$$

$$
\leq c_0 \cdot \frac{1}{2} \cdot \|\boldsymbol{\beta}_j + \boldsymbol{\beta}_k\|_2^2 \cdot \sup_{1 \leq j \leq k \leq d} \left| \boldsymbol{\gamma}_{jk}^\top \left(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\gamma}_{jk} \right|
$$

$$
+ c_0 \cdot \sup_{1 \leq j \leq d} \left| \boldsymbol{\beta}_j^\top \left(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\beta}_j \right|
$$

$$
\leq 3 c_0 \cdot \sup_{\boldsymbol{\eta} \in \mathcal{Y}_1} \left| \boldsymbol{\eta}^\top \left(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\eta} \right|,
$$

where the last inequality holds since $\|\boldsymbol{\beta}_j + \boldsymbol{\beta}_k\|_2^2 \leq 4$ and $\{\boldsymbol{\beta}_j, \boldsymbol{\gamma}_{jk}\}_{1 \leq j \leq k \leq d} \subset \mathcal{Y}_1$. Combining this inequality with the upper bound in (B.4), we can upper bound the $\ell_{\infty,\infty}$ norm of $\boldsymbol{A}$ in terms of the Gaussian width $\mathbb{W}(\mathcal{Y}_1)$,

$$
\|\boldsymbol{A}\|_{\infty,\infty} \leq 3 c_0 \cdot \sup_{\boldsymbol{\eta} \in \mathcal{Y}_1} \left| \boldsymbol{\eta}^\top \left(\frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{z}}_i^m \widetilde{\boldsymbol{z}}_i^{m\top} - \boldsymbol{I}_d \right) \boldsymbol{\eta} \right|
$$

$$
\leq 3 c_0 \left(c_1 \frac{\mathbb{W}(\mathcal{Y}_1)}{\sqrt{n}} + \delta \right),
$$

with probability at least $1 - e^{-\frac{c_2 n \delta^2}{\sigma^4}}$. Now we give an upper bound on the Gaussian width $\mathbb{W}(\mathcal{Y}_1)$ using the maximal inequality Wainwright (2019). Since the random variable $\langle \boldsymbol{g}, \boldsymbol{\eta} \rangle$ is Gaussian with zero mean and variance 1 for any $\boldsymbol{\eta} \in \mathcal{Y}_1 \subset \mathbb{S}^{d-1}$, and the cardinality of $\mathcal{Y}_1$ is finite less than $\frac{d(d+1)}{2}$, we have

$$
\mathbb{W}(\mathcal{Y}_1) = \mathbb{E}[\sup_{\boldsymbol{\eta} \in \mathcal{Y}_1} |\langle \boldsymbol{g}, \boldsymbol{\eta} \rangle|] \leq \sqrt{2 \log(d(d+1))}.
$$

Therefore, we have with probability at least $1 - e^{-\frac{c_2 n \delta^2}{\sigma^4}}$,

$$
\|\boldsymbol{A}\|_{\infty,\infty} \leq c_0 \cdot (c_1 \sqrt{\frac{\log(d)}{n}} + \delta), \tag{B.5}
$$

for some universal constants c_1 and c_2. Applying the same argument to $\boldsymbol{A}'$, we obtain the following upper bound

$$
\|\boldsymbol{A}'\|_{\infty,\infty} \leq c_0 \cdot (c_1 \sqrt{\frac{\log(d)}{K}} + \delta), \tag{B.6}
$$

with probability at least $1 - e^{-\frac{c_2 K \delta^2}{\sigma^2}}$. Combining inequalities (B.3), (B.5), and (B.6), and the observation that $\boldsymbol{A}'$-term is negligible in $\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$ relative to $\boldsymbol{A}$-term, we have with probability at least $1 - e^{-\frac{c_2 n \delta^2}{\sigma^4}}$,

$$
\|\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|_{\infty,\infty} \leq \frac{b}{b-1} \|\boldsymbol{A}\|_{\infty,\infty} + \frac{1}{b-1} \|\boldsymbol{A}'\|_{\infty,\infty} \leq c_0 \cdot c_1 (\sqrt{\frac{\log(d)}{n}} + \delta), \tag{B.7}
$$

for some universal constants c_1, c_2. Furthermore, we can bound the $\ell_{\infty,\infty}$ norm of $\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$ in the same way. In specific, with probability at least $1 - e^{-\frac{c_2 n \delta^2}{\sigma^4}}$, we have

