# High-dimensional Linear Discriminant Analysis: Optimality, Adaptive Algorithm, and Missing Data

T. Tony Cai

*University of Pennsylvania, Philadelphia, USA*

Linjun Zhang

*University of Pennsylvania, Philadelphia, USA*

**Summary.** This paper aims to develop an optimality theory for linear discriminant analysis in the high-dimensional setting. A data-driven and tuning-free classification rule, which is based on an adaptive constrained $\ell_1$ minimization approach, is proposed and analyzed. Minimax lower bounds are obtained and this classification rule is shown to be simultaneously rate optimal over a collection of parameter spaces. In addition, we consider classification with incomplete data under the missing completely at random (MCR) model. An adaptive classifier with theoretical guarantees is introduced and optimal rate of convergence for high-dimensional linear discriminant analysis under the MCR model is established. The technical analysis for the case of missing data is much more challenging than that for the complete data. We establish a large deviation result for the generalized sample covariance matrix, which serves as a key technical tool and can be of independent interest. An application to lung cancer and leukemia studies is also discussed.

*Keywords:* Adaptive classification rule; constrained $\ell_1$ minimization; high-dimensional data; linear discriminant analysis; missing data; optimal rate of convergence.

## 1. Introduction

Classification is one of the most important tasks in statistics and machine learning with applications in a broad range of fields. See, for example, Hastie et al. [2009]. The problem has been well studied in the low-dimensional setting. In particular, consider the Gaussian case where one wishes to classify a new random vector $\boldsymbol{Z}$ drawn with equal probability from one of two Gaussian distributions $N_p(\boldsymbol{\mu}_1, \Sigma)$ (class 1) and $N_p(\boldsymbol{\mu}_2, \Sigma)$ (class 2). In the ideal setting where all the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$ are known, Fisher's linear discriminant rule, which is given by

$$C_{\boldsymbol{\theta}}(\boldsymbol{Z}) = \begin{cases} 1, & (\boldsymbol{Z} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})^\top \Omega \boldsymbol{\delta} < 0 \\ 2, & (\boldsymbol{Z} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})^\top \Omega \boldsymbol{\delta} \geq 0, \end{cases} \tag{1}$$

where $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, and $\Omega = \Sigma^{-1}$ is the precision matrix, is well known to be optimal [Anderson, 2003]. Fisher's rule separates the two classes by a linear combination of features and its misclassification error is given by $R_{\text{opt}}(\boldsymbol{\theta}) = \Phi\left(-\frac{1}{2}\Delta\right)$, where $\Phi$ is the cumulative distribution function of the standard normal distribution and $\Delta = \sqrt{\boldsymbol{\delta}^\top \Omega \boldsymbol{\delta}}$ is the signal-to-noise ratio.

Although Fisher's rule can serve as a useful performance benchmark, it is not practical for real data analysis as the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\Sigma$ are typically unknown and need to be estimated from the data. In applications, it is desirable to construct a data-driven classification rule based on two observed random samples, $\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_{n_1}^{(1)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\boldsymbol{X}_1^{(2)}, ..., \boldsymbol{X}_{n_2}^{(2)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma)$. In the conventional low-dimensional setting, this is easily achieved by plugging in

Fisher's linear discriminant rule (1) the corresponding sample means and pooled sample covariance matrix for the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\Sigma$ respectively. This classification rule is asymptotically optimal when the dimension $p$ is fixed. See, for example, Anderson [2003].

Driven by many contemporary applications, much recent attention has been on the high-dimensional setting where the dimension is much larger than the sample size. In this case, the sample covariance matrix is not even invertible and it is difficult to estimate the precision matrix $\Omega$. The standard linear discriminant rule thus fails completely. Several regularized classification methods, including the regularized LDA [Wu et al., 2009], covariance-regularized classification [Witten and Tibshirani, 2009], and hard thresholding [Shao et al., 2011], have been proposed for classification of high-dimensional data. However, all these methods rely on the individual sparsity assumptions on $\Omega$ (or $\Sigma$) and $\boldsymbol{\delta}$. A fundamental quantity in LDA is the discriminant direction $\boldsymbol{\beta} = \Omega\boldsymbol{\delta}$ and a more flexible assumption is the sparsity of $\boldsymbol{\beta}$. In particular, Cai and Liu [2011], Mai et al. [2012] introduced a direct estimation method for the high-dimensional LDA based on the key observation that the ideal Fisher's discriminant rule given in (1) depends on the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\Sigma$ primarily through $\boldsymbol{\beta} = \Omega\boldsymbol{\delta}$. They proposed to estimate the discriminant direction $\boldsymbol{\beta}$ directly instead of estimating $\Sigma$ and $\boldsymbol{\delta}$ separately, under the assumption that $\boldsymbol{\beta}$ is sparse. The proposed classification rule was shown to be consistent.

Despite much recent progress in methodological development on high-dimensional classification problems, there has been relatively little fundamental study on the optimality theory for the discriminant analysis. Minimax study of high-dimensional discriminant analysis has been considered in Azizyan et al. [2013] and Li et al. [2017] in the special case where the covariance matrix $\Sigma = \sigma^2 I$ for some $\sigma > 0$. However, even in this relatively simple setting there is still a gap between the minimax upper and lower bounds. It is unclear what the optimal rate of convergence for the minimax misclassification risk is and which classification rule is rate optimal under the general Gaussian distribution. The first major goal of the present paper is to provide answers to these questions. Furthermore, although the problem of missing data arises frequently in the analysis of high-dimensional data, compared to the conventional low-dimensional setting, there is a paucity of methods for inference with incomplete high-dimensional data. The second goal of this paper is to develop an optimality theory for high-dimensional discriminant analysis with incomplete data and to construct in this setting a data-driven adaptive classifier with theoretical guarantees.

Given two random samples, $\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_{n_1}^{(1)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\boldsymbol{X}_1^{(2)}, ..., \boldsymbol{X}_{n_2}^{(2)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma)$, we wish to construct a classifier $\hat{C}$ to classify a future data point $\boldsymbol{Z}$ drawn from these two distributions with equal prior probabilities, into one of the two classes. Given the observed data, the performance of the classification rule is measured by the misclassification error

$$R_{\boldsymbol{\theta}}(\hat{C}) = \mathbb{P}_{\boldsymbol{\theta}}(\mathrm{label}(\boldsymbol{Z}) \neq \hat{C}(\boldsymbol{Z})), \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$, $\mathbb{P}_{\boldsymbol{\theta}}$ denotes the probability with respect to $\boldsymbol{Z} \sim \frac{1}{2}N_p(\boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}N_p(\boldsymbol{\mu}_2, \Sigma)$ and $\boldsymbol{Z}$ is independent of the observed $\boldsymbol{X}$'s. $\mathrm{label}(\boldsymbol{Z})$ denotes the true class of $\boldsymbol{Z}$. For a given classifier $\hat{C}$, we use the excess misclassification risk relative to the oracle rule (1), $R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta})$, to measure the performance of the classifier $\hat{C}$. Let $n = \min\{n_1, n_2\}$. We consider in this paper a collection of the parameter spaces $\mathcal{G}(s, M_{n,p})$ defined by

$$\begin{aligned} \mathcal{G}(s, M_{n,p}) = \{\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}, \Sigma \succ 0, \\ \|\boldsymbol{\beta}\|_0 \leq s, M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M, M_{n,p} \leq \Delta \leq 3M_{n,p}\}, \end{aligned} \tag{3}$$

where $M > 1$ is a constant, $M_{n,p} > 0$ can potentially grow with $n$ and $p$, and $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ are respectively the largest and smallest eigenvalue of $\Sigma$. The notation $\Sigma \succ 0$ means

that $\Sigma$ is symmetric and positive definite. Recall that $\Delta = \sqrt{\boldsymbol{\delta}^\top \Omega \boldsymbol{\delta}}$ and $\boldsymbol{\beta} = \Omega \boldsymbol{\delta}$. The sparsity constraint $\|\boldsymbol{\beta}\|_0 \le s$, according to the oracle rule (1), implies the belief that only a limited number of covariates have discriminating power and contribute to the classification task. In addition, our lower bound results in Theorem 7 show that this sparsity assumption is necessary in general without further constraints of parameter space. Furthermore, we also assume the eigenvalues of the covariance matrix $\Sigma$ are bounded from below and above. This assumption is commonly used in high-dimensional statistics, ranging from high-dimensional linear regression [Javanmard and Montanari, 2014], covariance matrix estimation [Cai and Yuan, 2012], classification [Cai and Liu, 2011] and clustering [Cai et al., 2018].

Combining the upper and lower bounds results given in Section 3 leads to the following minimax rates of convergence for the excess misclassification risk.

THEOREM 1. *Consider the parameter space $\mathcal{G}(s, M_{n,p})$, $s$ and $p$ approach infinity as $n$ grows to infinity, and $M_{n,p} = o\left(\sqrt{\frac{n}{s \log p}}\right)$ with $n \to \infty$,*

(a) *If $M_{n,p}$ is a fixed constant not depending on $n$ and $p$, then for any constant $\alpha \in (0, 1)$, we have*

$$\inf \left\{ r : \inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}\left( R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \ge r \right) \le 1 - \alpha \right\} \asymp \frac{s \log p}{n}.$$

(b) *If $M_{n,p} \to \infty$ as $n \to \infty$, then for sufficiently large $n$ and any constant $\alpha \in (0, 1)$,*

$$\inf \left\{ r : \inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}\left( R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \ge r \right) \le 1 - \alpha \right\} \asymp \frac{s \log p}{n} \cdot e^{-(\frac{1}{8} + o(1)) M_{n,p}^2}.$$

It is worth noting that $M_{n,p}$ represents the magnitude of $\Delta$, which is interpreted as the signal-to-noise ratio. As shown in the second case, when the signal-to-noise ratio grows, the classification problem becomes easier and our result precisely characterizes that the convergence rate is exponentially faster with an additional factor $\exp\left(-\left(1/8 + o(1)\right) M_{n,p}^2\right)$.

Furthermore, we propose a three-step data-driven classification rule, called *AdaLDA*, by using an adaptive constrained $\ell_1$ minimization approach which takes into account the variability of individual entries. This classification rule is shown to be simultaneously rate optimal over the collection of parameter spaces $\mathcal{G}(s, M_{n,p})$. To the best of our knowledge, this is the first optimality result for classification of high-dimensional Gaussian data. Furthermore, in contrast to many classification rules proposed in the literature, which require to choose tuning parameters, this procedure is data-driven and tuning-free.

