Supplement to "Optimality in High-Dimensional Tensor Discriminant Analysis"

Keqian Min and Qing Mai

This supplementary material contains proofs of the technical results in the paper "Optimality in High-Dimensional Tensor Discriminant Analysis" and a discussion about the sparsity assumption.

A Proof of Theorems

We first introduce some notations that are used in the proofs. For a vector $\mathbf{u} \in \mathbb{R}^p$, the l_0, l_1, l_2, l_∞ norm are defined as $\|\mathbf{u}\|_0$, $\|\mathbf{u}\|_1$, $\|\mathbf{u}\|_2$ and $\|\mathbf{u}\|_\infty$ respectively. Denote the support of a vector as $\text{supp}(\mathbf{u}) = \{j : u_j \neq 0\}$. For a matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$, the l_1 norm is $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$, and the l_∞ norm is $\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|$. The l_2 norm is $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$. The Frobenius norm is $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}$. The max norm is $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$. Define the support of a matrix as $\text{supp}(\mathbf{A}) = \{j : a_{ij} \neq 0 \text{ for some } i\}$. Denote $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_q)$ where $\mathbf{a}_h \in \mathbb{R}^p$ is the hth column of $\mathbf{A}$. Denote $\mathbf{A}_{j\cdot}$ as the jth row of $\mathbf{A}$. Denote $[p] = \{1, \ldots, p\}$. For a subset $T \subset [p]$, we define $\mathbf{A}_{T\cdot} = \left(a_{ij}\mathbf{1}_{\{i\in T, j\in[q]\}}\right)$, which is a sub-matrix of $\mathbf{A}$ restricted on the index set T. Similarly, define $\mathbf{A}_{\cdot T} = \left(a_{ij}\mathbf{1}_{\{i\in[p], j\in T\}}\right)$ and $\mathbf{A}_{T,T} = \left(a_{ij}\mathbf{1}_{\{i\in T, j\in T\}}\right)$. For a positive integer $s < p$, let $\Gamma(s) = \{\mathbf{u} \in \mathbb{R}^p : \|\mathbf{u}_{S^c}\|_1 \leq 2\|\mathbf{u}_S\|_1 \text{ for some } S \subset [p] \text{ with } |S| = s\}$, where $\mathbf{u}_S$ denotes the subvector of $\mathbf{u}$ restricted on the index set S. Define the restricted norm $\|\mathbf{u}\|_{2,s} = \sup_{\|\mathbf{v}\|_2=1, \mathbf{v}\in\Gamma(s)} \left|\mathbf{u}^T\mathbf{v}\right|$. For two numbers a and b, denote $a \vee b = \max\{a, b\}$.

Next, we introduce some tensor notation and operations. Let $\mathbf{A} \in \mathbb{R}^{p_1\times\cdots\times p_M}$ denote a tensor of order M. The vectorization of a tensor $\mathbf{A}$ is denoted by $\text{vec}(\mathbf{A}) \in \mathbb{R}^{p\times 1}$, $p = \prod_{m=1}^M p_m$, with $A_{i_1,\ldots,i_M}$ being its jth element, where $j = 1 + \sum_{m=1}^M (i_m - 1)\prod_{m'=1}^{m-1} p_{m'}$. Denote $p_{-m} = \prod_{m'=1, m'\neq m}^M p_{m'}$. The mode-$m$ matricization of $\mathbf{A}$ is denoted by $\mathbf{A}_{(m)} \in \mathbb{R}^{p_m\times p_{-m}}$, with $A_{i_1,\ldots,i_M}$ being its (i_m, j)-th element, where $j = 1 + \sum_{m'\neq m}(i_{m'} - 1)\prod_{l=1, l\neq m}^{m'-1} p_l$. Recall that the Tucker decomposition of $\mathbf{A}$ is denoted as $\mathbf{A} = [\![\mathbf{C}; \mathbf{G}_1, \ldots, \mathbf{G}_M]\!]$. We also have the fact that $\text{vec}([\![\mathbf{C}; \mathbf{G}_1, \ldots, \mathbf{G}_M]\!]) = (\mathbf{G}_M \otimes \cdots \otimes \mathbf{G}_1)\text{vec}(\mathbf{C}) = (\otimes_{m=M}^1 \mathbf{G}_m)\text{vec}(\mathbf{C})$ where $\otimes$ denotes the Kronecker product.

For ease of presentation, the proofs of the theoretical results in this paper are mainly given in the vector's form. Define the vectorized parameters as $\boldsymbol{\beta}_k = \text{vec}(\mathbf{B}_k)$, $\widehat{\boldsymbol{\beta}}_k = \text{vec}(\widehat{\mathbf{B}}_k)$, $\mathbf{v}_k = \text{vec}(\boldsymbol{\mu}_k)$ and $\widehat{\mathbf{v}}_k = \text{vec}(\widehat{\boldsymbol{\mu}}_k)$. We first present the objective function in the vector's form.

The vectorized HD-TDA estimator $\left(\widehat{\boldsymbol{\beta}}_2, \ldots, \widehat{\boldsymbol{\beta}}_K\right)$ corresponding to (2.5) can be rewritten as

$$\left(\widehat{\boldsymbol{\beta}}_2, \ldots, \widehat{\boldsymbol{\beta}}_K\right) = \arg\min_{\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_K} \left\{\sum_{k=2}^K \left(\boldsymbol{\beta}_k^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k - 2\boldsymbol{\beta}_k\widehat{\boldsymbol{\delta}}_k^T\right) + \lambda\sum_{j=1}^p \|\boldsymbol{\beta}_{\cdot j}\|_2\right\}, \tag{A.1}$$

where $\widehat{\boldsymbol{\delta}}_k = \widehat{\mathbf{v}}_k - \widehat{\mathbf{v}}_1$, $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_M \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_1$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)^T \in \mathbb{R}^{(K-1)\times p}$, $\boldsymbol{\beta}_{\cdot j} \in \mathbb{R}^{K-1}$ denotes the jth column vector of $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}_{\cdot j}\|_2 = \left(\sum_{k=2}^K \boldsymbol{\beta}_{k,j}^2\right)^{\frac{1}{2}}$ and $\boldsymbol{\beta}_{k,j}$ is the jth element of $\boldsymbol{\beta}_k$. After obtaining $\widehat{\boldsymbol{\beta}}_k$, we map $\widehat{\boldsymbol{\beta}}_k$ back to the original tensor to get $\widehat{\mathbf{B}}_k$. Since $\|\widehat{\mathbf{B}} - \mathbf{B}\|_F = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F$, it is equivalent to show that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F \lesssim \sqrt{s \sum_{m=1}^M \log p_m / n}$ with probability at least $1 - O(p^{-1})$.

The signal-to-noise ratio can be written as $\Delta_k = \sqrt{\boldsymbol{\beta}_k^T \boldsymbol{\Sigma} \boldsymbol{\beta}_k}$. Define the support of $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)^T$ as

$$\operatorname{supp}(\boldsymbol{\beta}) = \{j : \boldsymbol{\beta}_{k,j} \neq 0 \text{ for some } k\}.$$

Therefore, $\operatorname{supp}(\boldsymbol{\beta})$ and $\mathcal{D}$ refer to the same subgroup of important predictors and we have $|\mathcal{D}| = |\operatorname{supp}(\boldsymbol{\beta})|$.

For ease of presentation, throughout the rest of the proof we further assume that the diagonal elements of $\boldsymbol{\Sigma}_m, m = 1, \ldots, M$ are all ones. Otherwise, our estimated discriminant direction will be proportional to the truth by some constant depending on the trace of the covariance matrices. Even so, we are still estimating the same projection direction.

A.1 Proof of Theorem 3.1

Upper bound for estimation error. We first prove the conclusion for estimation error in Theorem 3.1. In the following lemma, we state a generic result regarding the estimator to the optimization problem in (A.1).

Lemma A.1. *Assume that $\sqrt{s \log p / n} \leq C$ for some constant C, and choose $\lambda \asymp \sqrt{\frac{\log p}{n}}$. If $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\delta}}_k$ in* (A.1) *satisfy*

(i) $\|\widehat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k\|_{\max} \lesssim \sqrt{\frac{\log p}{n}}$;

(ii) $\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_k\|_\infty \lesssim \sqrt{\frac{\log p}{n}}$;

(iii) $\operatorname{tr}\left\{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\} \gtrsim \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2$;

with probability at least $1 - O(p^{-1})$, *then we have*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F \lesssim \sqrt{\frac{s \log p}{n}}.$$

with probability at least $1 - O(p^{-1})$.

Lemma A.1 indicates that for any plug-in estimators $\widehat{\boldsymbol{\delta}}_k$ and $\widehat{\boldsymbol{\Sigma}}$ that satisfy the three conditions, the solution to the penalized objective function (A.1) converges to its truth at the rate of $\sqrt{s \log p / n}$. With Lemma A.1, to bound the estimation error remains to show that the sample estimators $\widehat{\boldsymbol{\delta}}_k$ and $\widehat{\boldsymbol{\Sigma}}$ defined in Section 2.3 satisfy the three conditions. Specifically, the first two conditions require the convergence rate of $\widehat{\boldsymbol{\delta}}_k$ and $\widehat{\boldsymbol{\Sigma}}$ to be no slower than $\sqrt{\log p / n}$. In fact, the convergence rate is strictly required and any slower rate will lead to sub-optimal convergence rate in the final result. The last condition requires that the smallest restricted eigenvalue of $\widehat{\boldsymbol{\Sigma}}$ is lower bounded by some positive constant.

We start from checking the first condition. We use the following results from Pan et al. (2019).

Lemma A.2 (Pan et al. (2019) c.f Lemma G.2, G.3, G.4)**.** *With probability at least $1 - O(p^{-1})$, we have* *(i)* $\|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_{\max} \lesssim \sqrt{\frac{\log p}{n}}$;

(ii) $\|\widehat{\boldsymbol{\Sigma}}_j^{(k)} - \boldsymbol{\Sigma}_j\|_{\max} \lesssim \sqrt{\frac{\log p_j}{np_{-j}}}$, $\|\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}_j\|_{\max} \lesssim \sqrt{\frac{\log p_j}{np_{-j}}}$, $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \lesssim \sqrt{\frac{\log p}{n}}$.

We can easily verify the first condition using Lemma A.2. By triangle inequality, we have that

$$\|\widehat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k\|_{\max} = \|\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_k + \boldsymbol{\mu}_1\|_{\max} \le \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_{\max} + \|\widehat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\|_{\max} \lesssim \sqrt{\frac{\log p}{n}}.$$

Next, we go to verify the second condition. While the second conclusion in Lemma A.2 can be used to bound the term in the second condition since $\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_k\|_\infty \le \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}\|\boldsymbol{\beta}_k\|_2$, the bound is not tight enough to reach the desired rate. Instead, after more careful analysis, we have the following lemma which verifies the second condition.

Lemma A.3. *With probability at least $1 - O(p^{-1})$, we have*

$$\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_k\|_\infty \lesssim \|\boldsymbol{\beta}_k\|_2\sqrt{\frac{\log p}{n}}.$$

Before we proceed to checking the third one, we state another important intermediate conclusion regarding $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. Denote $\Upsilon = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. We have the following sparse property of Υ.

Lemma A.4. *If $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\delta}}_k$ in* (A.1) *satisfy that*

(i) $\|\widehat{\boldsymbol{\delta}}_k - \boldsymbol{\delta}_k\|_{\max} \lesssim \sqrt{\frac{\log p}{n}}$;

(ii) $\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}_k\|_\infty \lesssim \sqrt{\frac{\log p}{n}}$,

with properly chosen $\lambda \asymp \sqrt{K-1}\sqrt{\frac{\log p}{n}}$, we have that

$$\sum_{j \in S^c} \|\Upsilon_{\cdot j}\|_2 \le 2\sum_{j \in S} \|\Upsilon_{\cdot j}\|_2.$$

When $K = 2$, Υ is a vector, and Lemma A.4 implies that $\Upsilon \in \Gamma(s)$. Lemma A.4 shows that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ also enjoys some sparsity property. With this sparsity property, the term $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ in the third condition is linked to the restricted eigenvalue of $\widehat{\boldsymbol{\Sigma}}$ in the proof. For any positive semi-definite matrix $\mathbf{A}$, the restricted eigenvalue is defined as

$$\phi_{\min}^{\mathbf{A}}(s) = \min_{\|\mathbf{u}\|_0 \le s, \mathbf{u} \ne 0} \frac{\mathbf{u}^T\mathbf{A}\mathbf{u}}{\mathbf{u}^T\mathbf{u}}, \quad \phi_{\max}^{\mathbf{A}}(s) = \max_{\|\mathbf{u}\|_0 \le s, \mathbf{u} \ne 0} \frac{\mathbf{u}^T\mathbf{A}\mathbf{u}}{\mathbf{u}^T\mathbf{u}}.$$

We have the following result for the restricted eigenvalue of $\widehat{\boldsymbol{\Sigma}}$.

Lemma A.5. *Assume $s\log p/n \le C_1$ for some constant C_1. For the sample estimator $\widehat{\boldsymbol{\Sigma}}$, with probability at least $1 - O(p^{-1})$, we have*

$$C_\lambda^{-M} - C\epsilon_s \le \phi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s) \le \phi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(s) \le C_\lambda^M + C\epsilon_s,$$

where $\epsilon_s = \sqrt{\frac{s\log p}{n}}$, and C is some positive constant.

The following lemma, together with Lemma A.4 and Lemma A.5, help us to verify the last condition.

Lemma A.6. *If we have*

(i) $\sum_{j\in S^c} \|\Upsilon_{\cdot j}\|_2 \leq 2\sum_{j\in S} \|\Upsilon_{\cdot j}\|_2$ *where* $\Upsilon = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$*;*

(ii) $C - c\epsilon_s \leq \phi^{\widehat{\boldsymbol{\Sigma}}}_{\min}(s) \leq \phi^{\widehat{\boldsymbol{\Sigma}}}_{\max}(s) \leq C + c\epsilon_s$ *for some* $\epsilon_s = o(1)$*,*

then we have

$$\sum_{k=2}^{K} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^T \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) \gtrsim \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2.$$

Since the two conditions in Lemma A.4 have been verified by Lemma A.2 and A.3, we complete the verification of the third condition by Lemma A.4–A.6. Therefore, we have proved that the sample estimators $\widehat{\boldsymbol{\delta}}_k$ and $\widehat{\boldsymbol{\Sigma}}$ satisfy the three conditions in Lemma A.1. By Lemma A.1, we complete the proof of the first conclusion in Theorem 3.1. The proofs of the lemmas can be found in Section B.

Upper bound for excess misclassification risk. To prove the second conclusion in Theorem 3.1, we first state a proposition below, which is from Lemma 7 in Cai & Zhang (2019).