$$\|\!|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|\!|_{\infty,\infty} \le c_0 \cdot c_1 (\sqrt{\frac{\log(d)}{N}} + \frac{\delta}{\sqrt{M}}), \tag{B.8}$$

for some constants c_1, c_2. Using the method of union bound for (B.7) for all $m \in [M]$ and (B.8) and the change of variable of δ, we have

$$\|\!|\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|\!|_{\infty,\infty} \le c_1 \sqrt{\frac{\log(dM/\delta)}{n}}, \quad \forall m,$$

$$\|\!|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|\!|_{\infty,\infty} \le c_1 \sqrt{\frac{\log(dM/\delta)}{N}}, \tag{B.9}$$

for some universal constant c_1, with probability at least $1 - \delta$.

Next, we derive an upper bound on the $\ell_{\infty,\infty}$ norm of the error $\widehat{\boldsymbol{U}}^m - \boldsymbol{U}$ for a fixed m. By Assumption 4.1, we can rewrite $\boldsymbol{U}_{\cdot k}$ in the following form

$$\boldsymbol{U}_{\cdot k} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)\boldsymbol{h}^k, \quad \forall k \in \{1, \ldots, K-1\},$$

where the vector $\boldsymbol{h}^k \in \mathbb{R}^K$ satisfies

$$\boldsymbol{h}_k^k = 1 - \frac{1}{K}, \boldsymbol{h}_j^k = -\frac{1}{K}, \forall j \ne k.$$

Using $\boldsymbol{h}^k$, we can express elements of $\widehat{\boldsymbol{U}}^m - \boldsymbol{U}$ as

$$[\widehat{\boldsymbol{U}}^m - \boldsymbol{U}]_{jk} = \widehat{\boldsymbol{U}}_{jk}^m - \boldsymbol{U}_{jk} = \widetilde{\boldsymbol{X}}_{j\cdot}^m \boldsymbol{h}^k - \boldsymbol{U}_{jk},$$

where $\widetilde{\boldsymbol{X}}^m = (\bar{\boldsymbol{x}}^{m1}, \ldots, \bar{\boldsymbol{x}}^{mK}) \in \mathbb{R}^{d \times K}$ is the standard unbiased estimator of $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$. Note that the (j,k)th element of $\widetilde{\boldsymbol{X}}^m$ is a linear combination of coordinates of $\boldsymbol{x}_i^m$ and thus $\widetilde{\boldsymbol{X}}_{jk}^m$ is sub-Gaussian with parameter

$$\sigma_{jk}^2 = \frac{\sigma^2 \boldsymbol{\Sigma}_{jj}}{b} \le \widetilde{\sigma}^2 := \frac{\sigma^2 c_0}{b},$$

where the inequality follows from the condition $\sup_j \boldsymbol{\Sigma}_{jj} \le c_0$. Furthermore, columns of $\widetilde{\boldsymbol{X}}^m$ are independent random vectors. Therefore, $[\widehat{\boldsymbol{U}}^m - \boldsymbol{U}]_{jk}$ is zero-mean sub-Gaussian variable with parameter

$$\|\boldsymbol{h}^k\|_2 \cdot \widetilde{\sigma} = \frac{\sqrt{K-1}}{K}\widetilde{\sigma} \le \widetilde{\sigma}.$$

Then by the maximal inequality, we have

$$\|\!|\widehat{\boldsymbol{U}}^m - \boldsymbol{U}\|\!|_{\infty,\infty} = \sup_{1 \le j \le d, 1 \le k \le K-1} \left|\widehat{\boldsymbol{U}}_{jk}^m - \boldsymbol{U}_{jk}\right| \le \delta, \tag{B.10}$$

with probability at least $1 - 2dKe^{-\frac{\delta^2 b}{2c_0 \sigma^2}}$. Applying the same argument to $\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{U}}^m - \boldsymbol{U}$, we have

$$\|\!|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{U}}^m - \boldsymbol{U}\|\!|_{\infty,\infty} \le \frac{\delta}{\sqrt{M}}, \tag{B.11}$$

with probability at least $1 - 2dKe^{-\frac{\delta^2 b}{2c_0 \sigma^2}}$. Using the method of union bound for (B.10) for all m and (B.11) and the change of variable of δ, we have with probability at least $1 - \delta$,

$$\|\!|\widehat{\boldsymbol{U}}^m - \boldsymbol{U}\|\!|_{\infty,\infty} \le c_2 \sqrt{\frac{K\log(dMK/\delta)}{n}}, \quad \forall m,$$

$$\|\!|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{U}}^m - \boldsymbol{U}\|\!|_{\infty,\infty} \le c_2 \sqrt{\frac{K\log(dMK/\delta)}{N}}, \tag{B.12}$$

for some universal constant c_2.