In addition, we also consider classification in the presence of missing data. As in the conventional low-dimensional setting, the problem of missing data also arises frequently in the analysis of high-dimensional data from in a range of fields such as genomics, epidemiology, engineering, and social sciences [Libbrecht and Noble, 2015, White et al., 2011, Graham, 2009]. Compared to the low-dimensional setting, there are relatively few inferential methods for missing data in the high-dimensional setting. Examples include high-dimensional linear regression [Loh and Wainwright, 2012], sparse principal component analysis [Lounici, 2013], covariance matrix estimation [Cai and Zhang, 2016], and vector autoregressive (VAR) processes [Rao et al., 2017]. In this paper, following the missing mechanism considered in the aforementioned papers, we investigate high-dimensional discriminant analysis in the presence of missing observations under the missing completely at random (MCR) model.

We construct a data-driven adaptive classifier with theoretical guarantees based on incomplete data and also develop an optimality theory for high-dimensional linear discriminant analysis

under the MCR model. The technical analysis for the case of missing data is much more challenging than that for the complete data, although the classification procedure and the resulting convergence rates look similar. To facilitate the theoretical analysis, we establish a key technical tool, which is a large deviation result for the generalized sample covariance matrix. This is related to the masked covariance matrix estimator considered in Levina and Vershynin [2012] and Chen et al. [2012], see further discussions in Section 2.3. This technical tool can be of independent interest as it is potentially useful for other related problems in high-dimensional statistical inference with missing data.

The proposed adaptive classification algorithms can be cast as linear programs and are thus easy to implement. Simulation studies are carried out to investigate the numerical performance of the classification rules. The results show that the proposed classifiers enjoy superior finite sample performance in comparison to existing methods for high-dimensional linear discriminant analysis. The proposed classifiers are also illustrated through an application to the analysis of lung cancer and leukemia datasets. The results show that they outperform existing methods.

The rest of the paper is organized as follows. In Section 2, after basic notation and definitions are reviewed, we introduce an adaptive algorithm for high-dimensional discriminant analysis with the complete data and then propose a more general procedure for the setting of incomplete data. Section 3 studies the theoretical properties of these classification rules and related estimators. In addition, minimax lower bounds are given. The upper and lower bounds together establish the optimal rates of convergence for the minimax misclassification risk. Numerical performance of the classification rules are investigated in Section 4 and an extension to the multiple-class LDA is discussed in Section 5. The proofs of the main results are given in Section 6. Technical lemmas are proved in the Supplementary Material [Cai and Zhang, 2018].

## 2.  Methodology

In this section, we firstly introduce an adaptive algorithm for high-dimensional linear discriminant analysis with the complete data. This algorithm is called AdaLDA (**Ada**ptive **L**inear **D**iscriminant **A**nalysis rule). We then propose a data-driven classifier, called ADAM (**A**daptive linear **D**iscriminant **A**nalysis with randomly **M**issing data), for the incomplete data under the MCR model.

### 2.1.  Notation and definitions

We begin with basic notation and definitions. Throughout the paper, vectors are denoted by boldface letters. For a vector $\boldsymbol{x} \in \mathbb{R}^p$, the usual vector $\ell_0, \ell_1, \ell_2$ and $\ell_\infty$ norms are denoted respectively by $\|\boldsymbol{x}\|_0, \|\boldsymbol{x}\|_1, \|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$. Here the $\ell_0$ norm counts the number of nonzero entries in a vector. The support of a vector $\boldsymbol{x}$ is denoted by $\mathrm{supp}(\boldsymbol{x})$. The symbol $\circ$ denotes the Hadamard product. For $p \in \mathbb{N}$, $[p]$ denotes the set $\{1, 2, ..., p\}$. For $j \in [p]$, denote by $\boldsymbol{e}_j$ the $j$-th canonical basis in $\mathbb{R}^p$. For a matrix $\Sigma = (\sigma_{ij})_{1 \le i,j \le p} \in \mathbb{R}^{p \times p}$, the Frobenius norm is defined as $\|\Sigma\|_F = \sqrt{\sum_{i,j} \sigma_{ij}^2}$ and the spectral norm is defined to be $\|\Sigma\|_2 = \sup_{\|\boldsymbol{x}\|_2=1} \|\Sigma \boldsymbol{x}\|_2$. The vector $\ell_\infty$ norm of the matrix $\Sigma$ is $|\Sigma|_\infty = \max_{i,j} |\sigma_{ij}|$. For a symmetric matrix $\Sigma$, we use $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ to denote respectively the largest and smallest eigenvalue of $\Sigma$. $\Sigma \succ 0$ means that $\Sigma$ is positive definite. For a positive integer $s < p$, let $\Gamma(s) = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}_{S^C}\|_1 \le \|\boldsymbol{u}_S\|_1$, for some $S \subset [p]$ with $|S| = s\}$, where $\boldsymbol{u}_S$ denotes the subvector of $\boldsymbol{u}$ confined to $S$. For two sequences of positive numbers $a_n$ and $b_n$, $a_n \lesssim b_n$ means that for some constant $c > 0$, $a_n \le c b_n$ for all $n$, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We say an event $\mathcal{A}_n$ holds with high probability if $\liminf_{n \to \infty} \mathbb{P}(\mathcal{A}_n) = 1$. Finally, $c_0, c_1, c_2, C, C_1, C_2, \dots$ denote generic positive constants

that may vary from place to place.

The complete data $\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_{n_1}^{(1)}$ and $\boldsymbol{X}_1^{(2)}, ..., \boldsymbol{X}_{n_2}^{(2)}$ are independent realizations of $\boldsymbol{X}^{(1)} \sim N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\boldsymbol{X}^{(2)} \sim N_p(\boldsymbol{\mu}_2, \Sigma)$. We assume $n_1 \asymp n_2$ and define $n = \min\{n_1, n_2\}$. In our asymptotic framework, we let $n$ be the driving asymptotic parameter, $s$ and $p$ approach infinity as $n$ grows to infinity. The missing completely at random (MCR) model assumes that one observes samples $\{\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_{n_1}^{(1)}\}$ and $\{\boldsymbol{X}_1^{(2)}, ..., \boldsymbol{X}_{n_2}^{(2)}\}$ with missing values, where the observed coordinates of $\boldsymbol{X}_t^{(k)}$ are indicated by an independent vector $\boldsymbol{S}_t^{(k)} \in \{0, 1\}^p$ for $t = 1, ..., n_k$, $k = 1, 2$, that is,

$$X_{tj}^{(k)} \text{ is observed if } S_{tj}^{(k)} = 1 \text{ and } X_{tj}^{(k)} \text{ is missing if } S_{tj}^{(k)} = 0; \ t \in [n_k], j \in [p], k = 1, 2. \quad (4)$$

Here $X_{tj}^{(k)}$ and $S_{tj}^{(k)}$ are respectively the $j$-th coordinate of the vectors $\boldsymbol{X}_t^{(k)}$ and $\boldsymbol{S}_t^{(k)}$. Generally, we use the superscript "$*$" to denote objects related to missing values. The incomplete samples with missing values are denoted by $\boldsymbol{X}^{(1)*} = \{\boldsymbol{X}_1^{(1)*}, ..., \boldsymbol{X}_{n_1}^{(1)*}\}$ and $\boldsymbol{X}^{(2)*} = \{\boldsymbol{X}_1^{(2)*}, ..., \boldsymbol{X}_{n_2}^{(2)*}\}$.

Regarding the mechanism for missingness, the MCR model is formally stated as below. This assumption is more general than the one considered previously by Loh and Wainwright [2012] and Lounici [2013].

ASSUMPTION 1. *(Missing Completely at Random (MCR))* $S = \left\{ \boldsymbol{S}_t^{(k)} \in \{0, 1\}^p : t = 1, ..., n_k, \ k = 1, 2 \right\}$ *is independent of the values of* $\boldsymbol{X}_t^{(1)}$ *and* $\boldsymbol{X}_t^{(2)}$ *for* $t = 1, ..., n_k$, $k = 1, 2$. *Here* $\boldsymbol{S}_t^{(k)}$ *can be either deterministic or random, but independent of* $\boldsymbol{X}_t^{(1)}$ *and* $\boldsymbol{X}_t^{(2)}$.

A major goal of the present paper is to construct a classification rule $\hat{C}$ in the high dimensional setting where $p \gg n$ for both complete and incomplete data.

## 2.2. Data-driven adaptive classifier for complete data

We first consider the case of complete data. In this setting, as mentioned in the introduction, a number of high-dimensional linear discriminant rules have been proposed in the literature. In particular, Cai and Liu [2011] introduced a classification rule called LPD (Linear Programming Discriminant) rule by directly estimating the discriminant direction $\boldsymbol{\beta}$ through solving the following optimization problem:

$$\hat{\boldsymbol{\beta}}_{\text{LPD}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{\beta}\|_1 : \text{subject to } \|\hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)\|_\infty \leq \lambda_n \right\}, \quad (5)$$

where $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\Sigma}$ are sample means and pooled sample covariance matrix respectively, and $\lambda_n = C\sqrt{\log p/n}$ is the tuning parameter with some constant $C$. Based on $\hat{\boldsymbol{\beta}}_{\text{LPD}}$, the LPD rule is then given by

$$\hat{C}_{\text{LPD}}(\boldsymbol{Z}) = \begin{cases} 1, & (\boldsymbol{Z} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2})^\top \hat{\boldsymbol{\beta}}_{\text{LPD}} < 0 \\ 2, & (\boldsymbol{Z} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2})^\top \hat{\boldsymbol{\beta}}_{\text{LPD}} \geq 0 \end{cases}. \quad (6)$$

The LPD rule is easy to implement and Cai and Liu [2011] proves the consistency of LPD when the tuning parameter $\lambda_n$ is appropriately chosen. However, it has three drawbacks. One major drawback of the LPD rule is that it uses a common constraint $\lambda_n$ for all coordinates of $\boldsymbol{a} = \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$. This essentially treats the random vector $\boldsymbol{a}$ as homoscedastic, while in fact $\boldsymbol{a}$ is intrinsically heteroscedastic and the coordinates could have a wide range of variability. The resulting estimator $\hat{\boldsymbol{\beta}}_{\text{LPD}}$ obtained in (5) of the discriminant direction $\boldsymbol{\beta}$ has yet to be

shown as rate optimal; secondly, the procedure is not adaptive in the sense that the tuning parameter $\lambda_n$ is not fully specified and needs to be chosen through an empirical method such as cross-validation. The third drawback is that the LPD rule $\hat{C}_{\text{LPD}}$ does not come with theoretical optimality guarantees.

To resolve these drawbacks, we introduce an adaptive algorithm for high-dimensional LDA with complete data, called AdaLDA (**Ada**ptive **L**inear **D**iscriminant **A**nalysis rule), which takes into account the heteroscedasticity of the random vector $\boldsymbol{a} = \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$. AdaLDA is fully data-driven and tuning-free and will be shown to be minimax rate optimal later. Before we describe the classifier in detail, it is helpful to state the following key technical result which provides the motivation for the new procedure.