Proposition A.1. *For two vectors* $\boldsymbol{\gamma}$ *and* $\widehat{\boldsymbol{\gamma}}$*, if* $\|\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}\|_2 = o(1)$ *as* $n \to \infty$*, and* $\|\boldsymbol{\gamma}\|_2 \geq c$ *for some constant* $c > 0$*, then, when* $n \to \infty$*,*

$$\|\boldsymbol{\gamma}\|_2 \|\widehat{\boldsymbol{\gamma}}\|_2 - \boldsymbol{\gamma}^{\mathrm{T}} \widehat{\boldsymbol{\gamma}} \asymp \|\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}\|_2^2.$$

Lemma A.7. *With probability at least* $1 - O(p^{-1})$*, we have*

$$\|\mathbf{v}_k - \widehat{\mathbf{v}}_k\|_{2,s} \lesssim \sqrt{\frac{s \log p}{n}}.$$

Next, we start to bound the excess misclassification risk. In the following proof, we use $\widehat{\boldsymbol{\beta}}$, $\widehat{\Delta}$ to denote $\widehat{\boldsymbol{\beta}}_2$, $\widehat{\Delta}_2$ for simplicity. Recall that

$$\widehat{C}(\mathbf{Z}) = \begin{cases} 1, & \left(\mathbf{Z} - \frac{\widehat{\mathbf{v}}_1 + \widehat{\mathbf{v}}_2}{2}\right)^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \log \frac{\widehat{\pi}_2}{\widehat{\pi}_1} < 0, \\ 2, & \left(\mathbf{Z} - \frac{\widehat{\mathbf{v}}_1 + \widehat{\mathbf{v}}_2}{2}\right)^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \log \frac{\widehat{\pi}_2}{\widehat{\pi}_1} \geq 0. \end{cases}$$

The misclassification risk $R_{\boldsymbol{\theta}}(\widehat{C})$ is calculated by

$$R_{\boldsymbol{\theta}}(\widehat{C}) = \pi_1 \mathrm{P}_{\boldsymbol{\theta}}(\widehat{C}(\mathbf{Z}) = 2 \mid \mathrm{label}(\mathbf{Z}) = 1) + \pi_2 \mathrm{P}_{\boldsymbol{\theta}}(\widehat{C}(\mathbf{Z}) = 1 \mid \mathrm{label}(\mathbf{Z}) = 2).$$

Define $\widehat{\Delta} = \sqrt{\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{\Sigma} \widehat{\boldsymbol{\beta}}}$, then $\mathrm{var}\left\{\left(\mathbf{Z} - \frac{\widehat{\mathbf{v}}_1 + \widehat{\mathbf{v}}_2}{2}\right)^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \log \frac{\widehat{\pi}_2}{\widehat{\pi}_1}\right\} = \widehat{\Delta}^2$. Let $\widehat{\mathbf{v}} = \frac{\widehat{\mathbf{v}}_1 + \widehat{\mathbf{v}}_2}{2}$. We have

$$\begin{aligned} \mathrm{P}\left\{\widehat{C}(\mathbf{Z}) = 2 \mid \mathrm{label}(\mathbf{Z}) = 1\right\} &= \mathrm{P}\left\{(\mathbf{Z} - \widehat{\mathbf{v}})^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \log \frac{\widehat{\pi}_2}{\widehat{\pi}_1} \geq 0 \mid \mathbf{Z} \sim \mathrm{N}(\mathbf{v}_2, \boldsymbol{\Sigma})\right\} \\ &= 1 - \Phi\left\{-\frac{(\mathbf{v}_1 - \widehat{\mathbf{v}})^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \log \frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right\} \\ &= \Phi\left\{\frac{(\mathbf{v}_1 - \widehat{\mathbf{v}})^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \log \frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right\} \\ &= \Phi\left\{-\frac{(\widehat{\mathbf{v}} - \mathbf{v}_1)^{\mathrm{T}} \widehat{\boldsymbol{\beta}} - \log \frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right\}, \end{aligned}$$

and

$$\begin{aligned}\mathrm{P}\left\{\widehat{C}(\mathbf{Z})=1 \mid \operatorname{label}(\mathbf{Z})=2\right\} &= \mathrm{P}\left\{(\mathbf{Z}-\widehat{\mathbf{v}})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}<0 \mid \mathbf{Z}\sim N(\mathbf{v}_1,\boldsymbol{\Sigma})\right\}\\ &= \Phi\left\{-\frac{(\mathbf{v}_2-\widehat{\mathbf{v}})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right\}\\ &= \bar{\Phi}\left\{\frac{(\mathbf{v}_2-\widehat{\mathbf{v}})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right\}\\ &= \bar{\Phi}\left\{-\frac{(\widehat{\mathbf{v}}-\mathbf{v}_2)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}-\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right\},\end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution and $\bar{\Phi}(x)=1-\Phi(x)$. Thus,

$$R_{\boldsymbol{\theta}}(\widehat{C})=\pi_1\Phi\left\{-\frac{(\widehat{\mathbf{v}}-\mathbf{v}_1)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}-\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1})}{\widehat{\Delta}}\right\}+\pi_2\bar{\Phi}\left\{-\frac{(\widehat{\mathbf{v}}-\mathbf{v}_2)^{\mathrm{T}}\widehat{\boldsymbol{\beta}}-\log\frac{\widehat{\pi}_2}{\widehat{\pi}_1}}{\widehat{\Delta}}\right\}.$$

Similarly, we can calculate the misclassification risk of oracle Fisher's rule which is

$$R_{\mathrm{opt}}(\boldsymbol{\theta})=\pi_1\Phi\left(-\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}\right)+\pi_2\bar{\Phi}\left(\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}\right).$$

Define an intermediate quantity

$$R^*=\pi_1\Phi\left(-\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}-\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right)+\pi_2\bar{\Phi}\left(\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right).$$

Let $\delta_n=\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2\vee\|\widehat{\mathbf{v}}_1-\mathbf{v}_1\|_{2,s}\vee\|\widehat{\mathbf{v}}_2-\mathbf{v}_2\|_{2,s}$. We first show that

$$R^*-R_{\mathrm{opt}}(\boldsymbol{\theta})\lesssim\exp\left(-\Delta^2/8\right)\Delta\delta_n^2. \tag{A.2}$$

Applying Taylor's expansion to the two terms in R^* at $-\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}$ and $\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}$ respectively, we obtain

$$\begin{aligned}&\Phi\left(-\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}-\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right)\\ =&\Phi\left(-\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}\right)+\Phi'\left(-\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}\right)\left\{\frac{\Delta}{2}-\frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}}+\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right\}+\mathrm{Re}_1\end{aligned}$$

and

$$\begin{aligned}&\bar{\Phi}\left(\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}+\log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right)\\ =&\,\bar{\Phi}\left(\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}\right)+\bar{\Phi}'\left(\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}\right)\left\{-\frac{\Delta}{2}+\frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}}+\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right\}+\mathrm{Re}_2\\ =&\,\bar{\Phi}\left(\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}\right)+\Phi'\left(\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}\right)\left\{\frac{\Delta}{2}-\frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}}-\log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}}-\frac{1}{\Delta}\right)\right\}+\mathrm{Re}_2\end{aligned}$$

where $\mathrm{Re}_1, \mathrm{Re}_2$ are the the residuals to be specified later. Thus,

$$R^* - R_{\mathrm{opt}}(\boldsymbol{\theta}) = \pi_1 \Phi' \left(-\frac{\Delta}{2} + \frac{\log \frac{\pi_2}{\pi_1}}{\Delta} \right) \left\{ \frac{\Delta}{2} - \frac{\boldsymbol{\delta}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}} + \log \frac{\pi_2}{\pi_1} \left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta} \right) \right\}$$
$$+ \pi_2 \Phi' \left(\frac{\Delta}{2} + \frac{\log \frac{\pi_2}{\pi_1}}{\Delta} \right) \left\{ \frac{\Delta}{2} - \frac{\boldsymbol{\delta}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}} - \log \frac{\pi_2}{\pi_1} \left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta} \right) \right\} + \pi_1 \mathrm{Re}_1 + \pi_2 \mathrm{Re}_2. \tag{A.3}$$

Note that

$$\mathrm{Re}_1 = \frac{1}{2} \left\{ \frac{\Delta}{2} - \frac{\boldsymbol{\delta}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}} + \log \frac{\pi_2}{\pi_1} \left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta} \right) \right\}^2 \Phi''(t_{1,n})$$

where $t_{1,n}$ is some constant satisfying that $t_{1,n}$ is between $-\frac{\Delta}{2} + \frac{\log \frac{\pi_2}{\pi_1}}{\Delta}$ and $-\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}} \widehat{\boldsymbol{\beta}} - \log \frac{\pi_2}{\pi_1}}{\widehat{\Delta}}$. By Proposition A.1, letting $\boldsymbol{\gamma} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}$ and $\widehat{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}^{1/2} \widehat{\boldsymbol{\beta}}$, we can obtain

$$\left| \Delta - \frac{\boldsymbol{\delta}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{\widehat{\Delta}} \right| = \left| \|\boldsymbol{\gamma}\|_2 - \frac{\boldsymbol{\gamma}^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}}{\|\widehat{\boldsymbol{\gamma}}\|_2} \right| = \left| \frac{\|\boldsymbol{\gamma}\|_2 \|\widehat{\boldsymbol{\gamma}}\|_2 - \boldsymbol{\gamma}^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}}{\|\widehat{\boldsymbol{\gamma}}\|_2} \right| \lesssim \frac{1}{\Delta} \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2 \lesssim \frac{1}{\Delta} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \lesssim \frac{1}{\Delta} \delta_n^2. \tag{A.4}$$

As $\delta_n \to 0$, $-\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}} \widehat{\boldsymbol{\beta}} - \log \frac{\pi_2}{\pi_1}}{\widehat{\Delta}} \to -\frac{\Delta}{2} + \frac{\log \frac{\pi_2}{\pi_1}}{\Delta}$, we have

$$|\Phi''(t_{1,n})| \asymp \left| \left(-\frac{\Delta}{2} + \frac{\log \frac{\pi_2}{\pi_1}}{\Delta} \right) \exp \left\{ -\frac{1}{2} \left(-\frac{\Delta}{2} + \frac{\log \frac{\pi_2}{\pi_1}}{\Delta} \right)^2 \right\} \right| \asymp \left| -\frac{\Delta}{2} \exp \left\{ -\frac{1}{2} \left(-\frac{\Delta}{2} \right)^2 \right\} \right|, \tag{A.5}$$

because Δ is bounded below by some constant. Thus,

$$|\Phi''(t_{1,n})| = O \left\{ \Delta \exp \left(\frac{-\Delta^2}{8} \right) \right\}.$$

In fact, we can check that

$$|\widehat{\Delta} - \Delta| \leq \sqrt{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Sigma} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})} \lesssim \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \delta_n. \tag{A.6}$$

Combining the results in (A.4) and (A.6), if $\delta_n = o(1)$, we have

$$\begin{aligned} \frac{\Delta}{2} - \frac{\boldsymbol{\delta}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}} + \log \frac{\pi_2}{\pi_1} \left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta} \right) &\leq \left| \log \frac{\pi_2}{\pi_1} \left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta} \right) \right| + \left| \frac{\Delta}{2} - \frac{\boldsymbol{\delta}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}} \right| \\ &\lesssim |\widehat{\Delta} - \Delta| + \frac{1}{\Delta} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \\ &\lesssim \delta_n + \frac{1}{\Delta} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \\ &\lesssim \delta_n. \end{aligned}$$

Thus, the reminder term satisfies

$$\mathrm{Re}_1 \lesssim \Delta \exp\left(\frac{-\Delta^2}{8}\right) \delta_n^2.$$

The result is also true for Re_2. We further expand (A.3),

$$\begin{aligned}
\frac{R^* - R_{\mathrm{opt}}(\boldsymbol{\theta})}{\sqrt{\pi_1 \pi_2}} &= e^{-\frac{1}{2}\left(-\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta}\right)^2 + \log \pi_1 - \frac{1}{2}\log(\pi_1\pi_2)} \left\{\frac{\Delta}{2} - \frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}} + \log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta}\right)\right\} \\
&+ e^{-\frac{1}{2}(\frac{\Delta}{2}+\frac{\log\frac{\pi_2}{\pi_1}}{\Delta})^2 + \log \pi_2 - \frac{1}{2}\log(\pi_1\pi_2)} \left\{\frac{\Delta}{2} - \frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}} - \log\frac{\pi_2}{\pi_1}\left(\frac{1}{\widehat{\Delta}} - \frac{1}{\Delta}\right)\right\} \\
&+ O\left\{\Delta \exp\left(\frac{-\Delta^2}{8}\right)\delta_n^2\right\} \\
&= e^{-\frac{\Delta^2}{8} - \frac{(\log\frac{\pi_2}{\pi_1})^2}{2\Delta^2}} \left(\frac{\Delta}{2} - \frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}}\right) + O\left\{\Delta \exp\left(\frac{-\Delta^2}{8}\right)\delta_n^2\right\} \\
&\lesssim \exp\left(-\frac{\Delta^2}{8}\right)\left|\frac{\Delta}{2} - \frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{2\widehat{\Delta}}\right| + O\left\{\Delta \exp\left(\frac{-\Delta^2}{8}\right)\delta_n^2\right\} \\
&\lesssim \Delta \exp(\frac{-\Delta^2}{8})\delta_n^2.
\end{aligned} \tag{A.7}$$

Since π_1 and π_2 are bounded, we have proved (A.2).