Finally, we conclude our proof by using the method of union bound for (B.9) and (B.12). $\qquad \square$

B.2 Proof of Lemma A.2

Proof of Lemma A.2. By definition of $\mathcal{L}_m(\cdot)$ and $\boldsymbol{W}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{U}$, the gradient of $\mathcal{L}_m(\cdot)$ at the point $\boldsymbol{W}^*$ is

$$\nabla\mathcal{L}_m(\boldsymbol{W}^*) = \widehat{\boldsymbol{\Sigma}}^m\boldsymbol{W}^* - \widehat{\boldsymbol{U}}^m = (\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma})\boldsymbol{W}^* + (\boldsymbol{U} - \widehat{\boldsymbol{U}}^m).$$

Then by the triangle inequality, we obtain the following inequalities,

$$\|\nabla\mathcal{L}_m(\boldsymbol{W}^*)\|_{\infty,\infty} \leq \|\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|_{\infty,\infty}\|\boldsymbol{W}^*\|_{1,1} + \|\widehat{\boldsymbol{U}}^m - \boldsymbol{U}\|_{\infty,\infty},$$

$$\|\frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{W}^*)\|_{\infty,\infty} \leq \|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}\|_{\infty,\infty}\|\boldsymbol{W}^*\|_{1,1} + \|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{U}}^m - \boldsymbol{U}\|_{\infty,\infty}$$

Substituting the upper bounds (A.3), (A.4), (A.5), and (A.6) in Lemma A.1 into the previous two inequalities, we have for all $m \in [M]$,

$$\|\nabla\mathcal{L}_m(\boldsymbol{W}^*)\|_{\infty,\infty} \leq c_1\sqrt{\frac{\log(dM/\delta)}{n}}\|\boldsymbol{W}^*\|_{1,1} + c_2\sqrt{\frac{K\log(dMK/\delta)}{n}},$$

$$\|\frac{1}{M}\sum_{m=1}^{M}\nabla\mathcal{L}_m(\boldsymbol{W}^*)\|_{\infty,\infty} \leq c_1\sqrt{\frac{\log(dM/\delta)}{N}}\|\boldsymbol{W}^*\|_{1,1} + c_2\sqrt{\frac{K\log(dMK/\delta)}{N}},$$

where the constants c_1, c_2 are defined in Lemma A.1. $\qquad\square$

B.3 Proof of Lemma A.3

Proof of Lemma A.3. We prove this lemma in two steps: first, we show that $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \mathcal{H}(s)$; second, we show that $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \widetilde{\mathcal{C}}(\mathcal{S}, 3)$, where $\mathcal{S} = \text{supp}(\boldsymbol{W}^*)$ and $s = |\mathcal{S}|$.

Step 1: $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \mathcal{H}(s)$. Before we give an proof, let us introduce some notations first. We use $\widehat{\boldsymbol{w}}_k^t$ and $\boldsymbol{w}_k^*$ to denote the kth column of the matrix $\widehat{\boldsymbol{W}}_t$ and $\boldsymbol{W}^*$, respectively. Then we use $\mathcal{S}_k = \text{supp}(\boldsymbol{w}_k^*)$ to denote the support of the kth column of $\boldsymbol{W}^*$. Note that the cardinality of $\mathcal{S}_k$ is smaller than that of $\mathcal{S} = \text{supp}(\boldsymbol{W}^*)$, which is s.