LEMMA 1. *Suppose $\{\boldsymbol{X}_t^{(1)}\}_{t=1}^{n_1}$ and $\{\boldsymbol{X}_t^{(2)}\}_{t=1}^{n_2}$ are i.i.d. random samples from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$ respectively with $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$. Let $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\boldsymbol{\beta} = \Omega\boldsymbol{\delta}$, $\Delta = \sqrt{\boldsymbol{\beta}^\top \boldsymbol{\delta}}$ and $\boldsymbol{a} = \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$, where $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\Sigma}$ are sample means and pooled sample covariance matrix respectively. Then*

$$\text{Var}(a_j) = \frac{n-1}{2n^2}(\sigma_{jj}\Delta^2 + \delta_j^2) + \frac{2}{n}\sigma_{jj}, \quad j = 1, ..., p.$$

*Furthermore, with probability at least $1 - 4p^{-1}$,*

$$|a_j| \leq 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\sigma_{jj}} \cdot \left(\sqrt{\frac{25\Delta^2}{2} + 1}\right), \quad j = 1, ..., p. \tag{7}$$

A major step in the construction of the AdaLDA classifier is using Lemma 1 to construct an element-wise constraint for $\hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_2)$, which relies on an accurate estimation of the right hand side of (7). In (7), $\sigma_{jj}$ can be easily estimated by the sample variances $\hat{\sigma}_{jj}$, but $\Delta^2$ is harder to estimate. In the following, we begin by constructing a preliminary estimator $\tilde{\boldsymbol{\beta}}$, estimating $\Delta^2$ by $|\tilde{\boldsymbol{\beta}}^\top(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)|$, and then applying Lemma 1 to refine the estimation of $\boldsymbol{\beta}$. The data-driven adaptive classifier AdaLDA is constructed in three steps.

**Step 1 (Estimating $\Delta^2$).** Fix $\lambda_0 = 25/2$, we estimate $\boldsymbol{\beta}$ by a preliminary estimator

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1$$

$$\text{subject to } |\boldsymbol{e}_j^\top \left( \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \right)| \leq 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}} \cdot (\lambda_0 \boldsymbol{\beta}^\top(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) + 1), \ j \in [p]. \tag{8}$$

Then we estimate $\Delta^2$ by $\hat{\Delta}^2 = |\tilde{\boldsymbol{\beta}}^\top(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)|$.

**Step 2 (Adaptive estimation of $\boldsymbol{\beta}$).** Given $\hat{\Delta}^2$, the final estimator $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$ of $\boldsymbol{\beta}$ is constructed through the following linear optimization

$$\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1$$

$$\text{subject to } |\boldsymbol{e}_j^\top \left( \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \right)| \leq 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}(\lambda_0 \hat{\Delta}^2 + 1)}, \ j \in [p]. \tag{9}$$

**Step 3 (Construction of AdaLDA).** The AdaLDA classification rule is obtained by plugging $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$ into Fisher's rule (1),

$$\hat{C}_{\text{AdaLDA}}(\boldsymbol{Z}) = \begin{cases} 1, & \left(\boldsymbol{Z} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)^\top \hat{\boldsymbol{\beta}}_{\text{AdaLDA}} < 0, \\ 2, & \left(\boldsymbol{Z} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)^\top \hat{\boldsymbol{\beta}}_{\text{AdaLDA}} \geq 0. \end{cases} \tag{10}$$

Note that there is a square root on $\Delta^2$ (or $\hat{\Delta}^2$) in both (7) and (9), but this square root is removed in (8). Intuitively, by removing the square root in (8), Step 1 becomes a linear program, which provides a computationally efficient but sub-optimal estimator. This estimator is then refined to be rate-optimal in Step 2 by adding back the square root. This two-step idea is in the similar spirit as that in Cai et al. [2016] for adaptive estimation of precision matrices. Despite this similarity in the ideas for the construction procedures, the problem considered and the technical tools applied in the present paper are very different from those in Cai et al. [2016].

This classification rule does not require a tuning parameter and the estimator $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$ is solved by optimizing a linear program with an element-wise constraint, adapting to individual variability of $\hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$. It will be shown in Section 3 that the AdaLDA classification rule is adaptively minimax rate optimal. Our theoretical analysis also shows that the resulting estimator $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$ is rate optimally adaptive whenever $\lambda_0$ is a sufficiently large constant. In particular, it can be taken as fixed at $\lambda_0 = 25/2$, which is derived from the concentration inequality given in Lemma 1.

REMARK 2.1. Note that the optimization problems (8) and (9) are both linear programs, so the proposed AdaLDA rule is computationally easy to implement. In contrast, the LPD uses a universal tuning parameter $\lambda_n = C\sqrt{\log p/n}$, whose value is usually chosen by cross-validation. This tuning procedure is computationally costly. In addition, cross-validation tends to overfit [Friedman et al., 2001]. Therefore, estimator obtained through cross-validation can be variable and its theoretical properties are unclear, while the AdaLDA procedure does not depend on any unknown parameter and the estimator will be shown to be minimax rate optimal.

### 2.3. ADAM with randomly missing data

We now turn to the case of incomplete data under the MCR model. To generalize AdaLDA to the incomplete data case, we proceed by firstly estimating $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\Sigma$. The following estimators follow the idea in Cai and Zhang [2016], and for completeness, we present their proposed estimators below. Let

$$n_{ij}^{(k)*} = \sum_{t=1}^{n_k} S_{ti}^{(k)} S_{tj}^{(k)}, \quad 1 \le i,j \le p, \; k=1,2.$$

Here $n_{ij}^{(k)*}$ is the number of vectors $\boldsymbol{X}_t^{(k)}$ in which the $i^{th}$ and $j^{th}$ entries are both observed. In addition, we denote $n_i^{(k)*} = n_{ii}^{(k)*}$ for simplicity and

$$n_{\min}^* = \min_{i,j,k} n_{ij}^{(k)*}. \tag{11}$$

In the presence of missing values, the usual sample mean and sample covariance matrix can no longer be calculated. Instead, the "generalized sample mean" is proposed, defined by

$$\hat{\boldsymbol{\mu}}_1 = (\hat{\mu}_{1i}^*)_{1 \le i \le p} \quad \text{with} \quad \hat{\mu}_{1i}^* = \frac{1}{n_i^{(1)*}} \sum_{t=1}^{n_1} X_{ti}^{(1)} S_{ti}^{(1)}, \quad 1 \le i \le p;$$

$$\hat{\boldsymbol{\mu}}_2 = (\hat{\mu}_{2i}^*)_{1 \le i \le p} \quad \text{with} \quad \hat{\mu}_{2i}^* = \frac{1}{n_i^{(2)*}} \sum_{t=1}^{n_2} X_{ti}^{(2)} S_{ti}^{(2)}, \quad 1 \le i \le p.$$

The "generalized sample covariance matrix" is then defined by $\hat{\Sigma} = (\hat{\sigma}_{ij}^*)_{1 \le i,j \le p}$ with

$$\hat{\sigma}_{ij}^* = \frac{1}{n_{ij}^{(1)*} + n_{ij}^{(2)*}} \left( \sum_{t=1}^{n_1} (X_{ti}^{(1)} - \hat{\mu}_{1i}^*)(X_{tj}^{(1)} - \hat{\mu}_{1j}^*) S_{ti}^{(1)} S_{tj}^{(1)} + \sum_{t=1}^{n_2} (X_{ti}^{(2)} - \hat{\mu}_{2i}^*)(X_{tj}^{(2)} - \hat{\mu}_{2j}^*) S_{ti}^{(2)} S_{tj}^{(2)} \right).$$

For these generalized estimators, we have the following bound under the MCR model.

LEMMA 2. *Let* $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\boldsymbol{\beta} = \Omega\boldsymbol{\delta}$, $\Delta = \sqrt{\boldsymbol{\delta}^\top \Omega \boldsymbol{\delta}}$ *and* $\boldsymbol{a}^* = \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$. *Then conditioning on* $\boldsymbol{S}$, *we have with high probability,*

$$|a_j^*| \le 4\sqrt{\frac{\log p}{n_{\min}^*}} \cdot \sqrt{\sigma_{jj}} \cdot \left(\sqrt{64\Delta^2 + 1}\right), \quad j = 1, ..., p. \tag{12}$$

REMARK 2.2. Although the above result has a form that is similar to Lemma 1, its derivation is quite different and relies on a new technical tool, the large deviation bound for $\hat{\Sigma}$. This is of independent interest and is related to that of the masked sample covariance estimator considered in Levina and Vershynin [2012] and Chen et al. [2012]. In particular, the masked sample covariance estimator considered in Chen et al. [2012] applies the mask matrix to the sample covariance maxtrix, while our proposed estimator $\hat{\Sigma}$ can be interpreted as applying the mask matrix to each *i.i.d.* sample, and thus is more general. The proof of Lemma 2 uses the idea of Lemma 2.1 in Cai and Zhang [2016], but yields a sharper bound. The detailed proof is given in Section A.3.2 in the supplement [Cai and Zhang, 2018].

We propose to estimate $\boldsymbol{\beta}$ adaptively and construct ADAM (**A**daptive linear **D**iscriminant **A**nalysis with randomly **M**issing data) in the following way:

**Step 1 (Estimating $\Delta^2$).** Fix $\lambda_1 = 64$. We estimate $\boldsymbol{\beta}$ by a preliminary estimator

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1$$

$$\text{subject to } |e_j^\top \left(\hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)\right)| \le 4\sqrt{\frac{\log p}{n_{min}^*}} \cdot \sqrt{\hat{\sigma}_{jj}^*} \cdot (\lambda_1 \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) + 1), \ j \in [p].$$
$$\tag{13}$$

Then we estimate $\Delta^2$ by $\hat{\Delta}^{*2} = |\tilde{\boldsymbol{\beta}}^\top (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)|$.