Next, we aim to show that

$$R_{\boldsymbol{\theta}}(\widehat{C}) - R^* \lesssim \Delta \exp\left(\frac{-\Delta^2}{8}\right)\delta_n^2.$$

Similarly, applying Taylor's expansion to $R_{\boldsymbol{\theta}}(\widehat{C})$ at $-\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}$ and $\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}$, we have

$$\begin{aligned}
R_{\boldsymbol{\theta}}(\widehat{C}) - R^* &= \pi_1 \Phi'\left(-\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right) \frac{-\left(\widehat{\mathbf{v}} - \mathbf{v}_1\right)^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1} + \frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}} \\
&+ \pi_2 \Phi'\left(\frac{\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}}\right) \frac{\left(\widehat{\mathbf{v}} - \mathbf{v}_2\right)^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \log\frac{\widehat{\pi}_2}{\widehat{\pi}_1} + \frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \log\frac{\pi_2}{\pi_1}}{\widehat{\Delta}} + \pi_1 \mathrm{Re}_1 + \pi_2 \mathrm{Re}_2.
\end{aligned} \tag{A.8}$$

Since $|\log \frac{\widehat{\pi}_2}{\widehat{\pi}_1} - \log \frac{\pi_2}{\pi_1}| \to 0$ as $n \to \infty$, we have

$$\begin{aligned}
&\left| \frac{-\left(\widehat{\mathbf{v}} - \mathbf{v}_1\right)^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \log \frac{\widehat{\pi}_2}{\widehat{\pi}_1} + \frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \log \frac{\pi_2}{\pi_1}}{\widehat{\Delta}} \right| \\
\lesssim &\left| \frac{-\left(\widehat{\mathbf{v}} - \mathbf{v}_1\right)^{\mathrm{T}} \widehat{\boldsymbol{\beta}} + \frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{\widehat{\Delta}} \right| \\
\lesssim &\left| \frac{\left(\widehat{\mathbf{v}}_1 + \widehat{\mathbf{v}}_2 - \mathbf{v}_1 - \mathbf{v}_2\right)^{\mathrm{T}} \widehat{\boldsymbol{\beta}}}{2\Delta} \right| \\
\lesssim &\frac{1}{\Delta} \left\{ \|\widehat{\mathbf{v}}_1 + \widehat{\mathbf{v}}_2 - \mathbf{v}_1 - \mathbf{v}_2\|_{2,s} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + |\left(\widehat{\mathbf{v}}_1 + \widehat{\mathbf{v}}_2 - \mathbf{v}_1 - \mathbf{v}_2\right)^{\mathrm{T}} \boldsymbol{\beta}| \right\} \\
\lesssim &\delta_n,
\end{aligned} \tag{A.9}$$

where we use the fact that $\|\boldsymbol{\beta}\|_2 \lesssim \Delta$ in the last inequality. Following the same idea of (A.5), we also have

$$|\Phi''(\cdot)| = O\left\{\Delta \exp\left(\frac{-\Delta^2}{8}\right)\right\}. \tag{A.10}$$

Therefore, (A.9) and (A.10) imply

$$Re_1 \asymp Re_2 \lesssim \Delta \exp(\frac{-\Delta^2}{8})\delta_n^2.$$

We further expand (A.8)

$$\begin{aligned}
\frac{R_{\boldsymbol{\theta}}(\widehat{C}) - R^*}{\sqrt{\pi_1 \pi_2}} &\lesssim \delta_n \left| e^{-\frac{(\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} - \log \frac{\pi_2}{\pi_1})^2}{2\widehat{\Delta}^2} - \frac{1}{2}\log\frac{\pi_2}{\pi_1}} - e^{-\frac{(\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \log \frac{\pi_2}{\pi_1})^2}{2\widehat{\Delta}^2} + \frac{1}{2}\log\frac{\pi_2}{\pi_1}} \right| + O(\mathrm{Re}_1) \\
&= \delta_n e^{-\frac{(\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^2 + \log \frac{\pi_2}{\pi_1}^2}{2\widehat{\Delta}^2}} \left| e^{\frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \log \frac{\pi_2}{\pi_1}}{2\widehat{\Delta}^2} - \frac{1}{2}\log\frac{\pi_2}{\pi_1}} - e^{-\frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \log \frac{\pi_2}{\pi_1}}{2\widehat{\Delta}^2} + \frac{1}{2}\log\frac{\pi_2}{\pi_1}} \right| + O(\mathrm{Re}_1).
\end{aligned} \tag{A.11}$$

When $\delta_n \to 0$, $\frac{(\frac{1}{2}\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^2 + \log \frac{\pi_2}{\pi_1}^2}{2\widehat{\Delta}^2} \to \frac{\Delta^2}{8} + \frac{\log \frac{\pi_2}{\pi_1}^2}{2\Delta^2}$. We further have $e^{-\frac{\Delta^2}{8} - \frac{\log \frac{\pi_2}{\pi_1}^2}{2\Delta^2}} \lesssim e^{-\frac{\Delta^2}{8}}$. Let $x = \frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} \log \frac{\pi_2}{\pi_1}}{2\widehat{\Delta}^2} - \frac{1}{2}\log\frac{\pi_2}{\pi_1}$, the absolute difference can be written as $|e^x - e^{-x}|$. In fact, using the result in (A.4), we have

$$x = \frac{\log \frac{\pi_2}{\pi_1}}{2}\left(\frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{\widehat{\Delta}^2} - 1\right) \lesssim \frac{1}{\Delta}\left|\frac{\boldsymbol{\delta}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}}{\widehat{\Delta}} - \Delta\right| \lesssim \frac{1}{\Delta^2}\delta_n^2 = o(1).$$

When $x = o(1)$, we have $|e^x - e^{-x}| \lesssim x \lesssim \frac{1}{\Delta^2}\delta_n^2$. Thus,

$$\begin{aligned}
\frac{R_{\boldsymbol{\theta}}(\widehat{C}) - R^*}{\sqrt{\pi_1 \pi_2}} &\lesssim \delta_n e^{-\frac{\Delta^2}{8}} \frac{1}{\Delta^2}\delta_n^2 + O\left\{\Delta \exp\left(\frac{-\Delta^2}{8}\right)\delta_n^2\right\} \\
&\lesssim \Delta \exp\left(\frac{-\Delta^2}{8}\right)\delta_n^2.
\end{aligned} \tag{A.12}$$

Combining (A.7) and (A.12), we obtain

$$R_{\boldsymbol{\theta}}(\widehat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) = (R_{\boldsymbol{\theta}}(\widehat{C}) - R^*) + (R^* - R_{\mathrm{opt}}(\boldsymbol{\theta})) \lesssim \Delta \exp(\frac{-\Delta^2}{8})\delta_n^2.$$

We have proved that $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\| \lesssim \sqrt{\frac{s\log p}{n}}$. By Lemma A.7, we know that $\delta_n \lesssim \sqrt{\frac{s\log p}{n}}$ with probability at least $1 - O(p^{-1})$. Since $C \le \Delta \le 3C$, we have, with probability at least $1 - O(p^{-1})$

$$R_{\boldsymbol{\theta}}(\widehat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \lesssim \frac{s\log p}{n}.$$

A.2 Proof of Theorem 3.2

Lower bound for estimation error. We follow the proof in Cai & Zhang (2019) by first stating two lemmas below. For two probability distribution P_1 and P_2, denote the Kullback-Leibler divergence as $\mathrm{KL}(\mathrm{P}_1, \mathrm{P}_2)$.

Lemma A.8 (Fano's Lemma, Tsybakov (2009)). *Suppose that $\boldsymbol{\Theta}_p$ is a parameter space consisting of T parameters $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots \boldsymbol{\theta}_T \in \boldsymbol{\Theta}_p$ for some $T > 0$, and $d(\cdot,\cdot) : \boldsymbol{\Theta}_p \times \boldsymbol{\Theta}_p \to \mathbb{R}^+$ is some distance. Denote $\mathrm{P}_{\boldsymbol{\theta}}$ to be some probability measure parameterized by $\boldsymbol{\theta}$. If for some constants $\alpha \in \left(0, \frac{1}{8}\right), \gamma > 0, \mathrm{KL}\left(\mathrm{P}_{\boldsymbol{\theta}_i}, \mathrm{P}_{\boldsymbol{\theta}_0}\right) \le \alpha \log(T)/n$ for all $1 \le i \le T$, and $d\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\right) \ge \gamma$ for all $0 \le i \ne j \le T$, then*

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{i\in[T]} \mathrm{E}_{\boldsymbol{\theta}_i}\left\{d_{\boldsymbol{\theta}_i}\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_i\right)\right\} \gtrsim \gamma.$$

Lemma A.9 (Tsybakov (2009)). *Define $\mathcal{A}_{p,s} = \{\mathbf{u} : \mathbf{u} \in \{0,1\}^p, \|\mathbf{u}\|_0 \le s\}$. If $p \ge 4s$, then there is a subset $\{\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_T\} \subset \mathcal{A}_{p,s}$ such that $\mathbf{u}_0 = \{0, \dots, 0\}^{\mathrm{T}}, \rho_{\mathrm{H}}\left(\mathbf{u}_i, \mathbf{u}_j\right) \ge s/2$ and $\log(T+1) \ge (s/5)\log(p/s)$, where ρ_{H} denotes the Hamming distance.*

We first construct a subspace of the whole parameter space and derive the lower bound of the estimation error in the subspace. Since $\log p/\log(p/s) = O(1)$, we have $s < p/4$ for sufficiently large p. Let $s' = \lceil s/2 \rceil$. Define $\mathbf{f}_0 \in \mathbb{R}^p$ as a vector with the last s' entries being $1/\sqrt{s'}$ and the rest being 0. Thus we have $\|\mathbf{f}_0\|_2 = 1$. Let $r = \lceil p/2 \rceil$. By Lemma A.9, there exists a collection of vectors $\widetilde{\mathcal{A}}_{r,s'} = \{\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_T\} \subset \mathcal{A}_{r,s'}$ such that $\mathbf{u}_i \in \mathbb{R}^r$, $\rho_{\mathrm{H}}\left(\mathbf{u}_i, \mathbf{u}_j\right) > s'/2, \log(T+1) \ge (s'/5)\log(r/s')$ and we denote $\mathbf{u}_0 = \mathbf{0}_r$. For any vector $\mathbf{a} \in \mathbb{R}^r$, let $\mathbf{F}_{\mathbf{a}}$ be the $p \times p$ symmetric matrix whose ith row and column are both $\epsilon a_i \mathbf{f}_0$ for $i \in \{1, \dots, r\}$, where ϵ will be specified later. Therefore, a_i determines whether the ith column and ith row in $\mathbf{F}_{\mathbf{a}}$ are all zeros or equal to $\epsilon \mathbf{f}_0$. And we always have the bottom-right block matrix of $\mathbf{F}_{\mathbf{a}}$ filled with zeros. Then we consider the parameter set

$$\begin{aligned}\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_M) : \pi_1 = \pi_2 = 1/2, \mathrm{vec}(\boldsymbol{\mu}_1) = \left(\mathbf{I}_p + \mathbf{F}_{\mathbf{u}}\right)\mathbf{f}_0,\\ \mathrm{vec}(\boldsymbol{\mu}_2) = -\left(\mathbf{I}_p + \mathbf{F}_{\mathbf{u}}\right)\mathbf{f}_0, \boldsymbol{\Sigma}_m = \mathbf{I}_{p_m}, m = 1, \dots, M; \mathbf{u} \in \widetilde{\mathcal{A}}_{r,s'} \cup \{\mathbf{0}_r\}\Big\}.\end{aligned}$$

As we consider the binary case, we denote $\boldsymbol{\beta}_2$ as $\boldsymbol{\beta}$ for simplicity. Then for fixed $\mathbf{u}$, we have $\boldsymbol{\beta}_{\mathbf{u}} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = -2(\mathbf{I}_p + \mathbf{F}_{\mathbf{u}})\mathbf{f}_0$. Then we have $\|\boldsymbol{\beta}_{\mathbf{u}}\|_0 \le \|\mathbf{f}_0\|_0 + \|\mathbf{F}_{\mathbf{u}}\mathbf{f}_0\|_0$. By the construction of $\mathbf{f}_0$, we have $\|\mathbf{f}_0\|_0 = s'$. Since the first $p - s'$ elements in $\mathbf{f}_0$ are zeros and the bottom-right block matrix in $\mathbf{F}_{\mathbf{u}}$ are zeros, $\|\mathbf{F}_{\mathbf{u}}\mathbf{f}_0\|_0$ only depends on the number of rows in $\mathbf{F}_{\mathbf{u}}$ that has non-zero inner product with $\mathbf{f}_0$ and is in the top half of $\mathbf{F}_{\mathbf{u}}$. By the construction of $\mathbf{F}_{\mathbf{u}}$, u_i determines whether the ith row is $\epsilon\mathbf{f}_0$ or $\mathbf{0}_p$. Thus we have $\|\mathbf{F}_{\mathbf{u}}\mathbf{f}_0\|_0 = \|\mathbf{u}\|_0 \le s'$. Therefore, $\|\boldsymbol{\beta}_{\mathbf{u}}\|_0 \le 2s' = s$. It also implies that

$$\|\boldsymbol{\beta}_{\mathbf{u}} - \boldsymbol{\beta}_{\widetilde{\mathbf{u}}}\|_2^2 = 4\left\|\left(\mathbf{F}_{\mathbf{u}} - \mathbf{F}_{\widetilde{\mathbf{u}}}\right)\mathbf{f}_0\right\|_2^2 = 4\rho_{\mathrm{H}}(\mathbf{u}, \widetilde{\mathbf{u}})\epsilon^2\left\|\mathbf{f}_0\right\|_2^2 \ge 4(s'/2)\epsilon^2 = s\epsilon^2. \tag{A.13}$$

In addition, if we take ϵ such that $\|\mathbf{F_u}\|_2 \leq \|\mathbf{F_u}\|_F \leq \sqrt{2s'\epsilon^2} = o(1)$, then for sufficiently large n, we have $\Delta = \sqrt{\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta}} = \sqrt{4\mathbf{f}_0^T(\mathbf{I}_p + \mathbf{F_u})^T(\mathbf{I}_p + \mathbf{F_u})\mathbf{f}_0} = O(1)$ since $\|\mathbf{f}_0\|_2 = 1$. Then, it implies that $\boldsymbol{\Theta}_0 \subset \mathcal{G}\left(s, p; c, C_\lambda, C\right)$.

Next, we aim to bound $\mathrm{KL}\left(\mathrm{P}_{\boldsymbol{\theta}_{\mathbf{u}_i}}, \mathrm{P}_{\boldsymbol{\theta}_{\mathbf{u}_0}}\right)$ for $i \in [T]$, where $\mathrm{P}_{\boldsymbol{\theta}_{\mathbf{u}_i}}$ and $\mathrm{P}_{\boldsymbol{\theta}_0}$ denote the distributions $\mathrm{N}\left((\mathbf{I}_p + \mathbf{F}_{\mathbf{u}_i})\,\mathbf{f}_0, \mathbf{I}_p\right)$ and $\mathrm{N}\left(\mathbf{f}_0, \mathbf{I}_p\right)$ respectively. We then have

$$\begin{aligned}
\mathrm{KL}\left(\mathrm{P}_{\boldsymbol{\theta}_{\mathbf{u}_i}}, \mathrm{P}_{\boldsymbol{\theta}_{\mathbf{u}_0}}\right) &= \frac{1}{2}\left\{\mathrm{tr}(\mathbf{I}_p) + (\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{u}_i}} - \boldsymbol{\mu}_{\boldsymbol{\theta}_0})^T\mathbf{I}_p^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{u}_i}} - \boldsymbol{\mu}_{\boldsymbol{\theta}_0}) - p + \log\frac{|\mathbf{I}_p|}{|\mathbf{I}_p|}\right\} \\
&= \frac{1}{2}(\mathbf{F}_{\mathbf{u}_i}\mathbf{f}_0)^T\mathbf{F}_{\mathbf{u}_i}\mathbf{f}_0 \\
&\leq \frac{1}{2}s'\epsilon^2 \\
&\leq \frac{1}{4}s\epsilon^2.
\end{aligned}$$

Note that $\log(T+1) \geq (s'/5)\log(r/s') = (s/10)\log(p/s)$. Furthermore, for sufficiently large p, we have $\log(T) \geq \frac{1}{2}\log(T{+}1) \geq (s/20)\log(p/s)$. Letting $\epsilon = \sqrt{\log(p/s)/(80n)}$, then $\mathrm{KL}\left(\mathrm{P}_{\boldsymbol{\theta}_{\mathbf{u}_i}}, \mathrm{P}_{\boldsymbol{\theta}_{\mathbf{u}_0}}\right) \leq (s\log(p/s))/(320n) \leq \alpha\log(T)/n$ for $\alpha = 1/16$.