For each k and t, we define the following transformed loss function

$$\widetilde{\ell}_{1k}(\boldsymbol{w}, \widehat{\boldsymbol{w}}_k^t) = \ell_{1k}(\boldsymbol{w}) + \left\langle \frac{1}{M}\sum_{m=1}^{M}\nabla\ell_{mk}(\widehat{\boldsymbol{w}}_k^t) - \nabla\ell_{1k}(\widehat{\boldsymbol{w}}_k^t), \boldsymbol{w} \right\rangle,$$

where

$$\ell_{mk}(\boldsymbol{w}) = \frac{1}{2}\left\langle \boldsymbol{w}, \widehat{\boldsymbol{\Sigma}}^m\boldsymbol{w} \right\rangle - \left\langle \widehat{\boldsymbol{U}}_{\cdot k}^m, \boldsymbol{w} \right\rangle.$$

Since the loss function $\widetilde{\mathcal{L}}_1(\boldsymbol{W}, \widehat{\boldsymbol{W}}_t)$ and the $\ell_{1,1}$ norm of $\boldsymbol{W}$ can be decomposed in the form of

$$\widetilde{\mathcal{L}}_1(\boldsymbol{W}, \widehat{\boldsymbol{W}}_t) = \sum_{k=1}^{K-1}\widetilde{\ell}_{1k}(\boldsymbol{w}_k, \widehat{\boldsymbol{w}}_k^t),$$

$$\|\boldsymbol{W}\|_{1,1} = \sum_{k=1}^{K-1}\|\boldsymbol{w}_k\|_1,$$

where $\boldsymbol{w}_k = \boldsymbol{W}_{\cdot k}$ is the kth column of $\boldsymbol{W}$, the optimization problem

$$\widehat{\boldsymbol{W}}_{t+1} = \underset{\boldsymbol{W}\in\mathbb{R}^{d\times(K-1)}}{\text{argmin}} \left\{ \widetilde{\mathcal{L}}_1(\boldsymbol{W}, \widehat{\boldsymbol{W}}_t) + \lambda_{t+1}\|\boldsymbol{W}\|_{1,1} \right\}$$

is equivalent to the following $K - 1$ minimization problems

$$\widehat{\boldsymbol{w}}_k^{t+1} = \underset{\boldsymbol{w}_k\in\mathbb{R}^d}{\text{argmin}} \left\{ \widetilde{\ell}_{1k}(\boldsymbol{w}_k, \widehat{\boldsymbol{w}}_k^t) + \lambda_{t+1}\|\boldsymbol{w}_k\|_1 \right\}, \quad \forall 1 \leq k \leq K-1. \tag{B.13}$$

Moreover, we have the following inequality

$$\|\nabla \widetilde{\ell}_{1k}(\boldsymbol{w}_k^*, \widehat{\boldsymbol{w}}_k^t)\|_\infty \leq \|\nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)\|_{\infty,\infty} \leq \frac{\lambda_{t+1}}{2},$$

since $\nabla \widetilde{\ell}_{1k}(\boldsymbol{w}_k^*, \widehat{\boldsymbol{w}}_k^t)$ is the kth column of $\nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)$ and λ_{t+1} is defined by (4.4).

Now let us prove the lemma by anaylzing the kth column of $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*$ with k and t fixed. By the triangle inequality, we have

$$\begin{aligned} \|\widehat{\boldsymbol{w}}_k^{t+1}\|_1 - \|\boldsymbol{w}_k^*\|_1 &= \|\boldsymbol{w}_k^* + (\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k^c} + (\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k}\|_1 - \|\boldsymbol{w}_k^*\|_1 \\ &\geq \|\boldsymbol{w}_k^* + (\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k^c}\|_1 - \|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k}\|_1 - \|\boldsymbol{w}_k^*\|_1 \\ &= \|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k^c}\|_1 - \|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k}\|_1, \end{aligned}$$

where $\mathcal{S}_k = \text{supp}(\boldsymbol{w}_k^*)$. By the optimality (B.13) of $\widehat{\boldsymbol{w}}_k^{t+1}$, we have

$$\widetilde{\ell}_{1k}(\widehat{\boldsymbol{w}}_k^{t+1}, \widehat{\boldsymbol{w}}_k^t) + \lambda_{t+1}\|\widehat{\boldsymbol{w}}_k^{t+1}\|_1 - \widetilde{\ell}_{1k}(\boldsymbol{w}_k^*, \widehat{\boldsymbol{w}}_k^t) - \lambda_{t+1}\|\boldsymbol{w}_k^*\|_1 \leq 0.$$

Combining the above two inequalities, we get

$$\widetilde{\ell}_{1k}(\widehat{\boldsymbol{w}}_k^{t+1}, \widehat{\boldsymbol{w}}_k^t) - \widetilde{\ell}_{1k}(\boldsymbol{w}_k^*, \widehat{\boldsymbol{w}}_k^t) + \lambda_{t+1}(\|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k^c}\|_1 - \|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k}\|_1) \leq 0.$$