**Step 2 (Adaptive estimation of $\boldsymbol{\beta}$).** Given $\hat{\Delta}^{*2}$, the final estimator $\hat{\boldsymbol{\beta}}_{\text{ADAM}}$ of $\boldsymbol{\beta}$ is constructed by the following linear optimization problem

$$\hat{\boldsymbol{\beta}}_{\text{ADAM}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1$$

$$\text{subject to } |e_j^\top \left(\hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)\right)| \le 4\sqrt{\frac{\log p}{n_{min}^*}} \cdot \sqrt{\hat{\sigma}_{jj}^*(\lambda_1 \hat{\Delta}^{*2} + 1)}, \ j \in [p]. \tag{14}$$

**Step 3 (Construction of ADAM).** Given the estimator $\hat{\boldsymbol{\beta}}_{\text{ADAM}}$ of the discriminant direction $\boldsymbol{\beta}$, we then construct the following ADAM classification rule by plugging $\hat{\boldsymbol{\beta}}_{\text{ADAM}}$ into the oracle rule (1):

$$\hat{C}_{\text{ADAM}}(\boldsymbol{Z}) = \begin{cases} 1, & \left(\boldsymbol{Z} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)^\top \hat{\boldsymbol{\beta}}_{\text{ADAM}} < 0, \\ 2, & \left(\boldsymbol{Z} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}\right)^\top \hat{\boldsymbol{\beta}}_{\text{ADAM}} \ge 0. \end{cases} \tag{15}$$

As shown in Section 3, $\hat{C}_{\text{ADAM}}$ has the similar theoretical performance as $\hat{C}_{\text{AdaLDA}}$.

REMARK 2.3. The ADAM algorithm is designed for the MCR model. Extensions to other missing mechanism such as missing not at random (MNAR) is possible but challenging. In such

a setting even parametric models are often not identifiable [Miao et al., 2016, Robins and Ritov, 1997]. Several authors have studied the problem of identification under MNAR with different conditions [Rotnitzky and Robins, 1997, Sun et al., 2016, Tchetgen Tchetgen and Wirth, 2017]. The consistency of our algorithm only relies on consistent estimation of the mean vectors and the covariance matrix. Therefore, if the means and the covariance matrix can be estimated consistently under some MNAR model, for example, by using EM algorithm and imputing the missing values [Schneider, 2001], we can then construct a consistent classification rule based on these estimators. However, such imputation techniques are computationally intensive [Lounici, 2014].

## 3. Theoretical properties of AdaLDA and ADAM

In this section, we develop an optimality theory for high-dimensional linear discriminant analysis for both the complete data and the incomplete data settings. We first investigate the theoretical properties of the AdaLDA and ADAM algorithms proposed in Section 2 and obtain the upper bounds for the excess misclassification risk. We then establish the lower bounds for the rate of convergence. The upper and lower bounds together yield the minimax rates of convergence and show that AdaLDA and ADAM are adaptively rate optimal.

### 3.1. Theoretical Analysis of AdaLDA

We begin by considering the properties of the estimator $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$ of the discriminant direction $\boldsymbol{\beta}$. The following theorem shows that $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$ attains the convergence rate of $M_{n,p}\sqrt{s\log p/n}$ over the class of sparse discriminating directions $\mathcal{G}(s, M_{n,p})$ defined in (3). The matching lower bound given in Section 3.3 implies that this rate is optimal. Therefore, AdaLDA adapts to both the sparsity pattern of $\boldsymbol{\beta}$ as well as the signal-to-noise ratio $\Delta$.

THEOREM 2. *Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$. Suppose $\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_{n_1}^{(1)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_1, \Sigma)$, $\boldsymbol{X}_1^{(2)}, ..., \boldsymbol{X}_{n_2}^{(2)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_2, \Sigma)$ and $n_1 \asymp n_2$. Assume that $M_{n,p}\sqrt{\frac{s\log p}{n}} = o(1)$. Then*

$$\sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2] \lesssim M_{n,p}\sqrt{\frac{s\log p}{n}}.$$

We then proceed to characterize the accuracy of the classification rule $\hat{C}_{\text{AdaLDA}}$, measured by the excess misclassification risk $R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta})$. Note that the conditional misclassification rate of $\hat{C}_{\text{AdaLDA}}$ given the two samples can be analytically calculated as

$$R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) = \frac{1}{2}\Phi\left(-\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\beta}}_{\text{AdaLDA}}}{\sqrt{\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}^\top \Sigma \hat{\boldsymbol{\beta}}_{\text{AdaLDA}}}}\right) + \frac{1}{2}\bar{\Phi}\left(-\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\beta}}_{\text{AdaLDA}}}{\sqrt{\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}^\top \Sigma \hat{\boldsymbol{\beta}}_{\text{AdaLDA}}}}\right),$$

where $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$ and $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$.

We are interested in the excess misclassification risk $R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta})$. That is, we compare $\hat{C}_{\text{AdaLDA}}$ with the oracle Fisher's rule, whose risk is given by

$$R_{\text{opt}}(\boldsymbol{\theta}) \overset{def}{=} R_{\boldsymbol{\theta}}(C_{\boldsymbol{\theta}}) = \Phi\left(-\frac{1}{2}\Delta\right).$$

The following theorem provides an upper bound for the excess misclassification risk of the AdaLDA rule.

THEOREM 3. *Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and assume the conditions in Theorem 2 hold.*

(a) *If $M_{n,p} \leq C_b$ for some $C_b > 0$, then there exists some constant $C > 0$,*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}\left( R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

(b) *If $M_{n,p} \to \infty$ as $n \to \infty$, then there exist some constant $C > 0$ and $\delta_n = o(1)$, such that*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}\left( R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot e^{-(\frac{1}{8} + \delta_n)M_{n,p}^2} \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

REMARK 3.1. The results in Theorem 3 improve the convergence rate of the misclassification risk of the LPD rule given in Cai and Liu [2011]. Consider the first case where $M_{n,p}$ is a constant not depending on $n$ and $p$, Theorem 3 of Cai and Liu [2011] shows that the convergence rate is $R_{\boldsymbol{\theta}}(\hat{C}_{\text{LPD}}) - R_{\text{opt}}(\boldsymbol{\theta}) = O_P((s \log p/n)^{1/2})$, while Theorem 3 here shows a faster rate $O_P((s \log p/n))$ when $M_{n,p}$ is a constant. Indeed, this improvement is due to a careful analysis of the misclassification error. In the proof of Theorem 3, it can be seen that the first order approximation error is vanishing, and only the second order approximation error, which has a faster convergence rate, remains. The lower bounds given in Section 3.3 show that both convergence rates in Theorem 3 are indeed optimal.

Similarly, upper bounds on the relative misclassification risk can be obtained.

PROPOSITION 3.2. *Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and assume the conditions in Theorem 2 hold.*

(a) *If $M_{n,p} \leq C_b$ for some $C_b > 0$, then there exists some constant $C > 0$,*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}\left( \frac{R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta})}{R_{\text{opt}}(\boldsymbol{\theta})} \leq C \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

(b) *If $M_{n,p} \to \infty$ as $n \to \infty$, then there exist some constant $C > 0$, such that*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}\left( \frac{R_{\boldsymbol{\theta}}(\hat{C}_{\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta})}{R_{\text{opt}}(\boldsymbol{\theta})} \leq C \cdot M_{n,p}^4 \cdot \frac{s \log p}{n} \right) \geq 1 - 8p^{-1}.$$

REMARK 3.3. The results in Proposition 3 show that the relative misclassification risk has a worse convergence rate when the magnitude of signal-to-noise ratio $M_{n,p}$ becomes larger. This is expected as when $M_{n,p}$ becomes larger, the classification problem itself becomes easier and the oracle misclassification risk is very small, making the oracle classification rule harder to be mimicked.

### 3.2. Theoretical Analysis of ADAM

We now investigate the theoretical properties of the ADAM procedure in the presence of missing data. Similar rates of convergence for estimation and excess misclassification risk can be obtained, but the technical analysis is much more involved under the MCR model.

Under the MCR model, suppose that the missingness pattern $S \in \{0,1\}^{n_1 \times p} \times \{0,1\}^{n_2 \times p}$ is a realization of a distribution $\mathcal{F}$. We consider the distribution space $\boldsymbol{\Psi}(n_0; n, p)$ given by

$$\boldsymbol{\Psi}(n_0; n, p) = \{\mathcal{F} : \mathbb{P}_{S \sim \mathcal{F}}(c_1 n_0 \leq n_{\min}^*(S) \leq c_2 n_0) \geq 1 - p^{-1}\},$$

for some constants $c_1, c_2 > 0$, and $n_{\min}^*(S)$ is defined for $S$ as in (11).

REMARK 3.4. This distribution space includes the missing uniformly and completely at random (MUCR) model considered in Loh and Wainwright [2012], Lounici [2013] and Lounici [2014]. More specifically, MUCR model assumes each entry $X_{i,j}^{(k)}$ ($k \in [2], i \in [n_k], j \in [p]$) is missing independently with probability $\epsilon$. As shown in Section A.6 in the supplement, when $\frac{1}{(1-\epsilon)^2}\sqrt{\frac{\log p}{n}} = o(1)$ as $n \to \infty$, the MUCR model is in the distribution space $\mathbf{\Psi}(n(1-\epsilon)^2; n, p)$.

In addition, this distribution space allows a more general variant of MUCR model that each entry $X_{i,j}^{(k)}$ is missing independently with different probabilities $\epsilon_{ij}^{(k)}$. If we assume $\tilde{c}_1 \cdot \epsilon \leq \min_{i,j,k} \epsilon_{ij}^{(k)} \leq \max_{i,j,k} \epsilon_{ij}^{(k)} \leq \tilde{c}_2 \cdot \epsilon$ for some constants $\tilde{c}_1, \tilde{c}_2 > 0$, then use the similar technique, this missingness pattern is included in $\mathbf{\Psi}(n(1-\epsilon)^2; n, p)$ when $\frac{1}{(1-\epsilon)^2}\sqrt{\frac{\log p}{n}} = o(1)$ as $n \to \infty$.

The following two theorems provide respectively the convergence rates for the discriminating direction estimator $\hat{\boldsymbol{\beta}}_{\text{ADAM}}$ and the excess misclassification rate of $\hat{C}_{\text{ADAM}}$ over the parameter space $\mathcal{G}(s, M_{n,p})$ for $\boldsymbol{\theta}$ and the distribution space $\mathbf{\Psi}(n_0; n, p)$.