In addition, let $\gamma = \sqrt{s\log(p/s)/(80n)}$. Then (A.13) implies that for any $0 \leq i \neq j \leq T$, we have

$$d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \|\boldsymbol{\beta}_{\mathbf{u}} - \boldsymbol{\beta}_{\widetilde{\mathbf{u}}}\|_2 \geq \sqrt{s}\epsilon = \sqrt{s\log(p/s)/(80n)} = \gamma.$$

Then by Lemma A.8, we have $\inf_{\widehat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\theta}\in\mathcal{G}}\mathrm{E}\left(\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_F\right) \gtrsim \sqrt{\frac{s\log(p/s)}{n}}$. Since $\log p/\log(p/s) = O(1)$, we have $\log(p/s) \asymp \log p$. Because $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_F = \|\widehat{\mathbf{B}} - \mathbf{B}\|_F$, we finally have

$$\inf_{\widehat{\mathbf{B}}}\sup_{\boldsymbol{\theta}\in\mathcal{G}}\mathrm{E}\left(\|\widehat{\mathbf{B}} - \mathbf{B}\|_F\right) \gtrsim \sqrt{\frac{s\log p}{n}}.$$

Lower bound for excess misclassification risk. Next, we prove the second conclusion for excess misclassification risk. For any classifier $\widehat{C}$, define

$$\mathrm{L}_{\boldsymbol{\theta}}(\widehat{C}) = \mathrm{P}_{\boldsymbol{\theta}}\left(\widehat{C}(\mathbf{Z}) \neq C_{\mathrm{opt}}(\mathbf{Z})\right).$$

where $C_{\mathrm{opt}}(\mathbf{Z})$ is the optimal Bayes' classifier. We have the following two lemmas which are from Lemma 3 and Lemma 4 in Cai & Zhang (2019).

Lemma A.10 (Cai & Zhang (2019), c.f. Lemma 3)**.** *Let* $\mathbf{Z} \sim \frac{1}{2}\mathrm{N}_p\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}\right) + \frac{1}{2}\mathrm{N}_p\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}\right)$ *with parameter* $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$*, where* $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p$ *and* $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$*. If a classifier* $\widehat{C}$ *satisfying* $\mathrm{L}_{\boldsymbol{\theta}}(\widehat{C}) = o(1)$ *as* $n \to \infty$*, then, for sufficiently large* n*,*

$$R_{\boldsymbol{\theta}}(\widehat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \geq \frac{\sqrt{2\pi}\Delta}{8}\exp\left(\frac{\Delta^2}{8}\right)\mathrm{L}^2_{\boldsymbol{\theta}}(\widehat{C}).$$

Lemma A.11 (Cai & Zhang (2019), c.f. Lemma 4)**.** *Let* $\boldsymbol{\theta} = (\boldsymbol{\mu}, -\boldsymbol{\mu}, \mathbf{I}_p)$ *and* $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\mu}}, -\widetilde{\boldsymbol{\mu}}, \mathbf{I}_p)$ *with* $\|\boldsymbol{\mu}\|_2 = \|\widetilde{\boldsymbol{\mu}}\|_2 = \Delta/2$*. For any classifier* C*, if* $\|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|_2 = o(1)$ *as* $n \to \infty$*, then, for sufficiently large* n*,*

$$\mathrm{L}_{\boldsymbol{\theta}}(C) + \mathrm{L}_{\widetilde{\boldsymbol{\theta}}}(C) \geq \frac{1}{\Delta}\exp\left(-\frac{\Delta^2}{8}\right)\|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|_2.$$

With Lemma A.10, the problem of bounding $R_{\boldsymbol{\theta}}(\widehat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta})$ is now reduced to bound $\mathrm{L}_{\boldsymbol{\theta}}(\widehat{C})$. The following lemma can be viewed as a variant of Fano's lemma that can be used to bound $\mathrm{L}_{\boldsymbol{\theta}}(\widehat{C})$.

Lemma A.12 (Tsybakov (2009))**.** *Let $T \geq 0$ and $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T \in \boldsymbol{\Theta}_p$ for $\boldsymbol{\Theta}_p$ defined in Lemma A.8. For some constants $\alpha_0 \in \left(0, \frac{1}{8}\right], \gamma > 0$, and any classifier $\widehat{C}$, if $\mathrm{KL}\left(\mathrm{P}_{\boldsymbol{\theta}_i}, \mathrm{P}_{\boldsymbol{\theta}_0}\right) \leq \alpha_0 \log(T)/n$ for all $1 \leq i \leq T$, and $\mathrm{L}_{\boldsymbol{\theta}_i}(\widehat{C}) < \gamma$ implies that $\mathrm{L}_{\boldsymbol{\theta}_j}(\widehat{C}) \geq \gamma$ for all $0 \leq i \neq j \leq T$, then*

$$\inf_{\widehat{C}} \sup_{i \in [T]} \mathrm{P}_{\boldsymbol{\theta}_i}\left(\mathrm{L}_{\boldsymbol{\theta}_i}(\widehat{C}) \geq \gamma\right) \geq \frac{\sqrt{T}}{\sqrt{T}+1}\left(1 - 2\alpha_0 - \sqrt{\frac{2\alpha_0}{\log T}}\right).$$

We aim to use Lemma A.12 to bound $\mathrm{L}_{\boldsymbol{\theta}_i}(\widehat{C})$ and then use Lemma A.10 to bound the excess misclassification risk. The first step is to construct a subspace of the whole parameter space and then derive the lower bound in the subspace. Let $\mathbf{e}_1$ be the basis vector in the standard Euclidean space whose first entry is 1 and 0 elsewhere. Let $s' = s - 1$. Denote $\check{\mathcal{A}}_{p,s'} = \left\{\mathbf{u} \in \{0,1\}^p : \mathbf{u}^T\mathbf{e}_1 = 0, \|\mathbf{u}\|_0 \leq s'\right\}$. By Lemma A.9, there exists a collection of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_T\} \subset \check{\mathcal{A}}_{p,s'}$ such that $\mathbf{u}_i \in \mathbb{R}^p$, $\rho_{\mathrm{H}}\left(\mathbf{u}_i, \mathbf{u}_j\right) > s'/2, \log(T+1) \geq (s'/5)\log((p-1)/s')$. We consider the subspace

$$\begin{aligned}\boldsymbol{\Theta}_1 = \{&\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M) : \pi_1 = \pi_2 = 1/2, \mathrm{vec}(\boldsymbol{\mu}_1) = \epsilon\mathbf{u} + \lambda\mathbf{e}_1,\\ &\mathrm{vec}(\boldsymbol{\mu}_2) = -\mathrm{vec}(\boldsymbol{\mu}_1), \boldsymbol{\Sigma}_m = \mathbf{I}_{p_m}, m = 1, \ldots, M-1, \boldsymbol{\Sigma}_M = \sigma^2\mathbf{I}_{p_M}; \mathbf{u} \in \check{\mathcal{A}}_{p,s}\Big\}.\end{aligned}$$

where $\epsilon = \sigma\sqrt{\log p/n}, \sigma^2 = O(1)$. We have $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_p$. We choose λ such that

$$C \leq \Delta^2 = \mathrm{vec}^T(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}\mathrm{vec}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = \frac{4\|\epsilon\mathbf{u} + \lambda\mathbf{e}_1\|_2^2}{\sigma^2} = \frac{4(\epsilon^2\|\mathbf{u}\|_2^2 + \lambda^2)}{\sigma^2} \leq 3C,$$

for some $C > 0$. Moreover, since $\mathbf{u}^T\mathbf{e}_1 = 0$, $\|\boldsymbol{\beta}\|_0 = \|2\sigma^2(\epsilon\mathbf{u} + \lambda\mathbf{e}_1)\|_0 \leq s' + 1 = s$. It ensures that $\boldsymbol{\Theta}_1 \in \mathcal{G}\left(s, p; c, C_\lambda, C\right)$. For $\mathbf{u}_i \in \check{\mathcal{A}}_{p,s}$, let $\mathrm{vec}(\boldsymbol{\mu}_{\boldsymbol{\theta}_i}) = \epsilon\mathbf{u}_i + \lambda\mathbf{e}_1$ and we consider the corresponding distribution $\mathrm{N}(\epsilon\mathbf{u}_i + \lambda\mathbf{e}_1, \sigma^2\mathbf{I}_p)$. For any $1 \leq i \neq j \leq T$, we have

$$\begin{aligned}\mathrm{KL}\left(\mathrm{P}_{\boldsymbol{\theta}_i}, \mathrm{P}_{\boldsymbol{\theta}_j}\right) &= \frac{1}{2}\left\{\mathrm{tr}(\mathbf{I}_p) + \mathrm{vec}^T(\boldsymbol{\mu}_{\boldsymbol{\theta}_i} - \boldsymbol{\mu}_{\boldsymbol{\theta}_j})(\sigma^2\mathbf{I}_p)^{-1}\mathrm{vec}(\boldsymbol{\mu}_{\boldsymbol{\theta}_i} - \boldsymbol{\mu}_{\boldsymbol{\theta}_j}) - p\right\}\\ &= \frac{1}{2\sigma^2}\|\mathrm{vec}(\boldsymbol{\mu}_{\boldsymbol{\theta}_i}) - \mathrm{vec}(\boldsymbol{\mu}_{\boldsymbol{\theta}_j})\|_2^2\\ &= \frac{1}{2\sigma^2}\epsilon^2\|\mathbf{u}_i - \mathbf{u}_j\|_2^2\\ &\leq \frac{s\log p}{n},\end{aligned}$$

since $\epsilon = \sigma\sqrt{\log p/n}$ and $\|\mathbf{u}_i - \mathbf{u}_j\|_2^2 \leq 2\|\mathbf{u}_i\|_2^2 \leq 2s$. On the other hand, by Lemma A.11, we have

$$\mathrm{L}_{\boldsymbol{\theta}_i}(C) + \mathrm{L}_{\boldsymbol{\theta}_j}(C) \gtrsim \|\mathrm{vec}(\boldsymbol{\mu}_{\boldsymbol{\theta}_i}) - \mathrm{vec}(\boldsymbol{\mu}_{\boldsymbol{\theta}_j})\|_2 = \epsilon\|\mathbf{u}_i - \mathbf{u}_j\|_2 \gtrsim \sqrt{\frac{s\log p}{n}},$$

since $\|\mathbf{u}_i - \mathbf{u}_j\|_2 \geq \sqrt{s'/2} \asymp \sqrt{s}$. Therefore, by taking $\gamma \asymp \sqrt{\frac{s\log p}{n}}$, Lemma A.12 implies that for any $0 < \alpha_1 < 1$, there exists some $C_{\alpha_1} > 0$ such that

$$\inf_{\widehat{C}} \sup_{i \in [T]} \mathrm{P}_{\boldsymbol{\theta}_i}\left(\mathrm{L}_{\boldsymbol{\theta}_i}(\widehat{C}) \geq C_{\alpha_1}\sqrt{\frac{s\log p}{n}}\right) \geq 1 - \alpha_1.$$

Combing this result with Lemma A.10, we complete the proof of Theorem 3.2.

A.3 Proof of Theorem 3.3

Proof of Theorem 3.3. The first conclusion has been proved in the proof of Theorem 3.1. For the second one, by definition

$$\begin{aligned}\overline{R}_{\boldsymbol{\theta}}(\widehat{C}) - \overline{R}_{\text{opt}}(\boldsymbol{\theta}) &= \sum_{k=1}^{K}\sum_{l\neq k}\pi_k\left\{\mathrm{P}_{\boldsymbol{\theta}}(\widehat{D}_k < \widehat{D}_l \mid \text{label}(\mathbf{Z}) = k) - \mathrm{P}_{\boldsymbol{\theta}}(D_k < D_l \mid \text{label}(\mathbf{Z}) = k)\right\}\\ &= \sum_{1\leq k<l\leq K}\Big\{\pi_k\mathrm{P}_{\boldsymbol{\theta}}(\widehat{D}_k < \widehat{D}_l \mid \text{label}(\mathbf{Z}) = k) - \pi_k\mathrm{P}_{\boldsymbol{\theta}}(D_k < D_l \mid \text{label}(\mathbf{Z}) = k)\\ &\qquad + \pi_l\mathrm{P}_{\boldsymbol{\theta}}(\widehat{D}_l < \widehat{D}_k \mid \text{label}(\mathbf{Z}) = l) - \pi_l\mathrm{P}_{\boldsymbol{\theta}}(D_l < D_k \mid \text{label}(\mathbf{Z}) = l)\Big\}.\end{aligned}$$

By arranging the probabilities in pair, we can apply the second part of the proof of Theorem 3.1 here. Then we have, for each pair of k, l,

$$\begin{aligned}&\pi_k\mathrm{P}_{\boldsymbol{\theta}}(\widehat{D}_k < \widehat{D}_l \mid \text{label}(\mathbf{Z}) = k) - \pi_k\mathrm{P}_{\boldsymbol{\theta}}(D_k < D_l \mid \text{label}(\mathbf{Z}) = k)\\ +&\pi_l\mathrm{P}_{\boldsymbol{\theta}}(\widehat{D}_l < \widehat{D}_k \mid \text{label}(\mathbf{Z}) = l) - \pi_l\mathrm{P}_{\boldsymbol{\theta}}(D_l < D_k \mid \text{label}(\mathbf{Z}) = l) \lesssim \frac{s\log p}{n}\end{aligned}$$

with probability at least $1 - O(p^{-1})$. Since K is fixed, the number of pairs is also fixed. Therefore, we obtain the convergence rate for $\overline{R}_{\boldsymbol{\theta}}(\widehat{C}) - \overline{R}_{\text{opt}}(\boldsymbol{\theta})$. □

B Proof of Lemmas

B.1 Proof of Lemma A.1

Proof of Lemma A.1. From the proof of Lemma A.4, we have that

$$\lambda\sum_{j=1}^{p}(\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2 - \|\boldsymbol{\beta}_{\cdot j}\|_2) \leq \sum_{k=2}^{K}\left\{-(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) - 2\langle\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k - \widehat{\boldsymbol{\delta}}_k, \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\rangle\right\}, \tag{B.1}$$

and

$$\sum_{j\in S^c}\|\Upsilon_{\cdot j}\|_2 \leq 2\sum_{j\in S}\|\Upsilon_{\cdot j}\|_2, \tag{B.2}$$

where $\Upsilon = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$.