By the convexity of $\widetilde{\ell}_{1k}(\cdot, \widehat{\boldsymbol{w}}_k^t)$ we have

$$\widetilde{\ell}_{1k}(\widehat{\boldsymbol{w}}_k^{t+1}, \widehat{\boldsymbol{w}}_k^t) - \widetilde{\ell}_{1k}(\boldsymbol{w}_k^*, \widehat{\boldsymbol{w}}_k^t) \geq \left\langle \nabla \widetilde{\ell}_{1k}(\boldsymbol{w}_k^*, \widehat{\boldsymbol{w}}_k^t), \widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^* \right\rangle.$$

Therefore,

$$\begin{aligned} 0 &\geq \left\langle \nabla \widetilde{\ell}_{1k}(\boldsymbol{w}_k^*, \widehat{\boldsymbol{w}}_k^t), \widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^* \right\rangle + \lambda_{t+1}(\|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k^c}\|_1 - \|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k}\|_1) \\ &\geq -\|\nabla \widetilde{\ell}_{1k}(\boldsymbol{w}_k^*, \widehat{\boldsymbol{w}}_k^t)\|_\infty \|\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*\|_1 + \lambda_{t+1}(\|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k^c}\|_1 - \|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k}\|_1). \end{aligned}$$

Since $\frac{\lambda_{t+1}}{2} \geq \|\nabla \widetilde{\ell}_{1k}(\boldsymbol{w}_k^*, \widehat{\boldsymbol{w}}_k^t)\|_\infty$, we have

$$\begin{aligned} 0 &\geq -\frac{1}{2}\|\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*\|_1 + \|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k^c}\|_1 - \|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k}\|_1 \\ &= \frac{1}{2}\|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k^c}\|_1 - \frac{3}{2}\|(\widehat{\boldsymbol{w}}_k^{t+1} - \boldsymbol{w}_k^*)_{\mathcal{S}_k}\|_1, \end{aligned}$$

which proves that $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \mathcal{H}(s)$ since $|\mathcal{S}_k| \leq |\mathcal{S}| = s$.

Step 2: $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \widetilde{\mathcal{C}}(\mathcal{S}, 3)$. The proof of this step is analogous to that of the first step. By the triangle inequality, we have

$$\begin{aligned} \|\widehat{\boldsymbol{W}}_{t+1}\|_{1,1} - \|\boldsymbol{W}^*\|_{1,1} &= \|\boldsymbol{W}^* + (\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}^c} + (\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}}\|_{1,1} - \|\boldsymbol{W}^*\|_{1,1} \\ &\geq \|\boldsymbol{W}^* + (\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}^c}\|_{1,1} - \|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}_k}\|_{1,1} - \|\boldsymbol{W}^*\|_{1,1} \\ &= \|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}^c}\|_{1,1} - \|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}}\|_{1,1}, \end{aligned}$$

where $\mathcal{S} = \text{supp}(\boldsymbol{W}^*)$. By the optimality of $\widehat{\boldsymbol{W}}_{t+1}$, we have

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{W}}_{t+1}, \widehat{\boldsymbol{W}}_t) + \lambda_{t+1}\|\widehat{\boldsymbol{W}}_{t+1}\|_{1,1} - \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t) - \lambda_{t+1}\|\boldsymbol{W}^*\|_{1,1} \leq 0.$$

Combining the above two inequalities, we get

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{W}}_{t+1}, \widehat{\boldsymbol{W}}_t) - \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t) + \lambda_{t+1}(\|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}^c}\|_{1,1} - \|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}}\|_{1,1}) \leq 0.$$

By the convexity of $\widetilde{\mathcal{L}}_1(\cdot, \widehat{\boldsymbol{W}}_t)$ we have

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{W}}_{t+1}, \widehat{\boldsymbol{W}}_t) - \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t) \geq \left\langle \nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t), \widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \right\rangle.$$

Therefore,

$$0 \geq \left\langle \nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t), \widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \right\rangle + \lambda_{t+1}(\|\!|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}^c}\|\!|_{1,1} - \|\!|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}}\|\!|_{1,1})$$

$$\geq -\|\!|\nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)\|\!|_{\infty,\infty}\|\!|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|\!|_{1,1} + \lambda_{t+1}(\|\!|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}^c}\|\!|_{1,1} - \|\!|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}}\|\!|_{1,1}).$$