THEOREM 4. *Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and the distribution space $\mathbf{\Psi}(n_0; n, p)$ with $M_{n,p}\sqrt{\frac{s \log p}{n_0}} = o(1)$. Suppose $\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_{n_1}^{(1)}$ and $\boldsymbol{X}_1^{(2)}, ..., \boldsymbol{X}_{n_2}^{(2)}$ are i.i.d. samples from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$ respectively. Assuming that $\boldsymbol{X}_1^{*(1)}, ..., \boldsymbol{X}_{n_1}^{*(1)}$ and $\boldsymbol{X}_1^{*(2)}, ..., \boldsymbol{X}_{n_2}^{*(2)}$ defined in (4) is observed and Assumption 1 with $S = \{\boldsymbol{S}_t^{(k)}\}_{t \in [n_k], k \in [2]}$ holds. Then the risk of estimating the discriminant direction $\boldsymbol{\beta}$ by ADAM satisfies*

$$\sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \mathbf{\Psi}(n_0; n, p)}} \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{ADAM}} - \boldsymbol{\beta}\|_2] \lesssim M_{n,p}\sqrt{\frac{s \log p}{n_0}}.$$

THEOREM 5. *Suppose the conditions of Theorem 4 hold.*

(a) *If $M_{n,p} \leq C_b$ for some $C_b > 0$, then there exists some constant $C > 0$, such that*

$$\inf_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \mathbf{\Psi}(n_0; n, p)}} \mathbb{P}\left(R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot \frac{s \log p}{n_0}\right) \geq 1 - 12p^{-1}.$$

(b) *If $M_{n,p} \to \infty$ as $n \to \infty$, then there exist some constant $C > 0$ and $\delta_n = o(1)$, such that*

$$\inf_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \mathbf{\Psi}(n_0; n, p)}} \mathbb{P}\left(R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot e^{-(\frac{1}{8}+\delta_n)M_{n,p}^2} \cdot \frac{s \log p}{n_0}\right) \geq 1 - 12p^{-1}.$$

In the complete data case, we have $n_0 = n$, so the rates of convergence shown in Theorem 4 and 5 match those in Theorems 2 and 3.

Similarly, upper bounds for the relative misclassification risks can be obtained.

PROPOSITION 3.5. *Suppose the conditions of Theorem 4 hold.*

(a) *If $M_{n,p} \leq C_b$ for some $C_b > 0$, then there exists some constant $C > 0$,*

$$\inf_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \mathbf{\Psi}(n_0; n, p)}} \mathbb{P}\left(\frac{R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta})}{R_{\text{opt}}(\boldsymbol{\theta})} \leq C \cdot \frac{s \log p}{n}\right) \geq 1 - 12p^{-1}.$$

(b) If $M_{n,p} \to \infty$ as $n \to \infty$, then there exist some constant $C > 0$, such that

$$\inf_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \boldsymbol{\Psi}(n_0; n, p)}} \mathbb{P}\left( \frac{R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta})}{R_{\text{opt}}(\boldsymbol{\theta})} \leq C \cdot M_{n,p}^4 \cdot \frac{s \log p}{n} \right) \geq 1 - 12p^{-1}.$$

In addition, in the special case of MUCR model, Theorem 4 and 5 imply the following result.

COROLLARY 1. *Under the conditions of Theorem 3 and consider the MUCR model with missing probability $\epsilon$. If $(M_{n,p}^2 \frac{s \log p}{n} \vee \sqrt{\frac{\log p}{n}}) \cdot \frac{1}{(1-\epsilon)^2} = o(1)$, then the risk of estimating the discriminant direction $\boldsymbol{\beta}$ by ADAM over the class $G(s, M_{n,p})$ satisfies*

$$\sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{ADAM}} - \boldsymbol{\beta}\|_2] \lesssim M_{n,p} \sqrt{\frac{s \log p}{n(1-\epsilon)^2}}.$$

*Moreover, there exist constant $C > 0$ and $\delta_n = o(1)$, such that the excess misclassification risk over the class $G(s, M_{n,p})$ satisfies*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}\left( R_{\boldsymbol{\theta}}(\hat{C}_{\text{ADAM}}) - R_{\text{opt}}(\boldsymbol{\theta}) \leq C \cdot e^{-(\frac{1}{8} + \delta_n) M_{n,p}^2} \cdot \frac{s \log p}{n(1-\epsilon)^2} \right) \geq 1 - 13p^{-1}.$$

This result shows that, although the sample size only loses a proportion of $\epsilon$, the convergence rates for the estimation risk and misclassification rate shrunk at the rate of $n(1-\epsilon)^2$ under the MUCR model.

### 3.3. Minimax lower bounds

To understand the difficulty of the classification problem and the related estimation problem as well as to establish the optimality for the AdaLDA and ADAM classifiers, it is essential to obtain the minimax lower bounds for the estimation risk and the excess misclassification risk. In this section, we only state the results for the missing data setting as the complete data setting can be treated as a special case. The following lower bound results show that the rates of convergence obtained by AdaLDA and ADAM algorithms are indeed optimal, for both estimation of the discriminant direction $\boldsymbol{\beta}$ and classification.

THEOREM 6. *Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and the distribution space $\boldsymbol{\Psi}(n_0; n, p)$ with $M_{n,p}\sqrt{\frac{s \log p}{n_0}} = o(1)$. For any $n_0 > 1$, suppose $1 \leq s \leq o(\frac{n_0}{\log p})$ and $\frac{\log p}{\log(p/s)} = O(1)$. Then under MCR model, the minimax risk of estimating the discriminant direction $\boldsymbol{\beta}$ over the class $G(s, M_{n,p})$ and $\boldsymbol{\Psi}(n_0; n, p)$ satisfies*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \boldsymbol{\Psi}(n_0; n, p)}} \mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2] \gtrsim M_{n,p} \sqrt{\frac{s \log p}{n_0}}.$$

THEOREM 7. *Consider the parameter space $\mathcal{G}(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and the distribution space $\boldsymbol{\Psi}(n_0; n, p)$ with $M_{n,p}\sqrt{\frac{s \log p}{n_0}} = o(1)$. For any $n_0 \geq 1$, suppose $1 \leq s \leq o(\frac{n_0}{\log p})$ and $\frac{\log p}{\log(p/s)} = O(1)$. Then under the MCR model, the minimax risk of the excess misclassification error over the class $G(s, M_{n,p})$ and $\boldsymbol{\Psi}(n_0; n, p)$ satisfies that*

(a) *If $M_{n,p} \leq C_b$ for some $C_b > 0$, then for any $\alpha > 0$, there are some constants $C_\alpha > 0$ such that*

$$\inf_{\hat{C}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \boldsymbol{\Psi}(n_0; n, p)}} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq C_\alpha \cdot \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

(b) *If $M_{n,p} \to \infty$ as $n \to \infty$, then for any $\alpha > 0$, there are some constants $C_\alpha > 0$ and $\tilde{\delta}_n = o(1)$ such that*

$$\inf_{\hat{C}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \boldsymbol{\Psi}(n_0; n, p)}} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq C_\alpha \cdot e^{-(\frac{1}{8} + \tilde{\delta}_n) M_{n,p}^2} \cdot \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

REMARK 3.6. In the complete data case, $n_{min}^* = \min\{n_1, n_2\} = n$, so Theorems 6 and 7 together with Theorems 1-4 imply that both AdaLDA and ADAM algorithms attain the optimal rates of convergence in terms of estimation and classification error.

We should also note that the proof of Theorem 7 is not straightforward. This is partially due to the fact that the excess risk $R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta})$ does not satisfy the triangle inequality that is required by standard lower bound techniques. A key technique here is to make a connection to an alternative risk function. For a generic classification rule $\hat{C}$, we define

$$L_{\boldsymbol{\theta}}(\hat{C}) = \mathbb{P}_{\boldsymbol{\theta}}(\hat{C}(\boldsymbol{Z}) \neq C_{\boldsymbol{\theta}}(\boldsymbol{Z})), \tag{16}$$

where $C_{\boldsymbol{\theta}}(\boldsymbol{Z})$ is the Fisher's linear discriminant rule in (1). The following lemma enables us to reduce the loss $R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta})$ to the risk function $L_{\boldsymbol{\theta}}(\hat{C})$.

LEMMA 3. *Let $\boldsymbol{Z} \sim \frac{1}{2} N_p(\boldsymbol{\mu}_1, \Sigma) + \frac{1}{2} N_p(\boldsymbol{\mu}_2, \Sigma)$ with parameter $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$. If a classifier $\hat{C}$ satisfying $L_{\boldsymbol{\theta}}(\hat{C}) = o(1)$ as $n \to \infty$, then for sufficiently large $n$,*

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq \frac{\sqrt{2\pi}\Delta}{8} e^{\Delta^2/8} \cdot L_{\boldsymbol{\theta}}^2(\hat{C}).$$

Lemma 3 shows the relationship between the risk function $R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta})$ and a more "standard" risk function $L_{\boldsymbol{\theta}}(\hat{C})$, who has the following property which served the same purpose as the triangle inequality.

LEMMA 4. *Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, -\boldsymbol{\mu}, I_p)$ and $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\mu}}, -\tilde{\boldsymbol{\mu}}, I_p)$ with $\|\boldsymbol{\mu}\|_2 = \|\tilde{\boldsymbol{\mu}}\|_2 = \Delta/2$. For any classifier $C$, if $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|_2 = o(1)$ as $n \to \infty$, then for sufficiently large $n$,*

$$L_{\boldsymbol{\theta}}(C) + L_{\tilde{\boldsymbol{\theta}}}(C) \geq \frac{1}{\Delta} e^{-\Delta^2/8} \cdot \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|_2.$$

Using Lemmas 3 and 4, we can then use Fano's inequality to complete the proof of Theorem 7. The details are shown in Section 6.

In addition, similar minimax lower bounds for estimating $\boldsymbol{\beta}$ and the excess misclassification error can be established under the MUCR model. The following result shows that the convergence rates in Corollary 1 are minimax rate optimal.

THEOREM 8. *Under the conditions of Theorem 6 and MUCR model with missing probability $\epsilon$, and further assume that $((M_{n,p}^2 \frac{s \log p}{n}) \vee \sqrt{\frac{\log p}{n}}) \cdot \frac{1}{(1-\epsilon)^2} = o(1)$, then the minimax risk of*

*estimating the discriminant direction $\boldsymbol{\beta}$ by ADAM over the class $G(s, M_{n,p})$ under the MUCR model satisfies*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\theta} \in G(s, M_{n,p})} \mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2] \gtrsim M_{n,p} \sqrt{\frac{s \log p}{n(1-\epsilon)^2}}.$$

*Moreover, if $M_{n,p} \to \infty$ and $\epsilon < 1 - c_B$ for some $c_B \in (0,1)$, the minimax risk of the misclassification error over the class $G(s, M_{n,p})$ satisfies that for any $\alpha, \delta > 0$, there are some constants $C_\alpha > 0$, such that*

$$\inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \geq C_\alpha \cdot e^{-(\frac{1}{8}+\delta)M_{n,p}^2} \cdot \frac{s \log p}{n(1-\epsilon)^2}) \geq 1 - \alpha.$$

## 4.  Numerical results

The proposed AdaLDA and ADAM classifiers are easy to implement, and the `MATLAB` code is available at `https://github.com/linjunz/ADAM`. We investigate in this section the numerical performance of AdaLDA and ADAM using both simulated and real data.

### 4.1.  Simulations

In all the simulations, the sample size is $n_1 = n_2 = 100$ while the dimension $p$ varies from $400, 800$ to $1200$. The probability of being in either of the two classes is equal. We consider the following six models for the covariance matrix $\Sigma$ and the discriminating direction $\boldsymbol{\beta}$.