Let $\lambda \asymp \sqrt{K-1}\sqrt{\frac{\log p}{n}} \asymp \sqrt{\frac{\log p}{n}}$. By reorganizing (B.1), we have

$$\begin{aligned}
\sum_{k=2}^{K}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k) &\leq 2\sum_{k=2}^{K}|\langle\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k-\widehat{\boldsymbol{\delta}}_k,\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\rangle| + 2\lambda\sum_{j=1}^{p}\left|\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2-\|\boldsymbol{\beta}_{\cdot j}\|_2\right| \\
&\leq 2\sum_{k=2}^{K}\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k-\widehat{\boldsymbol{\delta}}_k\|_\infty\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\|_1 + 2\lambda\sum_{j=1}^{p}\left|\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2-\|\boldsymbol{\beta}_{\cdot j}\|_2\right| \\
&\lesssim \sqrt{\frac{\log p}{n}}\sum_{k=2}^{K}\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\|_1 + \sqrt{K-1}\sqrt{\frac{\log p}{n}}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_2 \\
&\lesssim \sqrt{\frac{\log p}{n}}\left(\sum_{k=2}^{K}\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\|_1 + \sqrt{K-1}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_2\right). \qquad \text{(B.3)}
\end{aligned}$$

For the two terms in (B.3), using the result in (B.2), we can compute that

$$\begin{aligned}
\sum_{j=1}^{p}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_2 &= \sum_{j\in S}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_2 + \sum_{j\in S^c}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_2 \leq 3\sum_{j\in S}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_2 \\
&\leq 3\sqrt{s}\sqrt{\sum_{j\in S}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_2^2} \leq 3\sqrt{s}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F, \qquad \text{(B.4)}
\end{aligned}$$

and

$$\sum_{k=2}^{K}\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\|_1 = \sum_{j=1}^{p}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_1 \leq \sqrt{K-1}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_2 \leq 3\sqrt{s}\sqrt{K-1}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F. \qquad \text{(B.5)}$$

Plugging (B.4) and (B.5) into (B.3), we have

$$\sum_{k=2}^{K}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k) \lesssim \sqrt{K-1}\sqrt{\frac{s\log p}{n}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F \asymp \sqrt{\frac{s\log p}{n}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F. \qquad \text{(B.6)}$$

By Condition (iii), we have

$$\sum_{k=2}^{K}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k) \gtrsim \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F^2. \qquad \text{(B.7)}$$

Combining the results in (B.7) and (B.6), we have

$$\sum_{k=2}^{K}\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\|_2^2 \lesssim \sqrt{\frac{s\log p}{n}}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F,$$

which gives us

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_F \lesssim \sqrt{\frac{s\log p}{n}}.$$

□

B.2 Proof of Lemma A.3

Since $\widehat{\mathbf{\Sigma}} = \otimes_{m=M}^{1} \widehat{\mathbf{\Sigma}}_m$, we start from bounding $\|(\widehat{\mathbf{\Sigma}}_j^{(k)} - \mathbf{\Sigma}_j)\boldsymbol{\beta}_k\|_\infty$ and $\|(\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j)\boldsymbol{\beta}_k\|_\infty$. We show it in the later proof that $\widehat{\mathbf{\Sigma}}_j^{(k)}$ can be written as the summation of $n_k - 1$ independent matrix normal variables such that

$$\widehat{\mathbf{\Sigma}}_j^{(k)} = \frac{1}{(n_k - 1)p_{-j}} \sum_{i=1}^{n_k - 1} \mathbf{Z}^i \left(\mathbf{Z}^i\right)^T,$$

where $\mathbf{Z}^i$ is a linear combination of $\mathbf{X}_{(j)}^t$ for all t such that $Y^t = k$ and $\mathbf{Z}^i \sim \mathrm{MN}(\mathbf{0}; \mathbf{\Sigma}_j, \otimes_{m \neq j} \mathbf{\Sigma}_m)$. Then we directly bound $\|(\frac{1}{(n_k-1)p_{-j}} \sum_{i=1}^{n_k-1} \mathbf{Z}^i \left(\mathbf{Z}^i\right)^T - \mathbf{\Sigma}_j)\mathbf{a}\|_\infty$ for some vector $\mathbf{a}$. The results are given in the following lemma.

Lemma B.1. *For a vector* $\mathbf{a} \in \mathbb{R}^{p_j}$*, with probability at least* $1 - O(p_j^{-1})$*, we have*

$$\|(\widehat{\mathbf{\Sigma}}_j^{(k)} - \mathbf{\Sigma}_j)\mathbf{a}\|_\infty \lesssim \|\mathbf{a}\|_2 \sqrt{\frac{\log p_j}{(n_k - 1)p_{-j}}},$$

and

$$\|(\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j)\mathbf{a}\|_\infty \lesssim \|\mathbf{a}\|_2 \sqrt{\frac{\log p_j}{np_{-j}}}.$$

Proposition B.1 (Lyu et al. (2020), c.f Lemma S.1)**.** *Assume i.i.d. data* $\mathbf{X}^1, \ldots, \mathbf{X}^n \in \mathbb{R}^{p \times q}$ *follows the matrix-variate normal distribution such that* $\mathrm{vec}\left(\mathbf{X}^i\right) \sim \mathrm{N}\left(\mathbf{0}; \mathbf{\Psi} \otimes \mathbf{\Sigma}\right)$ *with* $\mathbf{\Psi} \in \mathbb{R}^{q \times q}$ *and* $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$*. Assume that* $0 < C_1 \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq 1/C_1 < \infty$ *and* $0 < C_2 \leq \lambda_{\min}(\mathbf{\Psi}) \leq \lambda_{\max}(\mathbf{\Psi}) \leq 1/C_2 < \infty$ *for some positive constants* C_1, C_2*. For two vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *such that* $\|\mathbf{a}\|_2 \leq C_3 \leq \infty$ *and* $\|\mathbf{b}\|_2 \leq C_4 \leq \infty$*, with high probability, we have*

$$\mathrm{P}\left\{\left|\frac{1}{np}\sum_{i=1}^{n} \mathbf{a}^T(\mathbf{X}^i)^T\mathbf{X}^i\mathbf{b} - \frac{1}{p}\mathbf{a}^T\mathrm{E}\left((\mathbf{X}^i)^T\mathbf{X}^i\right)\mathbf{b}\right| \geq t\|\mathbf{a}\|_2\|\mathbf{b}\|_2\right\} \leq c_1 \exp(-c_2npt^2).$$

Proposition B.2 (Kollo (2005))**.** *For tensors* $\mathbf{A}$*,* $\mathbf{B}$*,* $\mathbf{C}$ *and* $\mathbf{D}$*, we have the following basic properties of Kronecker product:*

(i) $(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T$*;*

(ii) $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$*;*

(iii) $(k\mathbf{A} \otimes \mathbf{B}) = \mathbf{A} \otimes (k\mathbf{B}) = k(\mathbf{A} \otimes \mathbf{B})$*;*

(iv) $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$*;*

(v) $\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}$.

Proof of Lemma B.1. Let $\mathbf{X}^{i_1}, \ldots, \mathbf{X}^{i_{n_k}}$ denote the n_k independent samples from the kth class where $Y^{i_l} = k$ for $l = 1, \ldots, n_k$. By definition, we have

$$\widehat{\mathbf{\Sigma}}_j^{(k)} = \frac{1}{(n_k - 1)p_{-j}} \sum_{l=1}^{n_k} (\mathbf{X}^{i_l} - \overline{\mathbf{X}}_k)_{(j)} \left(\mathbf{X}^{i_l} - \overline{\mathbf{X}}_k\right)_{(j)}^T. \tag{B.8}$$

Next we show that $\widehat{\boldsymbol{\Sigma}}_j^{(k)}$ can be rewritten as the summation of $n_k - 1$ independent matrix variables such that

$$\widehat{\boldsymbol{\Sigma}}_j^{(k)} = \frac{1}{(n_k-1)p_{-j}} \sum_{i=1}^{n_k-1} \mathbf{Z}^i \left(\mathbf{Z}^i\right)^T,$$

where $\mathbf{Z}^i$ will be specified later. We stack $\mathbf{X}_{(j)}^{i_l}$ together to get $\widetilde{\mathbf{X}}_{(j)} \in \mathbb{R}^{n_k p_{-j} \times p_j}$ such that

$$\widetilde{\mathbf{X}}_{(j)} = \begin{pmatrix} (\mathbf{X}_{(j)}^{i_1})^T \\ \vdots \\ (\mathbf{X}_{(j)}^{i_{n_k}})^T \end{pmatrix}.$$

By treating each $(\mathbf{X}_{(j)}^{i_l})^T$ as a block matrix and applying block matrix multiplication, the sample mean can be written as

$$\overline{\mathbf{X}}_{k(j)}^T = \frac{1}{n_k}(\mathbf{1}_{n_k}^T \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)}.$$

Denote $\mathbf{G} = \mathbf{I}_{n_k} - \frac{1}{n_k}\mathbf{1}_{n_k}\mathbf{1}_{n_k}^T$ as the centering matrix. Hence, the centered data can be expressed as

$$\begin{aligned}
\begin{pmatrix} (\mathbf{X}_{(j)}^{i_1})^T - \overline{\mathbf{X}}_{k(j)}^T \\ \vdots \\ (\mathbf{X}_{(j)}^{i_{n_k}})^T - \overline{\mathbf{X}}_{k(j)}^T \end{pmatrix} &= \widetilde{\mathbf{X}}_{(j)} - (\mathbf{1}_{n_k} \otimes \mathbf{I}_{p_{-j}})\overline{\mathbf{X}}_{k(j)}^T \\
&= (\mathbf{I}_{n_k} \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)} - \frac{1}{n_k}(\mathbf{1}_{n_k} \otimes \mathbf{I}_{p_{-j}})(\mathbf{1}_{n_k}^T \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)} \\
&= (\mathbf{I}_{n_k} \otimes \mathbf{I}_{p_{-j}} - \frac{1}{n_k}(\mathbf{1}_{n_k}\mathbf{1}_{n_k}^T \otimes \mathbf{I}_{p_{-j}}))\widetilde{\mathbf{X}}_{(j)} \\
&= ((\mathbf{I}_{n_k} - \frac{1}{n_k}\mathbf{1}_{n_k}\mathbf{1}_{n_k}^T) \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)} \\
&= (\mathbf{G} \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)}
\end{aligned}$$

where we use properties in Proposition B.2. Using the fact that $\mathbf{G}^T\mathbf{G} = \mathbf{G}$, (B.8) can be further expressed as

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_j^{(k)} &= \frac{1}{(n_k-1)p_{-j}} \left\{(\mathbf{G} \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)}\right\}^T \left\{(\mathbf{G} \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)}\right\} \\
&= \frac{1}{(n_k-1)p_{-j}} \widetilde{\mathbf{X}}_{(j)}^T(\mathbf{G}^T \otimes \mathbf{I}_{p_{-j}})(\mathbf{G} \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)} \\
&= \frac{1}{(n_k-1)p_{-j}} \widetilde{\mathbf{X}}_{(j)}^T(\mathbf{G} \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)} \\
&= \frac{1}{(n_k-1)p_{-j}} \widetilde{\mathbf{X}}_{(j)}^T(\mathbf{H}_{-1}^T\mathbf{H}_{-1} \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)} \\
&= \frac{1}{(n_k-1)p_{-j}} \widetilde{\mathbf{X}}_{(j)}^T(\mathbf{H}_{-1}^T \otimes \mathbf{I}_{p_{-j}})(\mathbf{H}_{-1} \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)}
\end{aligned}$$

where $\mathbf{H}_{-1}$ is the sub-matrix of an orthogonal matrix $\mathbf{H}(n_k)$ without the first row for $\mathbf{H}(n)$ defined as

$$\mathbf{H}(n) = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{1}{\sqrt{n(n-1)}} & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix}.$$

Let $\mathbf{h}_i$ denote the ith row of $\mathbf{H}_{-1}$. Therefore we have $(\mathbf{Z}^i)^T = (\mathbf{h}_i \otimes \mathbf{I}_{p_{-j}})\widetilde{\mathbf{X}}_{(j)}$. Since $\mathbf{h}_i\mathbf{h}_s^T = 0$ for $i \neq s$ and the n_k samples are independent, $\mathbf{Z}^i$ is independent of $\mathbf{Z}^s$. In addition, $\mathbf{Z}^i$ follows the matrix normal distribution $\mathrm{MN}(\mathbf{0}; \mathbf{\Sigma}_j, \otimes_{m\neq j}\mathbf{\Sigma}_m)$. Let $\{\mathbf{e}_l, 1 \leq l \leq p_j\}$ be the standard basis of Euclidean space $\mathbb{R}^{p_j}$. For fixed $\mathbf{e}_l$, Proposition B.1 implies

$$\mathrm{P}\left\{\left|\frac{1}{(n_k-1)p_{-j}}\sum_{i=1}^{n_k-1}\mathbf{e}_l^T\mathbf{Z}^i(\mathbf{Z}^i)^T\mathbf{a} - \frac{1}{p_{-j}}\mathbf{e}_l^T\mathrm{E}\left((\mathbf{Z}^1\mathbf{Z}^1)^T\right)\mathbf{a}\right| \geq t\|\mathbf{a}\|_2\right\} \leq c_1\exp\{-c_2(n_k-1)p_{-j}t^2\}.$$

By Lemma 2 in Pan et al. (2019),

$$\frac{1}{p_{-j}}\mathrm{E}\left\{(\mathbf{Z}^1\mathbf{Z}^1)^T\right\} = \frac{1}{p_{-j}}\mathbf{\Sigma}_j \cdot \mathrm{tr}(\otimes_{m\neq j}\mathbf{\Sigma}_m) = \mathbf{\Sigma}_j$$

since the diagonal elements are all ones by assumption. Therefore, we have

$$\begin{aligned} &\mathrm{P}\left\{\|(\widehat{\mathbf{\Sigma}}_j^{(k)} - \mathbf{\Sigma}_j)\mathbf{a}\|_\infty \geq t\|\mathbf{a}\|_2\right\} \\ \leq &p_j\mathrm{P}\left\{\left|\frac{1}{(n_k-1)p_{-j}}\sum_{i=1}^{n_k-1}\mathbf{e}_l^T\mathbf{Z}^i(\mathbf{Z}^i)^T\mathbf{a} - \frac{1}{p_{-j}}\mathbf{e}_l^T\mathrm{E}\left((\mathbf{Z}^1)^T\mathbf{Z}^1\right)\mathbf{a}\right| \geq t\|\mathbf{a}\|_2\right\} \\ \leq &c_1p_j\exp\{-c_2(n_k-1)p_{-j}t^2\}. \end{aligned}$$

By taking $t = \sqrt{\frac{2\log p_j}{c_2(n_k-1)p_{-j}}}$, we have that

$$\mathrm{P}\left(\|(\widehat{\mathbf{\Sigma}}_j^{(k)} - \mathbf{\Sigma}_j)\mathbf{a}\|_\infty \geq \sqrt{\frac{2\log p_j}{c_2(n_k-1)p_{-j}}}\|\mathbf{a}\|_2\right) \leq c_1p_j^{-1}.$$

Therefore, we show that the first conclusion holds with probability at least $1 - O(p_j^{-1})$. Since $n_k \asymp n$, by definition, we have with probability at least $1 - O(p_j^{-1})$

$$\begin{aligned} \|(\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j)\mathbf{a}\|_\infty &= \|\sum_{k=1}^K\frac{n_k-1}{n-K}(\widehat{\mathbf{\Sigma}}_j^{(k)} - \mathbf{\Sigma}_j)\mathbf{a}\|_\infty \\ &\leq \sum_{k=1}^K\frac{n_k-1}{n-K}\|(\widehat{\mathbf{\Sigma}}_j^{(k)} - \mathbf{\Sigma}_j)\mathbf{a}\|_\infty \\ &\lesssim \|\mathbf{a}\|_2\sqrt{\frac{\log p_j}{(n_k-1)p_{-j}}} \\ &\lesssim \|\mathbf{a}\|_2\sqrt{\frac{\log p_j}{np_{-j}}}. \end{aligned}$$

□

Lemma B.2. *Suppose we have three matrices* $\mathbf{B} \in \mathbb{R}^{p_1 \times p_1}$, $\mathbf{C} \in \mathbb{R}^{p_2 \times p_2}$ *and* $\mathbf{D} \in \mathbb{R}^{p_3 \times p_3}$. *Assume that* $\|\mathbf{Cu}\|_\infty \leq \|\mathbf{u}\|_2 \epsilon$ *for any* $\mathbf{u} \in \mathbb{R}^{p_2}$ *and some* $\epsilon > 0$. *We have the following conclusions:*

(i) $\|(\mathbf{B} \otimes \mathbf{C})\mathbf{a}\|_\infty \leq \|\mathbf{B}\|_{\max} \|\mathbf{a}\|_2 \sqrt{p_1} \epsilon$ *for any* $\mathbf{a} \in \mathbb{R}^{p_1 p_2}$*;*

(ii) $\|(\mathbf{C} \otimes \mathbf{B})\mathbf{a}\|_\infty \leq \|\mathbf{B}\|_{\max} \|\mathbf{a}\|_2 \sqrt{p_1} \epsilon$ *for any* $\mathbf{a} \in \mathbb{R}^{p_1 p_2}$*;*

(iii) $\|(\mathbf{B} \otimes \mathbf{C} \otimes \mathbf{D})\mathbf{b}\|_\infty \leq \|\mathbf{B}\|_{\max} \|\mathbf{D}\|_{\max} \|\mathbf{b}\|_2 \sqrt{p_1 p_3} \epsilon$ *for any* $\mathbf{b} \in \mathbb{R}^{p_1 p_2 p_3}$.