Since $\frac{\lambda_{t+1}}{2} \geq \|\!|\nabla \widetilde{\mathcal{L}}_1(\boldsymbol{W}^*, \widehat{\boldsymbol{W}}_t)\|\!|_{\infty,\infty}$, we have

$$0 \geq -\frac{1}{2}\|\!|\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*\|\!|_{1,1} + \|\!|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}^c}\|\!|_{1,1} - \|\!|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}}\|\!|_{1,1}$$

$$= \frac{1}{2}\|\!|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}^c}\|\!|_{1,1} - \frac{3}{2}\|\!|(\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^*)_{\mathcal{S}}\|\!|_{1,1},$$

which proves that $\widehat{\boldsymbol{W}}_{t+1} - \boldsymbol{W}^* \in \widetilde{\mathcal{C}}(\mathcal{S}, 3)$. □

B.4 Proof of Lemma A.4

Proof of Lemma A.4. By definition of $\mathcal{L}_1(\cdot)$, we have for each $\boldsymbol{\Delta} \in \mathcal{H}(s)$,

$$\mathcal{L}_1(\boldsymbol{W}^* + \boldsymbol{\Delta}) - \mathcal{L}_1(\boldsymbol{W}^*) - \langle \nabla \mathcal{L}_1(\boldsymbol{W}^*), \boldsymbol{\Delta} \rangle = \frac{1}{2}\left\langle \boldsymbol{\Delta}, \widehat{\boldsymbol{\Sigma}}^1 \boldsymbol{\Delta} \right\rangle.$$

$$= \frac{1}{2}\sum_{j=1}^{K-1} \boldsymbol{\Delta}_{\cdot j}^{\top}\boldsymbol{\Sigma}\boldsymbol{\Delta}_{\cdot j} \cdot \left\{ 1 + \frac{\boldsymbol{\Delta}_{\cdot j}^{\top}(\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})\boldsymbol{\Delta}_{\cdot j}}{\boldsymbol{\Delta}_{\cdot j}^{\top}\boldsymbol{\Sigma}\boldsymbol{\Delta}_{\cdot j}} \right\}, \tag{B.14}$$

where $\boldsymbol{\Delta}_{\cdot j}$ is the jth column of $\boldsymbol{\Delta}$. In the remain part of proof, we first establish an upper bound on the following term

$$\sup_{\boldsymbol{\Delta} \in \mathcal{H}(s), 1 \leq j \leq K-1} \frac{\left|\boldsymbol{\Delta}_{\cdot j}^{\top}(\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})\boldsymbol{\Delta}_{\cdot j}\right|}{\boldsymbol{\Delta}_{\cdot j}^{\top}\boldsymbol{\Sigma}\boldsymbol{\Delta}_{\cdot j}},$$

using Proposition B.1 and the $\text{RE}(\kappa_0, s)$ condition of $\boldsymbol{\Sigma}$, then we prove the strong restricted convexity of $\mathcal{L}_1$ by the equation (B.14) and the $\text{RE}(\kappa_0, s)$ condition of $\boldsymbol{\Sigma}$.

By definition of $\mathcal{H}(s)$, we have

$$\sup_{\boldsymbol{\Delta} \in \mathcal{H}(s), 1 \leq j \leq K-1} \frac{\left|\boldsymbol{\Delta}_{\cdot j}^{\top}(\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})\boldsymbol{\Delta}_{\cdot j}\right|}{\boldsymbol{\Delta}_{\cdot j}^{\top}\boldsymbol{\Sigma}\boldsymbol{\Delta}_{\cdot j}} \leq \sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A}, 3), |\mathcal{A}| \leq s} \frac{\left|\boldsymbol{\zeta}^{\top}(\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})\boldsymbol{\zeta}\right|}{\boldsymbol{\zeta}^{\top}\boldsymbol{\Sigma}\boldsymbol{\zeta}}.$$