**Model 1 Erdős-Rényi random graph:** Let $\tilde{\Omega} = (\tilde{\omega}_{ij})$ where $\tilde{\omega}_{ij} = u_{ij}\delta_{ij}$, $\delta_{ij} \sim \text{Ber}(\rho)$ being the Bernoulli random variable with success probability $\rho = 0.2$ and $u_{ij} \sim \text{Unif}[0.5, 1] \cup [-1, -0.5]$. After symmetrizing $\tilde{\Omega}$, set $\Omega = \tilde{\Omega} + \{\max(-\phi_{\min}(\tilde{\Omega}), 0) + 0.05\}I_p$ to ensure the positive definiteness. Finally, $\Omega$ is standardized to have unit diagonals and $\Sigma = \Omega^{-1}$. The discriminating direction $\boldsymbol{\beta} = (5/\sqrt{s}, \ldots, 5/\sqrt{s}, 0, \ldots, 0)^\top$ is sparse such that only the first $s$ entries are nonzero.

**Model 2 Block sparse model:** $\Omega = (\mathbf{B} + \delta I_p)/(1+\delta)$ where $b_{ij} = b_{ji} = 10 \times \text{Ber}(0.5)$ for $1 \leq i \leq p/2$, $i < j \leq p$; $b_{ij} = b_{ji} = 10$ for $p/2 + 1 \leq i < j \leq p$; $b_{ii} = 1$ for $1 \leq i \leq p$. Here $\delta = \max(-\phi_{\min}(\mathbf{B}), 0) + 0.05$. The matrix $\Omega$ is also standardized to have unit diagonals and $\Sigma = \Omega^{-1}$. The discriminating direction $\boldsymbol{\beta} = (2/\sqrt{s}, \ldots, 2/\sqrt{s}, 0, \ldots, 0)^\top$ where the first $s$ entries are nonzero.

**Model 3 AR(1) model:** $(\Sigma_{ij})_{p \times p}$ with $\Sigma_{ij} = 0.9^{|i-j|}$. The discriminating direction $\boldsymbol{\beta} = (2/\sqrt{s}, \ldots, 2/\sqrt{s}, 0, \ldots, 0)^\top$ where the first $s$ entries are nonzero.

**Model 4 Varying diagonals model:** We first let $(\Sigma_{ij})_{p \times p}$ with $\Sigma_{ij} = 0.9^{|i-j|}$. Then we add $\boldsymbol{d} = (10, 10, 10, 10, 10, U_6, \ldots, U_p)$ to the diagonal entries of $\Sigma$, where $U_6, .U_7, \ldots, U_p$ $i.i.d. \sim U(0,1)$. The discriminating direction $\boldsymbol{\beta} = (1/\sqrt{s}, \ldots, 1/\sqrt{s}, 0, \ldots, 0)^\top$ where the first $s$ entries are nonzero.

**Model 5 Approximately sparse $\boldsymbol{\beta}$:** Let $(\Sigma_{ij})_{p \times p}$ with $\Sigma_{ij} = 0.9^{|i-j|}$. The discriminating direction $\boldsymbol{\beta} = (0.75, (0.75)^2, (0.75)^3, \ldots, (0.75)^p)^\top$ being approximately sparse.

**Model 6 Sparse $\boldsymbol{\delta}$ and $\Sigma$:** Let $\Omega = (\mathbf{B} + \delta I_p)/(1+\delta)$ where $b_{ij} = b_{ji} = 10 \times \text{Ber}(0.2)$ for $1 \leq i \leq p/2$, $i < j \leq p$; $b_{ij} = b_{ji} = 10$ for $p/2 + 1 \leq i < j \leq p$; $b_{ii} = 1$ for $1 \leq i \leq p$. In other words, only the first $p/2$ rows and columns of $\Omega$ are sparse, whereas the rest of the matrix is

not sparse. Here $\delta = \max(-\phi_{\min}(\mathbf{B}), 0) + 0.05$. The matrix $\Omega$ is also standardized to have unit diagonals and $\Sigma = \Omega^{-1}$. The mean difference vector $\boldsymbol{\delta} = (2/\sqrt{s}, \ldots, 2/\sqrt{s}, 0, \ldots, 0)^\top$ where the first $s = 10$ entries are nonzero. Finally, let $\boldsymbol{\beta} = \Omega\boldsymbol{\delta}$.

Given the covariance matrix $\Sigma$ and the discriminating direction $\boldsymbol{\beta}$ generated by the model above, the means are $\boldsymbol{\mu}_1 = (0, \ldots, 0)^\top$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 - \Sigma\boldsymbol{\beta}$. The missing mechanism is chosen such that each entry $X_{ki}$ is observed with probability $p = 1 - \epsilon \in (0, 1)$. We change the missing proportion $\epsilon$ from 0 to 0.2. We apply AdaLDA rule when the data is complete, i.e. $\epsilon = 0$, and apply ADAM rule when $\epsilon > 0$. The AdaLDA rule is then compared with the LPD [Cai and Liu, 2011], SLDA [Shao et al., 2011], FAIR [Fan and Fan, 2008], and NSC [Tibshirani et al., 2002] rules whose tuning parameters are chosen by five-fold cross-validation over the grid $\{\sqrt{\log p/n}, \frac{3}{2}\sqrt{\log p/n}, 2\sqrt{\log p/n}, \ldots, 5\sqrt{\log p/n}\}$. We also note that one commonly used method, the Naive Bayes rule is a special case of the NSC rule with tuning parameter $\lambda_\Delta = 0$, so it's not included in the comparison. In the following tables, the fitting times (in seconds) on a regular computer (Intel Core i7-3770, 3.40 GHz) and misclassification errors (in %) of different algorithms are recorded. The misclassification error of a classifier $\hat{C}$ is computed as

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\text{label}(\boldsymbol{Z}_i) \neq \hat{C}(\boldsymbol{Z}_i)\},$$

where $Z_i$'s are $N$ fresh samples from the same distribution as the training data. Here we let $N = 200$ and $\boldsymbol{Z}_i$ are drawn from the two classes with the same probability. Due to different signal-to-noise ratios across different models, the misclassification rates for the Fisher's rule vary in different models. For each setting, the number of repetition is set to 100.

According to the simulation results in Tables 1–5, the proposed AdaLDA algorithm, which is purely data-driven and tuning-free, has a much shorter fitting time than that of LPD, which requires choosing tuning parameters via cross-validation. In addition, due to the element-wise constraints in the optimization, the AdaLDA algorithm adapts to the heteroscedasticity of $\boldsymbol{a}$ in Lemma 1, and has a better, if not comparable, performance than that of the LPD algorithm with optimally chosen tuning parameters and outperforms all the other methods. This advantage is further demonstrated in Model 4, where the diagonals of covariance matrices $\Sigma$ vary significantly. According to Table 4, the AdaLDA algorithm has a significant improvement over the LPD rule. Furthermore, we considered simulation settings where $\boldsymbol{\beta}$ is not sparse. In Table 5, the AdaLDA algorithm still performs well and outperforms all the other methods when $\boldsymbol{\beta}$ is approximately sparse in Model 5. Under Model 6 where $\boldsymbol{\delta}$ and $\Sigma$ are individually sparse and $\Sigma$ is diagonally dominant, which is a setting favoring SLDA and FAIR. In this setting, the numerical performance of AdaLDA is not as good as SLDA and FAIR, but the differences are small.

In addition, we also investigate the numerical performance of ADAM for incomplete data. According to Table 6, which shows the performance of ADAM across different missing proportions $\epsilon$ under Model 1, ADAM does not lose much accuracy in the presence of missing data when the missing proportion $\epsilon$ is small. As expected, the misclassification errors of ADAM grows when $\epsilon$ increases. Since the pattern of the performances of ADAM are similar across different models, the simulation results of ADAM under Models 2-6 are given in the supplementary material.

### 4.2. Real data analysis

In addition to the simulation studies, we also illustrate the merits of the AdaLDA and ADAM classifiers in an analysis of two real datasets to further investigate the numerical performance of the proposed methods. One dataset, available at www.chestsurg.org, is the Lung cancer data analyzed by Gordon et al. (2002). Another dataset is the Leukemia data from high-density

**Table 1.** Misclassification errors (%) and model fitting times for Model 1 with complete data

| $(s,p)$ | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|---|---|---|---|---|---|---|
| (10,400) | 17.50(1.51) | 18.50(0.42) | 42.77(1.89) | 30.42(1.34) | 34.05(1.21) | 12.60(0.51) |
|  | [0.58s] | [55.02s] | [2.18s] | [1.40s] | [42.56s] |  |
| (20,400) | 19.73(0.54) | 20.65(0.72) | 41.60(2.10) | 25.92(0.71) | 26.87(0.75) | 11.05(0.61) |
|  | [0.58s] | [49.73s] | [2.50s] | [2.96s] | [43.54s] |  |
| (10,800) | 20.15(1.24) | 25.37(1.62) | 41.46(2.14) | 29.55(0.81) | 33.60(1.01) | 15.13(0.42) |
|  | [3.39s] | [187.03s] | [8.74s] | [5.15s] | [111.60s] |  |
| (20,800) | 28.30(1.07) | 29.10(1.63) | 43.68(2.49) | 31.58(0.97) | 31.62(0.86) | 14.30(0.74) |
|  | [3.35s] | [195.59s] | [7.18s] | [5.62s] | [115.15s] |  |
| (10,1200) | 26.10(0.73) | 26.32(0.80) | 42.26(2.45) | 31.78(0.75) | 34.73(0.71) | 16.00(0.60) |
|  | [9.90s] | [531.74s] | [28.21s] | [21.43s] | [244.57s] |  |
| (20,1200) | 32.96(1.72) | 35.70(1.68) | 44.23(2.65) | 37.48(2.31) | 36.67(1.01) | 18.90(0.58) |
|  | [9.94s] | [493.31s] | [28.14s] | [26.41s] | [244.28s] |  |
| (10,1600) | 24.40(0.52) | 28.44(2.41) | 43.14(3.16) | 32.48(0.89) | 34.55(0.99) | 19.90(0.51) |
|  | [21.77s] | [809.22s] | [56.78s] | [34.36s] | [333.84s] |  |
| (20,1600) | 26.20(0.71) | 30.87(2.05) | 44.24(2.49) | 38.52(2.56) | 35.15(0.92) | 17.35(0.39) |
|  | [21.75s] | [1019.35s] | [54.92s] | [34.64s] | [421.86s] |  |