Proof of Lemma B.2. We prove the first result. Let $\mathbf{a} \in \mathbb{R}^{p_1 p_2}$. Divide $\mathbf{a}$ into p_1 sub-vectors of equal length such that $\mathbf{a} = (\mathbf{a}_1^T, \ldots, \mathbf{a}_{p_1}^T)^T$ and $\mathbf{a}_j \in \mathbb{R}^{p_2}$. We have

$$(\mathbf{B} \otimes \mathbf{C})\mathbf{a} = \begin{pmatrix} b_{11}\mathbf{C} & \cdots & b_{1n}\mathbf{C} \\ \vdots & \ddots & \vdots \\ b_{p_1 1}\mathbf{C} & \cdots & b_{p_1 p_1}\mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_{p_1} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{p_1} b_{1j}\mathbf{C}\mathbf{a}_j \\ \vdots \\ \sum_{j=1}^{p_1} b_{p_1 j}\mathbf{C}\mathbf{a}_j \end{pmatrix}.$$

Then we have

$$\begin{aligned} \|(\mathbf{B} \otimes \mathbf{C})\mathbf{a}\|_\infty &= \max_i \| \sum_{j=1}^{p_1} b_{ij}\mathbf{C}\mathbf{a}_j \|_\infty \leq \max_i \sum_{j=1}^{p_1} \|b_{ij}\mathbf{C}\mathbf{a}_j\|_\infty \leq \max_i \max_j |b_{ij}| \sum_{j=1}^{p_1} \|\mathbf{C}\mathbf{a}_j\|_\infty \\ &\leq \|\mathbf{B}\|_{\max} \sum_{j=1}^{p_1} \|\mathbf{C}\mathbf{a}_j\|_\infty \leq \|\mathbf{B}\|_{\max} \sum_{j=1}^{p_1} \|\mathbf{a}_j\|_2 \epsilon \leq \|\mathbf{B}\|_{\max} \|\mathbf{a}\|_2 \sqrt{p_1} \epsilon, \end{aligned}$$

where we use $\|\mathbf{C}\mathbf{a}_j\|_\infty \leq \|\mathbf{a}_j\|_2 \epsilon$ by assumption and we use Cauchy inequality in the last step.

Next, we show the second result. For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, denote $\mathbf{K}^{m,n}$ as the commutation matrix such that $\mathbf{K}^{m,n}\text{vec}(\mathbf{A}) = \text{vec}(\mathbf{A}^T)$. Note that $\mathbf{K}^{m,n} \in \mathbb{R}^{mn \times mn}$ has exactly one nonzero element that equals to 1 in each row and column. By Proposition 1.3.12 in Kollo (2005), we have

$$\mathbf{C} \otimes \mathbf{B} = \mathbf{K}^{p_1, p_1}(\mathbf{B} \otimes \mathbf{C})\mathbf{K}^{p_2, p_2}.$$

Therefore, we have

$$\begin{aligned} \|(\mathbf{C} \otimes \mathbf{B})\mathbf{a}\|_\infty &= \|\mathbf{K}^{p_1, p_1}(\mathbf{B} \otimes \mathbf{C})\mathbf{K}^{p_2, p_2}\mathbf{a}\|_\infty \\ &\leq \|\mathbf{K}^{p_1, p_1}\|_\infty \|(\mathbf{B} \otimes \mathbf{C})\mathbf{a}'\|_\infty \\ &\leq \|(\mathbf{B} \otimes \mathbf{C})\mathbf{a}'\|_\infty \\ &\leq \|\mathbf{B}\|_{\max} \|\mathbf{a}'\|_2 \sqrt{p_1} \epsilon \\ &= \|\mathbf{B}\|_{\max} \|\mathbf{a}\|_2 \sqrt{p_1} \epsilon, \end{aligned}$$

where $\mathbf{a}' = \mathbf{K}^{p_2, p_2}\mathbf{a}$ is a permutation of $\mathbf{a}$ and we use that fact that $\|\mathbf{a}'\|_2 = \|\mathbf{a}\|_2$.

To prove the third result, we first divide the vector $\mathbf{b} \in \mathbb{R}^{p_1 p_2 p_3}$ into p_1 sub-vectors of equal length such that $\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_{p_1}^T)^T$ and $\mathbf{b}_j \in \mathbb{R}^{p_2 p_3}$. Using the previous two conclusions, we have

$$\begin{aligned}
\|(\mathbf{B} \otimes \mathbf{C} \otimes \mathbf{D})\mathbf{b}\|_\infty &\le \|\mathbf{B}\|_{\max} \sum_{j=1}^{p_1} \|(\mathbf{C} \otimes \mathbf{D})\mathbf{b}_j\|_\infty \\
&\le \|\mathbf{B}\|_{\max} \sum_{j=1}^{p_1} \|\mathbf{D}\|_{\max} \|\mathbf{b}_j\|_2 \sqrt{p_3}\epsilon \\
&\le \|\mathbf{B}\|_{\max} \|\mathbf{D}\|_{\max} \sqrt{p_3} \sum_{j=1}^{p_1} \|\mathbf{b}_j\|_2 \epsilon \\
&\le \|\mathbf{B}\|_{\max} \|\mathbf{D}\|_{\max} \sqrt{p_1 p_3} \|\mathbf{b}\|_2 \epsilon.
\end{aligned}$$

□

Proof of Lemma A.3. By triangle inequality, we have

$$\|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\boldsymbol{\beta}_k\|_\infty \le \sum_{m=1}^{M} \left\| \left(\mathbf{\Sigma}_M \otimes \cdots \otimes \mathbf{\Sigma}_{m+1} \otimes (\widehat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m) \otimes \widehat{\mathbf{\Sigma}}_{m-1} \otimes \cdots \otimes \widehat{\mathbf{\Sigma}}_1 \right) \boldsymbol{\beta}_k \right\|_\infty .$$

Since $\lambda_{\max}(\mathbf{\Sigma}_j) \le C_\lambda$, we have that

$$\|\mathbf{\Sigma}_M \otimes \cdots \otimes \mathbf{\Sigma}_{m+1}\|_{\max} \le \prod_{j>m} \|\mathbf{\Sigma}_j\|_{\max} \le C_\lambda^{M-m}. \tag{B.9}$$

By Lemma A.2, we have that with probability at least $1 - O(p^{-1})$, $\|\widehat{\mathbf{\Sigma}}_j\|_{\max} \le C_\lambda + C\sqrt{\frac{\log p_j}{np_{-j}}}$. Since $\sqrt{\frac{\log p_j}{np_{-j}}} \le C_1$ for some constant C_1, there exists a constant C_2 such that $\|\widehat{\mathbf{\Sigma}}_j\|_{\max} \le C_2$. Similarly, we have

$$\|\widehat{\mathbf{\Sigma}}_{m-1} \otimes \cdots \otimes \widehat{\mathbf{\Sigma}}_1\|_{\max} \le \prod_{j<m} \|\widehat{\mathbf{\Sigma}}_j\|_{\max} \le C_2^{m-1}. \tag{B.10}$$

With the results in (B.9), (B.10) and Lemma B.1, applying Lemma B.2, we have that

$$\begin{aligned}
\|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\boldsymbol{\beta}_k\|_\infty &\le \sum_{m=1}^{M} \left\| \left(\mathbf{\Sigma}_M \otimes \cdots \otimes \mathbf{\Sigma}_{m+1} \otimes (\widehat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m) \otimes \widehat{\mathbf{\Sigma}}_{m-1} \otimes \cdots \otimes \widehat{\mathbf{\Sigma}}_1 \right) \boldsymbol{\beta}_k \right\|_\infty \\
&\lesssim \sum_{m=1}^{M} \prod_{j>m} \|\mathbf{\Sigma}_j\|_{\max} \prod_{j<m} \|\widehat{\mathbf{\Sigma}}_j\|_{\max} \sqrt{p_{-m}} \|\boldsymbol{\beta}_k\|_2 \sqrt{\frac{\log p_m}{np_{-m}}} \\
&\lesssim \sum_{m=1}^{M} \|\boldsymbol{\beta}_k\|_2 \sqrt{\frac{\log p_m}{n}} \\
&\lesssim \|\boldsymbol{\beta}_k\|_2 \sqrt{\frac{\log p}{n}}.
\end{aligned}$$

Therefore, with probability at least $1 - O(p^{-1})$, we have

$$\|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\boldsymbol{\beta}_k\|_\infty \lesssim \|\boldsymbol{\beta}_k\|_2 \sqrt{\frac{\log p}{n}}.$$

□

B.3 Proof of Lemma A.4

Proof of Lemma A.4. Since $\widehat{\boldsymbol{\beta}}$ is the minimizer to the objective function A.1, we have that

$$\begin{aligned}
\lambda\sum_{j=1}^{p}(\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2-\|\boldsymbol{\beta}_{\cdot j}\|_2) &\le \sum_{k=2}^{K}\Big\{\boldsymbol{\beta}_k^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k-2\boldsymbol{\beta}_k^T\widehat{\boldsymbol{\delta}}_k-(\widehat{\boldsymbol{\beta}}_k^T\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}}_k-2\widehat{\boldsymbol{\beta}}_k^T\widehat{\boldsymbol{\delta}}_k)\Big\}\\
&=\sum_{k=2}^{K}\Big\{-(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)-2\langle\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k-\widehat{\boldsymbol{\delta}}_k,\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\rangle\Big\}\\
&\le\sum_{k=2}^{K}|\langle\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k-\widehat{\boldsymbol{\delta}}_k,\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\rangle|\\
&\le\sum_{k=2}^{K}\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k-\widehat{\boldsymbol{\delta}}_k\|_{\max}\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\|_1.
\end{aligned}$$

By conditions (i) and (ii), since $\boldsymbol{\beta}_k=\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_k$, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k-\widehat{\boldsymbol{\delta}}_k\|_\infty &\le \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_k-\boldsymbol{\Sigma}\boldsymbol{\beta}_k+\boldsymbol{\Sigma}\boldsymbol{\beta}_k-\widehat{\boldsymbol{\delta}}_k\|_\infty\\
&\le\|(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\beta}_k\|_\infty+\|\boldsymbol{\delta}_k-\widehat{\boldsymbol{\delta}}_k\|_\infty\\
&\lesssim\sqrt{\frac{\log p}{n}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lambda\sum_{j=1}^{p}(\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2-\|\boldsymbol{\beta}_{\cdot j}\|_2) &\le C\sqrt{\frac{\log p}{n}}\sum_{k=2}^{K}\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\|_1\\
&\le C\sqrt{\frac{\log p}{n}}\sqrt{K-1}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\beta}}_{\cdot j}-\boldsymbol{\beta}_{\cdot j}\|_2\\
&=C\sqrt{K-1}\sqrt{\frac{\log p}{n}}\sum_{j=1}^{p}\|\Upsilon_{\cdot j}\|_2. \qquad \text{(B.11)}
\end{aligned}$$

On the other hand, letting $S=\mathrm{supp}(\boldsymbol{\beta})$, we have

$$\begin{aligned}
\lambda\sum_{j=1}^{p}(\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2-\|\boldsymbol{\beta}_{\cdot j}\|_2) &= \lambda\sum_{j\in S}(\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2-\|\boldsymbol{\beta}_{\cdot j}\|_2)+\lambda\sum_{j\in S^c}\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2\\
&\ge\lambda\sum_{j\in S^c}\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2-\lambda\sum_{j\in S}\Big|\|\widehat{\boldsymbol{\beta}}_{\cdot j}\|_2-\|\boldsymbol{\beta}_{\cdot j}\|_2\Big|\\
&\ge\lambda\sum_{j\in S^c}\|\Upsilon_{\cdot j}\|_2-\lambda\sum_{j\in S}\|\Upsilon_{\cdot j}\|_2. \qquad \text{(B.12)}
\end{aligned}$$

Taking $\lambda\ge 3C\sqrt{K-1}\sqrt{\frac{\log p}{n}}$, then (B.11) together with (B.12) give us

$$\sum_{j\in S^c}\|\Upsilon_{\cdot j}\|_2-\sum_{j\in S}\|\Upsilon_{\cdot j}\|_2\le\frac{1}{3}\sum_{j=1}^{p}\|\Upsilon_{\cdot j}\|_2=\frac{1}{3}(\sum_{j\in S^c}\|\Upsilon_{\cdot j}\|_2+\sum_{j\in S}\|\Upsilon_{\cdot j}\|_2),$$

which implies

$$\sum_{j\in S^c}\|\Upsilon_{\cdot j}\|_2 \le 2\sum_{j\in S}\|\Upsilon_{\cdot j}\|_2.$$

□

B.4 Proof of Lemma A.5

To prove Lemma A.5, we first prove two important technical results that are used in the proof.

Lemma B.3. *Suppose* $\mathbf{X}^1, \ldots, \mathbf{X}^n \in \mathbb{R}^{p\times q}$ *are independent samples from the matrix normal distribution* $\mathrm{MN}(\mathbf{0}; \mathbf{\Sigma}, \mathbf{\Psi})$. *Suppose that* $C_\lambda^{-1} \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le C_\lambda$ *and* $C_\lambda^{-1} \le \lambda_{\min}(\mathbf{\Psi}) \le \lambda_{\max}(\mathbf{\Psi}) \le C_\lambda$ *for some constant* $C_\lambda > 1$. *Assume that* $\frac{s\log p}{nq} \le C_1$ *for some constant* $C_1 > 0$. *Let* $\widehat{\mathbf{\Sigma}} = \frac{1}{nq}\sum_{i=1}^n \mathbf{X}^i\left(\mathbf{X}^i\right)^T$. *Then, for any* $C' > 0$, *there exists a constant* $C > 0$ *only depending on* C_1, C', C_λ, *such that for* $\epsilon_s = \sqrt{\frac{s\log p}{nq}}$, *we have*

$$C_\lambda^{-1} - C\epsilon_s \le \phi_{\min}^{\widehat{\mathbf{\Sigma}}}(s) \le \phi_{\max}^{\widehat{\mathbf{\Sigma}}}(s) \le C_\lambda + C\epsilon_s,$$

with probability at least $1 - C'\exp(-s\log p)$.