Similar to the proof of Lemma A.1, by definition of $\widehat{\boldsymbol{\Sigma}}^1$ and the triangle inequality, we have

$$\left|\boldsymbol{\zeta}^{\top}(\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})\boldsymbol{\zeta}\right| \leq \frac{Kb}{Kb - K}\left|\boldsymbol{\zeta}^{\top}\boldsymbol{A}\boldsymbol{\zeta}\right| + \frac{K}{Kb - K}\left|\boldsymbol{\zeta}^{\top}\boldsymbol{A}'\boldsymbol{\zeta}\right|, \tag{B.15}$$

where $\boldsymbol{A}$ and $\boldsymbol{A}'$ are defined in (B.1) and (B.2) for $m = 1$. By Proposition B.1, for some constants c_1 and c_2, we have

$$\sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A}, 3), |\mathcal{A}| \leq s} \frac{\left|\boldsymbol{\zeta}^{\top}\boldsymbol{A}\boldsymbol{\zeta}\right|}{\boldsymbol{\zeta}^{\top}\boldsymbol{\Sigma}\boldsymbol{\zeta}} = \sup_{\boldsymbol{\eta} = \frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}\|_2}, \boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A}, 3), |\mathcal{A}| \leq s} \left|\boldsymbol{\eta}^{\top}(\frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{z}}_i^1 \widetilde{\boldsymbol{z}}_i^{1\top} - \boldsymbol{I}_d)\boldsymbol{\eta}\right|$$

$$\leq c_1 \frac{\mathbb{W}(\mathcal{Y}_2)}{\sqrt{n}} + \delta, \tag{B.16}$$

with probability at least $1 - e^{-\frac{c_2 n \delta^2}{\sigma^4}}$, where

$$\mathcal{Y}_2 = \left\{ \boldsymbol{\eta} = \frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}\|_2} : \boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A}, 3), |\mathcal{A}| \leq s \right\} \subset \mathbb{S}^{d-1}.$$

Now let us upper bound the Gaussian width $\mathbb{W}(\mathcal{Y}_2)$. The Gaussian width $\mathbb{W}(\mathcal{Y}_2)$ is

$$\mathbb{W}(\mathcal{Y}_2) = \mathbb{E}[\sup_{\boldsymbol{\eta} \in \mathcal{Y}_2} |\langle \boldsymbol{g}, \boldsymbol{\eta} \rangle|] = \mathbb{E}[\sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A},3,), |\mathcal{A}| \leq s, \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}\|_2=1} |\langle \boldsymbol{g}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta} \rangle|].$$

Then by Hölder's inequality, we have

$$\mathbb{W}(\mathcal{Y}_2) \leq \mathbb{E}[\sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A},3), |\mathcal{A}| \leq s, \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}\|_2=1} \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{g}\|_\infty \|\boldsymbol{\zeta}\|_1]$$

$$\overset{(i)}{\leq} \mathbb{E}[\sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A},3), |\mathcal{A}| \leq s, \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}\|_2=1} \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{g}\|_\infty \|\boldsymbol{\zeta}\|_2] \cdot 4\sqrt{s}$$

$$\overset{(ii)}{\leq} \mathbb{E}[\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{g}\|_\infty] \cdot 4\sqrt{\frac{s}{\kappa_0}},$$

where the inequality (i) follows from the property

$$\|\boldsymbol{\zeta}\|_1 = \|\boldsymbol{\zeta}_\mathcal{A}\|_1 + \|\boldsymbol{\zeta}_{\mathcal{A}^c}\|_1 \leq 4\|\boldsymbol{\zeta}_\mathcal{A}\|_1 \leq 4\sqrt{|\mathcal{A}|}\|\boldsymbol{\zeta}_\mathcal{A}\|_2 \leq 4\sqrt{s}\|\boldsymbol{\zeta}\|_2,$$

and the inequality (ii) holds due to the $\text{RE}(\kappa_0, s)$ condition of $\boldsymbol{\Sigma}$ and $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\zeta}\|_2 = 1$. Then using the maximal inequality and the assumption $\sup_j \boldsymbol{\Sigma}_{jj} \leq c_0$, we can bound the Gaussian width $\mathbb{W}(\mathcal{Y}_2)$ as follows

$$\mathbb{W}(\mathcal{Y}_2) \leq \sqrt{\frac{32c_0 s \log(2d)}{\kappa_0}}. \tag{B.17}$$