**Table 2.** Misclassification errors (%) and model fitting times for Model 2 with complete data

| $(s,p)$ | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|---|---|---|---|---|---|---|
| (10,400) | 11.88(0.16) | 12.57(0.15) | 14.05(0.66) | 17.52(0.70) | 17.58(0.78) | 11.35(0.56) |
|  | [0.59s] | [66.8s] | [3.12s] | [1.56s] | [38.52s] |  |
| (20,400) | 10.53(0.94) | 11.28(0.67) | 12.03(0.43) | 12.28(0.41) | 12.25(0.40) | 7.40(0.45) |
|  | [0.62s] | [71.8s] | [1.72s] | [1.25s] | [37.15s] |  |
| (10,800) | 13.40(1.01) | 16.60(1.78) | 15.10(0.66) | 18.48(0.72) | 21.98(0.67) | 13.35(0.64) |
|  | [3.44s] | [232.64s] | [9.52s] | [5.77s] | [114.5s] |  |
| (20,800) | 13.45(0.98) | 16.85(1.75) | 14.48(0.68) | 15.28(0.75) | 16.53(0.68) | 9.85(0.41) |
|  | [3.34s] | [245.62s] | [8.58s] | [5.93s] | [111.56s] |  |
| (10,1200) | 15.20(0.21) | 17.57(1.04) | 18.20(0.26) | 18.88(0.53 | 21.68(0.73) | 12.93(0.50) |
|  | [9.87s] | [577.16s] | [17.16s] | [11.08s] | [243.50s] |  |
| (20,1200) | 14.27(0.82) | 15.20(0.87) | 17.40(0.42) | 15.93(0.71) | 17.68(1.01) | 9.72(0.28) |
|  | [9.90s] | [600.76s] | [18.72s] | [9.67s] | [245.75s] |  |
| (10,1600) | 14.38(0.43) | 15.74(1.01) | 15.35(0.42) | 16.08(0.73) | 22.07(0.68) | 11.77(0.42) |
|  | [21.87s] | [1215.31s] | [29.02s] | [18.88s] | [420.08s] |  |
| (20,1600) | 15.74(0.65) | 17.60(1.47) | 16.46(0.53) | 16.97(0.71) | 19.90(0.81) | 12.03(0.30) |
|  | [21.75s] | [1212.03s] | [30.26s] | [17.16s] | [413.55s] |  |

**Table 3.** Misclassification errors (%) and model fitting times for Model 3 with complete data

| $(s,p)$ | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|---|---|---|---|---|---|---|
| (10,400) | 27.98(0.90) | 29.65(1.12) | 33.77(0.88) | 37.15(1.16) | 30.00(1.00) | 23.12(0.82) |
| | [0.61s] | [79.42s] | [3.43s] | [0.62s] | [38.22s] | |
| (20,400) | 35.17(0.82) | 36.40(0.80) | 40.45(0.71) | 43.73(0.85) | 37.78(0.90) | 29.08(0.93) |
| | [0.62s] | [73.49s] | [2.50s] | [2.65s] | [37.75s] | |
| (10,800) | 28.45(0.80) | 34.75(0.79) | 34.83(0.66) | 42.18(1.10) | 31.87(0.67) | 22.50(0.50) |
| | [3.36s] | [255.77s] | [10.14s] | [5.77s] | [112.40s] | |
| (20,800) | 34.25 (0.68) | 41.52 (0.67) | 41.85 (1.32) | 44.98 (1.46) | 39.13 (0.68) | 29.83 (0.49) |
| | [3.39s] | [250.27s] | [9.52s] | [5.15s] | [113.40s] | |
| (10,1200) | 28.23(0.59) | 34.10(1.05) | 34.53(0.83) | 41.08(0.87) | 30.15(0.56) | 21.15(0.72) |
| | [9.81s] | [903.35s] | [24.18s] | [14.98s] | [250.96s] | |
| (20,1200) | 34.65(0.95) | 41.68(1.33) | 41.53(1.48) | 46.45(1.05) | 38.07(1.42) | 28.05(1.15) |
| | [9.93s] | [929.33s] | [24.03s] | [16.0s] | [256.52s] | |
| (10,1600) | 27.88(0.89) | 34.13(1.02) | 35.57(1.23) | 41.32(1.12) | 30.73(0.78) | 22.17(0.58) |
| | [21.71s] | [1321.28s] | [27.92s] | [17.78s] | [414.58s] | |
| (20,1600) | 33.45(0.72) | 37.82(0.50) | 41.80(1.52) | 46.05(1.13) | 38.70(0.95) | 27.65(0.42) |
| | [21.06s] | [1864.71s] | [41.03s] | [28.70s] | [478.12s] | |

**Table 4.** Misclassification errors (%) and model fitting times for Model 4 with complete data

| $(s,p)$ | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|---|---|---|---|---|---|---|
| (10,400) | 10.80(0.45) | 16.03(0.57) | 23.87(1.04) | 16.27(0.58) | 12.33(0.77) | 8.18 (0.40) |
| | [0.59s] | [77.95s] | [5.30s] | [2.34s] | [38.02s] | |
| (20,400) | 16.42(0.68) | 24.05(0.62) | 32.27(0.90) | 23.20(1.01) | 20.50(0.80) | 11.22(0.59) |
| | [0.60s] | [78.06s] | [4.68s] | [2.62s] | [37.75s] | |
| (10,800) | 12.03(0.57) | 21.17(0.70) | 29.00(0.81) | 23.95(0.40) | 17.20(1.00) | 10.05(0.33) |
| | [3.41s] | [252.72s] | [7.33s] | [6.08s] | [111.58s] | |
| (20,800) | 18.48(0.85) | 25.40(0.75) | 36.20(0.71) | 26.28(0.69) | 26.22(0.53) | 11.20(0.72) |
| | [3.35s] | [249.07s] | [9.67s] | [5.15s] | [111.34] | |
| (10,1200) | 13.98(0.58) | 21.90(0.82) | 28.63(0.49) | 25.52(0.48) | 17.40(0.70) | 9.27(0.42) |
| | [9.89s] | [630.88s] | [15.91s] | [12.32s] | [244.53s] | |
| (20,1200) | 21.70(0.90) | 29.77(0.46) | 34.33(0.64) | 28.40(0.88) | 25.82(0.72) | 10.70(0.61) |
| | [9.95s] | [631.24s] | [14.82s] | [9.67s] | [245.78s] | |
| (10,1600) | 13.90(0.42) | 22.07(0.65) | 27.05(0.71) | 28.73(0.90) | 20.20(0.37) | 11.67(0.49) |
| | [21.29s] | [1846.61s] | [40.56s] | [27.77s] | [480.78s] | |
| (20,1600) | 23.95(1.03) | 29.57(0.92) | 34.60(0.78) | 28.98(1.09) | 23.05(0.45) | 15.90(0.82) |
| | [21.56s] | [1215.36s] | [27.46s] | [17.47s] | [414.57s] | |

**Table 5.** Misclassification errors (%) and model fitting times for Model 5 (the first four rows) and 6 (the last four rows) with complete data

| Method | AdaLDA | LPD | SLDA | FAIR | NSC | Oracle |
|---|---|---|---|---|---|---|
| $p = 400$ | 27.98(0.66) | 31.68(0.46) | 46.72(0.83) | 34.68(0.61) | 31.90(0.60) | 18.55(0.56) |
| | [0.61s] | [84.19s] | [4.28s] | [2.03s] | [39.23s] | |
| $p = 800$ | 29.50(0.73) | 34.15(1.03) | 46.30(0.64) | 37.20(0.87) | 33.63(0.96) | 19.35(0.73) |
| | [3.37s] | [253.53s] | [10.76s] | [7.49s] | [112.38s] | |
| $p = 1200$ | 27.43(1.25) | 36.40(0.78) | 46.68(0.82) | 39.20(0.74) | 33.45(0.72) | 18.53(0.88) |
| | [10.00s] | [642.98s] | [15.76s] | [12.01s] | [248.86s] | |
| $p = 1600$ | 27.38(0.64) | 37.82(0.65) | 47.30(0.80) | 41.38(1.12) | 35.98(1.11) | 20.62(0.68) |
| | [21.67s] | [1314.69s] | [26.83s] | [17.94s] | [424.20s] | |
| $p = 400$ | 18.27(0.74) | 28.52(0.70) | 20.60(0.64) | 18.98(0.68) | 21.12(0.78) | 4.77(0.25) |
| | [0.60s] | [56.94s] | [4.84s] | [4.62s] | [37.66s] | |
| $p = 800$ | 19.10(0.99) | 26.68(2.06) | 21.30(0.89) | 18.82(0.73) | 21.10(0.61) | 4.30(0.23) |
| | [3.34s] | [202.69s] | [7.18s] | [5.62s] | [111.74s] | |
| $p = 1200$ | 18.50(0.83) | 25.60(2.12) | 18.90(0.60) | 17.25(0.52) | 20.18(0.73) | 4.25(0.33) |
| | [9.93s] | [514.57s] | [15.60s] | [11.39s] | [244.46s] | |
| $p = 1600$ | 18.23(1.08) | 24.65(1.94) | 18.80(0.68) | 17.32(0.60) | 22.95(0.76) | 4.62(0.24) |
| | [21.65s] | [1105.70s] | [27.46s] | [17.32s] | [420.25s] | |

**Table 6.** Misclassification errors (%) and model fitting times for Model 1 with missing proportion $\epsilon$

| Method | ADAM | | | | AdaLDA |
|---|---|---|---|---|---|
| $(s, p)\backslash\epsilon$ | 0.2 | 0.15 | 0.1 | 0.05 | 0 |
| (10,400) | 20.55(0.91) | 19.70(1.11) | 19.20(1.34) | 18.40(0.78) | 17.50(1.51) |
| | [0.68s] | [0.65s] | [0.66s] | [0.65s] | [0.58s] |
| (20,400) | 24.95(1.34) | 23.60(0.26) | 23.67(0.10) | 21.28(0.91) | 19.73(0.54) |
| | [0.67s] | [0.66s] | [0.64s] | [0.64s] | [0.58s] |
| (10,800) | 26.28(1.10) | 25.30(0.72) | 22.15(0.74) | 21.18(1.08) | 20.15(1.24) |
| | [3.60s] | [3.60s] | [3.59s] | [3.59s] | [3.39s] |
| (20,800) | 34.00(1.02) | 33.55(1.56) | 32.97(1.14) | 31.90(0.95) | 28.30(1.07) |
| | [3.63s] | [3.61s] | [3.64s] | [3.60s] | [3.35s] |
| (10,1200) | 30.62(1.12) | 30.39(1.08) | 29.65(1.72) | 27.42(1.16) | 26.10(0.73) |
| | [10.56s] | [10.73s] | [10.56s] | [10.53s] | [9.90s] |
| (20,1200) | 35.62(0.81) | 34.10(1.62) | 33.95(0.91) | 33.77(1.04) | 32.96(1.72) |
| | [10.51s] | [10.54s] | [10.52s] | [10.52s] | [9.94s] |
| (10,1600) | 33.47(1.59) | 30.53(0.79) | 27.40(1.61) | 26.40(1.52) | 24.40(0.52) |
| | [23.01s] | [22.94s] | [23.11s] | [23.12s] | [21.77s] |
| (20,1600) | 37.40(0.91) | 33.79(0.74) | 32.70(1.12) | 31.77(1.01) | 26.20(0.71) |
| | [23.00s] | [22.95s] | [23.01s] | [22.96s] | [21.75s] |