Proof of Lemma B.3. Let $T \subseteq \{1, \ldots, p\}$ be a set of indices with cardinality s. Let $\mathbf{X}^i_{T\cdot}$ be a sub-matrix of $\mathbf{X}^i$ such that the rows with indices in T are kept. We use $\mathbf{X}^i_T$ for short in this proof.

By Lemma 5.2 in Vershynin (2010), we can find a $\frac{1}{4}$-covering with the Euclidean metric of the unit Euclidean sphere $\mathcal{S}^{p-1} = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^p, \|\mathbf{x}\|_2 = 1\}$ such that the covering number satisfies

$$\mathcal{N}(\mathcal{S}^{p-1}, \frac{1}{4}) \le 9^p.$$

For fixed T, by applying Proposition B.1 and union bound, we have

$$\begin{aligned}\mathrm{P}\left\{\max_{\mathbf{a}\in\mathcal{S}^{s-1}}\left|\frac{1}{nq}\sum_{i=1}^n \mathbf{a}^T\mathbf{X}^i_T(\mathbf{X}^i_T)^T\mathbf{a} - \mathbf{a}^T\mathbf{\Sigma}_{T,T}\mathbf{a}\right| \ge t\right\} \le& C_1 9^s\exp(-C_2 nqt^2)\\ =& C_1\exp(s\log 9 - C_2 nqt^2).\end{aligned}$$

By Lemma 5.4 in Vershynin (2010),

$$\left\|\widehat{\mathbf{\Sigma}}_{T,T} - \mathbf{\Sigma}_{T,T}\right\|_2 \le 2\sup_{\mathbf{a}\in\mathcal{N}(\mathcal{S}^{s-1},\frac{1}{4})}\left|\mathbf{a}^T(\widehat{\mathbf{\Sigma}}_{T,T} - \mathbf{\Sigma}_{T,T})\mathbf{a}\right|.$$

Therefore, we have

$$\begin{aligned}\mathrm{P}\left(\left\|\widehat{\mathbf{\Sigma}}_{T,T} - \mathbf{\Sigma}_{T,T}\right\|_2 \ge t\right) \le& \mathrm{P}\left\{2\max_{\mathbf{a}\in\mathcal{S}^{s-1}}\left|\frac{1}{nq}\sum_{i=1}^n \mathbf{a}^T\mathbf{X}^i_T(\mathbf{X}^i_T)^T\mathbf{a} - \mathbf{a}^T\mathbf{\Sigma}_{T,T}\mathbf{a}\right| \ge t\right\}\\ =& C_1\exp\left(s\log 9 - \frac{C_2}{4}nqt^2\right).\end{aligned}$$

Next, using the bound $\begin{pmatrix} p \\ s \end{pmatrix} < (ep/s)^s$, we have

$$\mathrm{P}\left(\sup_{T\subseteq\{1,\ldots,p\}} \left\|\widehat{\mathbf{\Sigma}}_{T,T} - \mathbf{\Sigma}_{T,T}\right\|_2 \geq t\right) \leq C_1(ep/s)^s \exp\left(s\log 9 - \frac{C_2}{4}nqt^2\right)$$
$$\leq C_1 \exp\left(C_3 s\log p - \frac{C_2}{4}nqt^2\right).$$

By taking $t = \sqrt{\frac{4(C_3+1)s\log p}{C_2 nq}}$, with probability greater than $1 - C_1\exp(-s\log p)$,

$$\max_{T\subseteq\{1,\ldots,p\}} \|\widehat{\mathbf{\Sigma}}_{T,T} - \mathbf{\Sigma}_{T,T}\|_2 \leq C\sqrt{\frac{s\log p}{nq}} \tag{B.13}$$

for some constant C. Since $C_\lambda^{-1} \leq \lambda_{\min}(\mathbf{\Sigma}_{T,T}) \leq \lambda_{\max}(\mathbf{\Sigma}_{T,T}) \leq C_\lambda$ for any T, from (B.13), we have

$$C_\lambda^{-1} - C\sqrt{\frac{s\log p}{nq}} \leq \phi_{\min}^{\widehat{\mathbf{\Sigma}}}(s) \leq \phi_{\max}^{\widehat{\mathbf{\Sigma}}}(s) \leq C_\lambda + C\sqrt{\frac{s\log p}{nq}}.$$

□

Lemma B.4. *For two positive definite matrices $\mathbf{B} \in \mathbb{R}^{p_1\times p_1}$ and $\mathbf{C} \in \mathbb{R}^{p_2\times p_2}$, we have*

$$\max_{\|\mathbf{a}\|_2=1,\|\mathbf{a}\|_0\leq s} \left|\mathbf{a}^T(\mathbf{B}\otimes\mathbf{C})\mathbf{a}\right| \leq \max_{\|\mathbf{a}_1\|_2=1,\|\mathbf{a}_1\|_0\leq s} \left|\mathbf{a}_1^T\mathbf{B}\mathbf{a}_1\right| \cdot \max_{\|\mathbf{a}_2\|_2=1,\|\mathbf{a}_2\|_0\leq s} \left|\mathbf{a}_2^T\mathbf{C}\mathbf{a}_2\right|.$$

where $\mathbf{a} \in \mathbb{R}^{p_1p_2}, \mathbf{a}_1 \in \mathbb{R}^{p_1}, \mathbf{a}_2 \in \mathbb{R}^{p_2}$.

Proof of Lemma B.4. For a vector $\mathbf{a} \in \mathbb{R}^{p_1p_2}$, we reshape $\mathbf{a}$ into a matrix $\mathbf{A} \in \mathbb{R}^{p_1\times p_2}$ such that $\mathrm{vec}(\mathbf{A}) = \mathbf{a}$. We have $\mathbf{a}^T(\mathbf{B}\otimes\mathbf{C})\mathbf{a} = \mathrm{tr}(\mathbf{A}^T\mathbf{B}\mathbf{A}\mathbf{C})$, where $\mathrm{tr}(\cdot)$ denotes the trace of the matrix. Since there are at most s nonzero elements in $\mathbf{a}$, let T_1, T_2 denote two index sets that contain the row indices and column indices of all the nonzero elements in $\mathbf{A}$ and we have $|T_1| \leq s, |T_2| \leq s$. Then we have

$$\mathrm{tr}(\mathbf{A}^T\mathbf{B}\mathbf{A}\mathbf{C}) = \mathrm{tr}(\mathbf{A}_{T_1,T_2}^T\mathbf{B}_{T_1,T_1}\mathbf{A}_{T_1,T_2}\mathbf{C}_{T_2,T_2}) = (\mathbf{a}')^T(\mathbf{B}_{T_1,T_1}\otimes\mathbf{C}_{T_2,T_2})\mathbf{a}'$$

where $\mathbf{a}' = \mathrm{vec}(\mathbf{A}_{T_1,T_2}^T)$. By Proposition 1.3.12 in Kollo (2005), we have that $\lambda_{\max}(\mathbf{B}_{T_1,T_1}\otimes\mathbf{C}_{T_2,T_2}) = \lambda_{\max}(\mathbf{B}_{T_1,T_1})\cdot\lambda_{\max}(\mathbf{C}_{T_2,T_2})$. Then we have

$$\begin{aligned}
\mathbf{a}^T(\mathbf{B}\otimes\mathbf{C})\mathbf{a} &= (\mathbf{a}')^T(\mathbf{B}_{T_1,T_1}\otimes\mathbf{C}_{T_2,T_2})\mathbf{a}' \\
&\leq \max_{\mathbf{t}\in\mathbb{R}^{|T_1||T_2|},\|\mathbf{t}\|_2=1} \mathbf{t}^T(\mathbf{B}_{T_1,T_1}\otimes\mathbf{C}_{T_2,T_2})\mathbf{t} \\
&= \max_{\mathbf{t}_1\in\mathbb{R}^{|T_1|},\|\mathbf{t}_1\|_2=1} \mathbf{t}_1^T\mathbf{B}_{T_1,T_1}\mathbf{t}_1 \cdot \max_{\mathbf{t}_2\in\mathbb{R}^{|T_2|},\|\mathbf{t}_2\|_2=1} \mathbf{t}_2^T\mathbf{C}_{T_2,T_2}\mathbf{t}_2 \\
&\leq \max_{\|\mathbf{a}_1\|_2=1,\|\mathbf{a}_1\|_0\leq s} \left|\mathbf{a}_1^T\mathbf{B}\mathbf{a}_1\right| \cdot \max_{\|\mathbf{a}_2\|_2=1,\|\mathbf{a}_2\|_0\leq s} \left|\mathbf{a}_2^T\mathbf{C}\mathbf{a}_2\right|.
\end{aligned}$$

Therefore, we complete the proof. □

Proof of Lemma A.5. By Lemma B.3, with probability greater than $1 - O(p^{-1})$, we have

$$C_\lambda^{-1} - C\epsilon_s \le \phi_{\min}^{\widehat{\mathbf{\Sigma}}_j^{(k)}}(s) \le \phi_{\max}^{\widehat{\mathbf{\Sigma}}_j^{(k)}}(s) \le C_\lambda + C\epsilon_s,$$

where $\epsilon_s = \sqrt{\frac{s \log p_j}{(n_k-1)p_{-j}}}$. Moreover, we have

$$\begin{aligned}
\max_{\|\mathbf{a}\|_2=1, \|\mathbf{a}\|_0 \le s} \left| \mathbf{a}^T(\widehat{\mathbf{\Sigma}}_j - \mathbf{\Sigma}_j)\mathbf{a} \right| &= \max_{\|\mathbf{a}\|_2=1, \|\mathbf{a}\|_0 \le s} \left| \sum_{k=1}^K \frac{n_k - 1}{n - K} \mathbf{a}^T(\widehat{\mathbf{\Sigma}}_j^{(k)} - \mathbf{\Sigma}_j)\mathbf{a} \right| \\
&\le \sum_{k=1}^K \frac{n_k - 1}{n - K} \max_k \max_{\|\mathbf{a}\|_2=1, \|\mathbf{a}\|_0 \le s} \left| \mathbf{a}^T(\widehat{\mathbf{\Sigma}}_j^{(k)} - \mathbf{\Sigma}_j)\mathbf{a} \right| \\
&\lesssim \max_k \sqrt{\frac{s \log p_j}{(n_k - 1)p_{-j}}} \\
&\lesssim \sqrt{\frac{s \log p_j}{np_{-j}}}. \qquad \text{(B.14)}
\end{aligned}$$

Next, we bound the restricted eigenvalue of $\widehat{\mathbf{\Sigma}}$. By triangle inequality, we have

$$\begin{aligned}
\left| \mathbf{a}^T(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{a} \right| &= \left| \mathbf{a}^T(\otimes_{m=M}^1 \widehat{\mathbf{\Sigma}}_m - \otimes_{m=M}^1 \mathbf{\Sigma}_m)\mathbf{a} \right| \\
&\le \sum_{m=1}^M \left| \mathbf{a}^T \left(\mathbf{\Sigma}_M \otimes \cdots \otimes \mathbf{\Sigma}_{m+1} \otimes (\widehat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m) \otimes \widehat{\mathbf{\Sigma}}_{m-1} \otimes \cdots \otimes \widehat{\mathbf{\Sigma}}_1 \right) \mathbf{a} \right|.
\end{aligned}$$

By Lemma B.4, we have that

$$\begin{aligned}
&\max_{\|\mathbf{a}\|_2=1, \|\mathbf{a}\|_0 \le s} \left| \mathbf{a}^T \left(\mathbf{\Sigma}_M \otimes \cdots \otimes \mathbf{\Sigma}_{m+1} \otimes (\widehat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m) \otimes \widehat{\mathbf{\Sigma}}_{m-1} \otimes \cdots \otimes \widehat{\mathbf{\Sigma}}_1 \right) \mathbf{a} \right| \\
&\le \prod_{j>m} \max_{\|\mathbf{a}_j\|_2=1, \|\mathbf{a}_j\|_0 \le s} \mathbf{a}_j^T \mathbf{\Sigma}_j \mathbf{a}_j \cdot \max_{\|\mathbf{a}_m\|_2=1, \|\mathbf{a}_m\|_0 \le s} \left| \mathbf{a}_m^T(\widehat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m)\mathbf{a}_m \right| \cdot \prod_{j<m} \max_{\|\mathbf{a}_j\|_2=1, \|\mathbf{a}_j\|_0 \le s} \mathbf{a}_j^T \widehat{\mathbf{\Sigma}}_j \mathbf{a}_j.
\end{aligned}$$

Since $\lambda_{\max}(\mathbf{\Sigma}_m) \le C_\lambda$, using the result in (B.14), we have that

$$\max_{\|\mathbf{a}_m\|_2=1, \|\mathbf{a}_m\|_0 \le s} \left| \mathbf{a}_m^T(\widehat{\mathbf{\Sigma}}_m - \mathbf{\Sigma}_m)\mathbf{a}_m \right| \lesssim \sqrt{\frac{s \log p_m}{np_{-m}}} \lesssim \sqrt{\frac{s \log p}{n}},$$

and

$$\begin{aligned}
\max_{\|\mathbf{a}_j\|_2=1, \|\mathbf{a}_j\|_0 \le s} \mathbf{a}_j^T \mathbf{\Sigma}_j \mathbf{a}_j &\le C_\lambda, \\
\max_{\|\mathbf{a}_j\|_2=1, \|\mathbf{a}_j\|_0 \le s} \mathbf{a}_j^T \widehat{\mathbf{\Sigma}}_j \mathbf{a}_j &\le C_\lambda + c\sqrt{\frac{s \log p}{n}}.
\end{aligned}$$

Therefore, we have

$$\max_{\|\mathbf{a}\|_2=1, \|\mathbf{a}\|_0 \le s} \left| \mathbf{a}^T(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{a} \right| \lesssim C_\lambda^{M-m} \sqrt{\frac{s \log p}{n}} \left(C_\lambda + c\sqrt{\frac{s \log p}{n}} \right)^{m-1} \lesssim \sqrt{\frac{s \log p}{n}}.$$

Since $C_\lambda^{-M} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq C_\lambda^M$, we have that, with probability greater than $1 - O(p^{-1})$

$$C_\lambda^{-M} - C\epsilon_s \leq \phi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s) \leq \phi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(s) \leq C_\lambda^M + C\epsilon_s,$$

where $\epsilon_s = \sqrt{\frac{s\log p}{n}}$. □

B.5 Proof of Lemmas A.6

Proof of Lemma A.6. Note that $\Upsilon = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. We have

$$\sum_{k=2}^{K} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^T \widehat{\boldsymbol{\Sigma}} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) = \operatorname{tr}(\Upsilon \widehat{\boldsymbol{\Sigma}} \Upsilon^T) = \|\widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \Upsilon^T\|_F^2.$$