Combining the upper bound (B.17) of the Gaussian width $\mathbb{W}(\mathcal{Y}_2)$ with the bound (B.16), we have with probability at least $1 - e^{-\frac{c_2 n \delta^2}{\sigma^4}}$,

$$\sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A},3), |\mathcal{A}| \leq s} \frac{|\boldsymbol{\zeta}^\top \boldsymbol{A}\boldsymbol{\zeta}|}{\boldsymbol{\zeta}^\top \boldsymbol{\Sigma}\boldsymbol{\zeta}} \leq c_1 \sqrt{\frac{s \log(d)}{n}} + \delta,$$

for some constants c_1, c_2. Apply the same argument to $\boldsymbol{A}'$ and we get with probability at least $1 - e^{-\frac{c_2 K b \delta^2}{\sigma^4}}$,

$$\sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A},3), |\mathcal{A}| \leq s} \frac{|\boldsymbol{\zeta}^\top \boldsymbol{A}'\boldsymbol{\zeta}|}{\boldsymbol{\zeta}^\top \boldsymbol{\Sigma}\boldsymbol{\zeta}} \leq c_1 \sqrt{\frac{s \log(d)}{K}} + \sqrt{b}\delta,$$

for some contants c_1 and c_2. Combining these two upper bounds with the inequality (B.15), we have with probability at least $1 - 2e^{-\frac{c_2 n \delta^2}{\sigma^4}}$,

$$\sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A},3), |\mathcal{A}| \leq s} \frac{|\boldsymbol{\zeta}^\top (\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})\boldsymbol{\zeta}|}{\boldsymbol{\zeta}^\top \boldsymbol{\Sigma}\boldsymbol{\zeta}} \leq c_1 \sqrt{\frac{s \log(d)}{n}} + \delta,$$

for some contants c_1 and c_2. Note again that $\boldsymbol{A}'$-term is negligible relative to $\boldsymbol{A}$-term. By change of variable of δ, we have with probability at least $1 - \delta$,

$$\sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A},3), |\mathcal{A}| \leq s} \frac{|\boldsymbol{\zeta}^\top (\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})\boldsymbol{\zeta}|}{\boldsymbol{\zeta}^\top \boldsymbol{\Sigma}\boldsymbol{\zeta}} \leq c_1 \sqrt{\frac{s \log(d/\delta)}{n}}.$$

Then there exists some constant c_3, if $n \geq c_3 s \log(d/\delta)$,

$$\left| 1 + \frac{\boldsymbol{\Delta}_{\cdot j}^\top (\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})\boldsymbol{\Delta}_{\cdot j}}{\boldsymbol{\Delta}_{\cdot j}^\top \boldsymbol{\Sigma}\boldsymbol{\Delta}_{\cdot j}} \right|$$

$$\geq 1 - \sup_{\boldsymbol{\zeta} \in \mathcal{C}(\mathcal{A},3), |\mathcal{A}| \leq s} \frac{|\boldsymbol{\zeta}^\top (\widehat{\boldsymbol{\Sigma}}^1 - \boldsymbol{\Sigma})\boldsymbol{\zeta}|}{\boldsymbol{\zeta}^\top \boldsymbol{\Sigma}\boldsymbol{\zeta}}$$

$$\geq 1/2,$$

with probability at least $1 - \delta$.

By the RE(κ_0, s) condition of $\boldsymbol{\Sigma}$ and the property that $\boldsymbol{\Delta}_{\cdot j} \in \mathcal{C}(\mathcal{A}, 3)$ for some $|\mathcal{A}| \leq s$, we have

$$\boldsymbol{\Delta}_{\cdot j}^{\top} \boldsymbol{\Sigma} \boldsymbol{\Delta}_{\cdot j} \geq \kappa_0 \|\boldsymbol{\Delta}_{\cdot j}\|_2^2.$$

Thus we have with probability at least $1 - \delta$,

$$\mathcal{L}_1(\boldsymbol{W}^* + \boldsymbol{\Delta}) - \mathcal{L}_1(\boldsymbol{W}^*) - \langle \nabla \mathcal{L}_1(\boldsymbol{W}^*), \boldsymbol{\Delta} \rangle \geq \frac{1}{4} \sum_{j=1}^{K-1} \boldsymbol{\Delta}_{\cdot j}^{\top} \boldsymbol{\Sigma} \boldsymbol{\Delta}_{\cdot j}$$

$$\geq \frac{\kappa_0}{4} \sum_{j=1}^{K-1} \|\boldsymbol{\Delta}_{\cdot j}\|_2^2 = \frac{\kappa_0}{4} \|\boldsymbol{\Delta}\|_{2,2}^2,$$

if $n \geq c_3 s \log(d/\delta)$ for some constant c_3. $\qquad \square$