We follow the proof of Theorem 4.2 in Gao et al. (2017) to decompose the matrix $\Upsilon^T = (\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2, \ldots, \widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K) \in \mathbb{R}^{p\times(K-1)}$ in the following way. Let $S = \operatorname{supp}(\boldsymbol{\beta})$ and $J_0 = S$. Let $J_1 = \{j_1, \ldots, j_t\} \subset S^c$ correspond to the t rows with the largest ℓ_2 norm in $\Upsilon^T_{S^c\cdot}$ and we define $\widetilde{S} = S \cup J_1$. Then we partition $\widetilde{S}^c$ into disjoint subsets $J_2, \ldots, J_L$ of size t (with $J_L \leq t$), such that J_l is the set of indices corresponding to the t rows with the largest ℓ_2 norm in Υ^T outside $\widetilde{S} \cup \bigcup_{j=2}^{l-1} J_j$. Hence we have $\widetilde{S}^c = \bigcup_{j=2}^{L} J_j$, where $L > 2$, $J_l = t$, $\forall l = 2, \ldots, L-1$ and $J_L \leq t$. Then by triangle inequality, we have

$$\|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \Upsilon^T\|_F \geq \|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \Upsilon^T_{\widetilde{S}\cdot}\|_F - \sum_{l\geq 2} \|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \Upsilon^T_{J_l\cdot}\|_F.$$

Note that

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \Upsilon^T_{\widetilde{S}\cdot}\|_F = \sqrt{\operatorname{tr}(\Upsilon_{\widetilde{S}\cdot} \widehat{\boldsymbol{\Sigma}} \Upsilon^T_{\widetilde{S}\cdot})} = \sqrt{\sum_{k=2}^{K} (\boldsymbol{v}_k^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{v}_k)} \\ \geq \sqrt{\phi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s+t) \sum_{k=2}^{K} \|\boldsymbol{v}_k\|_2^2} = \sqrt{\phi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s+t)} \|\Upsilon^T_{\widetilde{S}\cdot}\|_F\end{aligned}$$

where $\boldsymbol{v}_k$ denotes the kth column of $\Upsilon^T_{\widetilde{S}\cdot}$. Similarly, we have

$$\|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \Upsilon^T_{J_l\cdot}\|_F \leq \sqrt{\phi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(t)} \|\Upsilon^T_{J_l\cdot}\|_F.$$

Thus,

$$\|\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \Upsilon^T\|_F \geq \sqrt{\phi_{\min}^{\widehat{\boldsymbol{\Sigma}}}(s+t)} \|\Upsilon^T_{\widetilde{S}\cdot}\|_F - \sqrt{\phi_{\max}^{\widehat{\boldsymbol{\Sigma}}}(t)} \sum_{l\geq 2} \|\Upsilon^T_{J_l\cdot}\|_F.$$

Moreover,

$$\begin{aligned}\sum_{l\geq 2} \|\Upsilon^T_{J_l\cdot}\|_F &\leq \sqrt{t} \sum_{l\geq 2} \max_{j\in J_l} \|\Upsilon^T_{j\cdot}\|_2 \leq \sqrt{t} \sum_{l\geq 2} \frac{1}{t} \sum_{j\in J_{l-1}} \|\Upsilon^T_{j\cdot}\|_2 && \text{(B.15)} \\ &\leq t^{-1/2} \sum_{j\in S^c} \|\Upsilon^T_{j\cdot}\|_2 \leq 2t^{-1/2} \sum_{j\in S} \|\Upsilon^T_{j\cdot}\|_2 && \text{(B.16)} \\ &\leq 2\sqrt{\frac{s}{t}} \sqrt{\sum_{j\in S} \|\Upsilon^T_{j\cdot}\|_2^2} \leq 2\sqrt{\frac{s}{t}} \|\Upsilon^T_{\widetilde{S}\cdot}\|_F. && \text{(B.17)}\end{aligned}$$

where we use the condition the $\sum_{j\in S^c}\|\Upsilon_{j\cdot}^T\| \le 2\sum_{j\in S}\|\Upsilon_{j\cdot}^T\|$ in (B.16). In (B.15), since the rows in $\Upsilon_{S^c\cdot}^T$ are sorted in descending order in terms of ℓ_2 norm, we have $\max_{j\in J_l}\|\Upsilon_{j\cdot}^T\|_2 \le \frac{1}{t}\sum_{j\in J_{l-1}}\|\Upsilon_{j\cdot}^T\|_2$. Hence,

$$\|\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}\Upsilon^T\|_F \ge \left(\sqrt{\phi_{\min}^{\widehat{\mathbf{\Sigma}}}(s+t)} - 2\sqrt{\frac{s}{t}}\sqrt{\phi_{\max}^{\widehat{\mathbf{\Sigma}}}(t)}\right)\|\Upsilon_{\widetilde{S}\cdot}^T\|_F.$$

Let $t = c_1 s$ for some constant c_1. By Condition (ii) that $C - c\epsilon_s \le \phi_{\min}^{\widehat{\mathbf{\Sigma}}}(s) \le \phi_{\max}^{\widehat{\mathbf{\Sigma}}}(s) \le C + c\epsilon_s$, we have

$$\kappa = \sqrt{\phi_{\min}^{\widehat{\mathbf{\Sigma}}}(s+t)} - 2\sqrt{\frac{s}{t}}\sqrt{\phi_{\max}^{\widehat{\mathbf{\Sigma}}}(t)} \ge \sqrt{\frac{1}{C} - O(\epsilon_s)} - 2\sqrt{\frac{1}{c_1}\left\{C + O(\epsilon_s)\right\}}.$$

Since $\epsilon_s = o(1)$, choosing c_1 to be sufficiently large, κ can be lower bounded by some constant only depending on C. Hence,

$$\|\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}\Upsilon^T\|_F \gtrsim \|\Upsilon_{\widetilde{S}\cdot}^T\|_F.$$

Since

$$\|\Upsilon^T\|_F \le \|\Upsilon_{\widetilde{S}\cdot}^T\|_F + \sum_{l\ge 2}\|\Upsilon_{J_l\cdot}^T\|_F \le \left(1 + \frac{2}{\sqrt{c_1}}\right)\|\Upsilon_{\widetilde{S}\cdot}^T\|_F,$$

we have

$$\|\widehat{\mathbf{\Sigma}}^{\frac{1}{2}}\Upsilon^T\|_F \gtrsim \|\Upsilon^T\|_F.$$

This is equivalent to

$$\sum_{k=2}^{K}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^T\widehat{\mathbf{\Sigma}}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) = \mathrm{tr}(\Upsilon\widehat{\mathbf{\Sigma}}\Upsilon^T) \gtrsim \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_F^2.$$

□

B.6 Proof of Lemmas A.7

Lemma B.5. *If* $\mathbf{u} \in \Gamma(s)$ *and* $\|\mathbf{u}\|_2 = 1$*, then*

$$\|\mathbf{u}\|_1 \lesssim \sqrt{s}.$$

Proof of Lemma B.5. Since $\mathbf{u} \in \Gamma(s)$, we have

$$\|\mathbf{u}\|_1 = \|\mathbf{u}_S\|_1 + \|\mathbf{u}_{S^C}\|_1 \le 3\|\mathbf{u}_S\|_1 \le 3\sqrt{s}\|\mathbf{u}_S\|_2 \lesssim \sqrt{s}.$$

□

Proof of Lemma A.7. By Lemma A.2, with high probability, we have that

$$\|\widehat{\mathbf{v}}_k - \mathbf{v}_k\|_\infty \lesssim \sqrt{\frac{\log p}{n}}.$$

Therefore, by applying Lemma B.5, we have

$$\begin{aligned}\|\widehat{\mathbf{v}}_k - \mathbf{v}_k\|_{2,s} &\le \|\widehat{\mathbf{v}}_k - \mathbf{v}_k\|_\infty \cdot \sup_{\|\mathbf{u}\|_2=1,\mathbf{u}\in\Gamma(s)}\|\mathbf{u}\|_1 \\ &\lesssim \sqrt{s}\|\widehat{\mathbf{v}}_k - \mathbf{v}_k\|_\infty \lesssim \sqrt{\frac{s\log p}{n}}.\end{aligned}$$

□

C Discussion about the Sparsity Assumption

Molstad & Rothman (2019) considered a different type of sparsity assumption that the mean difference and the precision matrices are sparse. We are going to compare this assumption with our assumption that the discriminant coefficient tensor is sparse. For ease of presentation, we focus on $K = 2$, but our discussion easily extends to problems with any number of classes.

We first define the sparsity of a vector and a matrix in the following way. If a vector $\mathbf{a}$ has at most k nonzero entries, we call that $\mathbf{a}$ is k-sparse. For a matrix $\mathbf{A}$, we say $\mathbf{A}$ is k-sparse if its columns and rows have at most k nonzero entries. Denote $\boldsymbol{\Omega}_m = \boldsymbol{\Sigma}_m^{-1}$ as the inverse of the covariance matrix. Denote $\boldsymbol{\zeta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ as the mean difference in the tensor form.

Lemma C.1. *Assume that* $\mathrm{vec}(\boldsymbol{\zeta})$ *is* k_0*-sparse and* $\boldsymbol{\Omega}_m$ *is* k_m*-sparse for* $m = 1, \ldots, M$*. We have that* $\mathbf{B}$ *has at most* $k_0 k_1 \cdots k_M$ *nonzero elements.*

By Lemma C.1, we have that when $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and $\boldsymbol{\Omega}_m, m = 1, \ldots, M$ are sparse (i.e, k_0 is much smaller than $\prod_{m=1}^{M} p_m$, and k_l is much smaller than p_l for $l = 1, \ldots, M$), we typically have sparse $\mathbf{B}$. Therefore, the sparsity assumption in Molstad & Rothman (2019) that $\boldsymbol{\Omega}_m, \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ are all sparse actually implies that $\mathbf{B}$ is sparse, which is our sparsity assumption. On the contrary, we do not need all of $\boldsymbol{\Omega}_m, \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ to be sparse to have sparse $\mathbf{B}$. Thus, our sparsity is indeed weaker. We will illustrate this point in the following example.

Example 1. Consider the TDA model with $K = 2$, $M = 2$. Let $\boldsymbol{\zeta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \in \mathbb{R}^{p_1 \times p_1}$. Consider that

$$\boldsymbol{\zeta} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

and

$$\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0.2 & 0 & 0 & 0 & \cdots & 0 \\ 0.2 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0.2 & 0.2 & \cdots & 0.2 \\ 0 & 0 & 0.2 & 1 & 0.2 & \cdots & 0.2 \\ 0 & 0 & 0.2 & 0.2 & 1 & \cdots & 0.2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0.2 & 0.2 & 0.2 & \cdots & 1 \end{pmatrix}.$$

In other words, there is a 2×2 block in $\boldsymbol{\zeta}$ that is nonzero, and $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are block diagonal matrices with one small block and a large block (when p_m is large), respectively. The large blocks make the precision matrices non-sparse, but $\mathbf{B}$ is still sparse as follows

$$\mathbf{B} = \begin{pmatrix} 1.44 & 1.44 & 0 & \cdots & 0 \\ 1.44 & 1.44 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

C.1 Proof of Lemma C.1

Proof of Lemma C.1. We first consider $M = 2$. Denote $\boldsymbol{\zeta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. Then $\boldsymbol{\zeta}$ is of dimension $p_1 \times p_2$ and it has at most k_0 nonzero elements in total. Without loss of generality, assume that the k_0 nonzero elements are located in an upper left block of size $a \times b$. For $1 \le i \le b$, the ith column in the block has at most d_i nonzero elements where $0 < d_i \le a$, $\sum_i d_i = k_0$. Therefore, $\boldsymbol{\zeta}_i$ is d_i-sparse for $1 \le i \le b$.

By assumption, $\boldsymbol{\Omega}_m$ is k_m-sparse for $m = 1, 2$. Then $\boldsymbol{\Omega}_1\boldsymbol{\zeta}$ has at most b nonzero columns and the ith column is at most $k_1 d_i$-sparse when $k_1 d_i \le p_1, i \le b$. Now for $\boldsymbol{\Omega}_1\boldsymbol{\zeta}\boldsymbol{\Omega}_2$, to measure the sparsity, we can take a look at $\boldsymbol{\Omega}_2(\boldsymbol{\Omega}_1\boldsymbol{\zeta})^T$. Let $\mathbf{r}_i$ be the ith column of $(\boldsymbol{\Omega}_1\boldsymbol{\zeta})^T$. Then $\boldsymbol{\Omega}_2\mathbf{r}_i$ has at most $k_2\|\mathbf{r}_i\|_0$ nonzero elements. Therefore, $\boldsymbol{\Omega}_1\boldsymbol{\zeta}\boldsymbol{\Omega}_2$ has at most $\sum_i k_2\|\mathbf{r}_i\|_0 = k_2\|\boldsymbol{\Omega}_1\boldsymbol{\zeta}\|_0 = k_2 k_1 \sum_{i=1}^b d_i = k_0 k_1 k_2$ nonzero elements. Note that when $k_2 b > p_2$ or $k_1 d_i > p_1$ for some i, $\boldsymbol{\Omega}_1\boldsymbol{\zeta}\boldsymbol{\Omega}_2$ will become sparser and has less than $k_0 k_1 k_2$ nonzero elements.

Consider $\mathbf{B} = [\![\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_M^{-1}]\!]$ for a general M. By matricization, we have $\mathbf{B}_{(1)} = \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(1)}(\boldsymbol{\Sigma}_M^{-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_2^{-1})^T = \boldsymbol{\Omega}_1(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(1)}(\boldsymbol{\Omega}_M \otimes \cdots \otimes \boldsymbol{\Omega}_2)^T$. Since $\boldsymbol{\Omega}_m$ is k_m-sparse, by the definition of Kronecker product, it is easy to see that $\boldsymbol{\Omega}_M \otimes \cdots \otimes \boldsymbol{\Omega}_2$ is at most $k_2 k_3 \cdots k_M$-sparse. Note that $\boldsymbol{\Omega}_1$ is k_1-sparse and $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{(1)}$ has at most k_0 nonzero elements. Using the previous result for $M = 2$, we have that $\mathbf{B}$ has at most $k_0 k_1 \cdots k_M$ nonzero elements.

□

References

Cai, T. & Zhang, L. (2019), 'High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**(4), 675–705.

Gao, C., Ma, Z. & Zhou, H. H. (2017), 'Sparse cca: Adaptive estimation and computational barriers', *The Annals of Statistics* **45**(5), 2074–2101.

Kollo, T. (2005), *Advanced multivariate statistics with matrices*, Springer.

Lyu, X., Sun, W. W., Wang, Z., Liu, H., Yang, J. & Cheng, G. (2020), 'Tensor graphical model: Nonconvex optimization and statistical inference', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**(8), 2024–2037.

Molstad, A. J. & Rothman, A. J. (2019), 'A penalized likelihood method for classification with matrix-valued predictors', *Journal of Computational and Graphical Statistics* **28**(1), 11–22.

Pan, Y., Mai, Q. & Zhang, X. (2019), 'Covariate-adjusted tensor classification in high dimensions', *Journal of the American Statistical Association* **114**(527), 1305–1319.

Tsybakov, A. B. (2009), *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer New York.

Vershynin, R. (2010), 'Introduction to the non-asymptotic analysis of random matrices', *arXiv preprint arXiv:1011.3027* .