**Table 7.** Classification error of Lung cancer data by various methods

|  | ADAM$_{(\epsilon=0.1)}$ | ADAM$_{(\epsilon=0.05)}$ | AdaLDA | LPD | SLDA | FAIR | NSC |
|---|---|---|---|---|---|---|---|
| Testing error | 5.53% | 3.22% | 2.09% | 2.11% | 4.88% | 3.64% | 7.30% |

**Table 8.** Classification error of Leukemia data by various methods

|  | ADAM$_{(\epsilon=0.1)}$ | ADAM$_{(\epsilon=0.05)}$ | AdaLDA | LPD | SLDA | FAIR | NSC |
|---|---|---|---|---|---|---|---|
| Testing error | 8.47% | 7.53% | 2.94% | 3.09% | 5.76% | 2.94% | 8.82% |

Affymetrix oligonucleotide arrays that was previously analyzed in Golub et al. (1999), and is available at www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. These two datasets were frequently used for illustrating the empirical performance of the classifier for high-dimensional data in recent literature. We will compare AdaLDA and ADAM with the existing methods.

*4.2.1. Lung cancer data*

We evaluate the proposed methods by classifying between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples (31 MPM and 150 ADCA) and each sample is described by 12533 genes in the lung cancer dataset in Gordon et al. (2002). This dataset has been analyzed in Fan and Fan (2008) using FAIR and NSC. In this section we apply the AdaLDA and ADAM rules to this dataset for disease classification. When ADAM rule is used, we make each entry in the dataset missing uniformly and independently with probability $\epsilon$. In the simulation, given the small sample size, we choose $\epsilon = 0.05$ and $\epsilon = 0.1$.

The sample variances of the genes range over a wide interval. We first compute the sample variances for each gene and drop the lower and upper 6-quantiles to control the condition number of $\hat{\Sigma}$. The average misclassification errors are computed by using 5-fold cross-validation for various methods with 50 repetitions. To reduce the computational costs, in each repetition, only 1500 genes with the largest absolute values of the two sample $t$ statistics are used. We then apply all the aforementioned methods to this reduced dimensional dataset. As seen in the Table 7, the classification result of AdaLDA is better than existing methods, including LPD [Cai and Liu, 2011], SLDA [Shao et al., 2011], FAIR [Fan and Fan, 2008], and NSC [Tibshirani et al., 2002] methods, although only 1500 genes were used. Moreover, in the incomplete data case, ADAM still has satisfactory accuracy.

*4.2.2. Leukemia data*

Golub et al. (1999) applied gene expression microarray techniques to study human acute leukemia and discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). There are 72 tissue samples (47 ALL and 25 AML) and 7129 genes in the Leukemia dataset. In this section, we apply the AdaLDA rule to this dataset and compare the classification results with those obtained by LPD [Cai and Liu, 2011], SLDA [Shao et al., 2011], FAIR [Fan and Fan, 2008], and NSC [Fan and Fan, 2008] methods. Same as the analysis of lung cancer data, when ADAM rule is used, we make each entry in the dataset missing independently with probability $\epsilon \in \{0.05, 0.1\}$

As in the analysis of the lung cancer data, we first drop genes with extreme sample variances out of lower and upper 6-quantiles. Similar to the analysis of the lung cancer data, the average misclassification errors are computed by using two-fold cross-validation for various methods with 50 repetitions, and to control the computational costs, we use 2000 genes with the largest absolute values of the two sample $t$ statistics in each repetition. After the application of all methods to the same reduced dimensional dataset, classification results are then summarized in Table 8. The

AdaLDA has the similar performance as the LPD rule and FAIR, as obtain the misclassification error of about 3%. In contrast, the navie-Bayes rule misclassifies 20.59% testing samples and SLDA misclassifies 5.76% testing samples. Fan and Fan [2008] report a test error rate of 2.94% for FAIR and a test error rate of 8.82% for NSC proposed by Tibshirani et al. [2002]. In the presence of missing data, ADAM misclassifies 7.53% and 8.47% testing samples when the missing proportion is 0.05 and 0.1 respectively.

## 5.  Extension to multiple-class LDA

We have so far focused on the two-class high-dimensional LDA. The procedure can be extended to the following $K$-class setting:

$$X_1^{(k)}, ..., X_{n_k}^{(k)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_k, \Sigma), \text{ for } k = 1, ..., K.$$

For ease of presentation, we focus on the complete data case in this section. For an independent observation $\boldsymbol{Z}$ drawn from these $K$ distributions with prior probabilities $\pi_1, ..., \pi_K$, let $L_k(\boldsymbol{Z}) = \pi_k \cdot |\Sigma|^{-1/2} \exp(-1/2(\boldsymbol{Z} - \boldsymbol{\mu}_K)^\top \Sigma^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}_K))$. Then

$$\mathbb{P}(\text{label}(\boldsymbol{Z}) = k \mid \boldsymbol{Z}) = \frac{\mathbb{P}(\boldsymbol{Z} \mid \text{label}(\boldsymbol{Z}) = k) \cdot \mathbb{P}(\text{label}(\boldsymbol{Z}) = k)}{\sum_{k=1}^{K} \mathbb{P}(\boldsymbol{Z} \mid \text{label}(\boldsymbol{Z}) = k) \cdot \mathbb{P}(\text{label}(\boldsymbol{Z}) = k)} = \frac{L_k(\boldsymbol{Z})}{\sum_{k=1}^{K} L_k(\boldsymbol{Z})},$$

and the oracle classification rule is given by

$$\begin{aligned}
C_{\boldsymbol{\theta}}(\boldsymbol{Z}) &= \arg\max_{k \in [K]} \mathbb{P}(\text{label}(\boldsymbol{Z}) = k \mid \boldsymbol{Z}) = \arg\max_{k \in [K]} L_k(\boldsymbol{Z}) = \arg\max_{k \in [K]} \frac{L_k(\boldsymbol{Z})}{L_1(\boldsymbol{Z})} \\
&= \arg\max_{k \in [K]} \log(\frac{L_k(\boldsymbol{Z})}{L_1(\boldsymbol{Z})}) = \arg\max_{k \in [K]} D_k,
\end{aligned} \tag{17}$$

where $D_1 = 0$ and $D_k = (\boldsymbol{Z} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_K}{2})^\top \boldsymbol{\beta}_k + \log(\frac{\pi_k}{\pi_1})$ for $k = 2, ..., K$, with $\boldsymbol{\beta}_k = \Omega(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$.

A similar data-driven adaptive classifier, called K-class AdaLDA, can then be constructed based on the estimation of $\boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_K$ and $\Delta_k = \sqrt{\boldsymbol{\beta}_k^\top \Sigma \boldsymbol{\beta}_k}$, as follows.

Let $\hat{\boldsymbol{\mu}}_k, k \in [K]$ and $\hat{\Sigma}$ be the sample means and pooled sample covariance matrix respectively.

**Step 1 (Estimating $\Delta_k^2$).** Fix $\lambda_0 = 25/2$. For $k = 2, ..., K$, we estimate $\boldsymbol{\beta}_k$ by a preliminary estimator

$$\tilde{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1$$

$$\text{subject to } |\boldsymbol{e}_j^\top \left(\hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)\right)| \le 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}} \cdot (\lambda_0 \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1) + 1), \ j \in [p]. \tag{18}$$

Then we estimate $\Delta_k^2$ by $\hat{\Delta}_k^2 = |\tilde{\boldsymbol{\beta}}_k^\top (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)|, \ k = 2, .., K$.

**Step 2 (Adaptive estimation of $\boldsymbol{\beta}_k$).** Given $\hat{\Delta}^2$, the final estimator $\hat{\boldsymbol{\beta}}_k$ of $\boldsymbol{\beta}_k$ is constructed through the following linear optimization

$$\hat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1$$

$$\text{subject to } |\boldsymbol{e}_j^\top \left(\hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}}_1)\right)| \le 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}(\lambda_0 \hat{\Delta}_k^2 + 1)}, \ j \in [p]. \tag{19}$$

**Step 3 (Construction of $K$-class AdaLDA).** The $K$-class AdaLDA classification rule is obtained by plugging $\hat{\boldsymbol{\beta}}_k$ into Fisher's rule (17),

$$\hat{C}_{\text{K}-\text{AdaLDA}}(\boldsymbol{Z}) = \arg\max_{k \in [K]} \hat{D}_k, \tag{20}$$

where $\hat{D}_1 = 0$ and $\hat{D}_k = (\boldsymbol{Z} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_K}{2})^\top \hat{\boldsymbol{\beta}}_k + \log(\hat{\pi}_k/\hat{\pi}_1)$ with $\hat{\pi}_k = n_k/\sum_{j=1}^K n_j$ for $k = 2, ..., K$.

For theoretical analysis, we consider the following parameter space $\mathcal{G}_K(s, M_{n,p})$ defined by

$$\mathcal{G}_K(s, M_{n,p}) = \{\boldsymbol{\theta} = (\pi_1, ... \pi_K, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \Sigma) : \boldsymbol{\mu}_k \in \mathbb{R}^p, \pi_k \in (c, 1-c), \sum_{k=1}^K \pi_k = 1, \Sigma \in \mathbb{R}^{p \times p}, \Sigma \succ 0,$$

$$\|\boldsymbol{\beta}_k\|_0 \le s, M^{-1} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le M, M_{n,p} \le \Delta \le 3M_{n,p}\},$$

where $M > 1$ and $c \in (0, 1/2)$ are some constants, $M_{n,p} > 0$ can potentially grow with $n$ and $p$.

Theoretical properties of $K$-class AdaLDA can be established by applying the same technical argument as before.

THEOREM 9. *Consider the parameter space $\mathcal{G}_K(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$. Suppose $\boldsymbol{X}_1^{(k)}, ..., \boldsymbol{X}_{n_k}^{(k)} \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}_k, \Sigma)$ for $k = 1, ..., K$. Assume that $M_{n,p}\sqrt{\frac{s \log p}{n}} = o(1)$. Then for $k \in [K]$,*

$$\sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{E}[\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_2] \lesssim M_{n,p}\sqrt{\frac{s \log p}{n}}.$$

The following theorem provides an upper bound for the excess misclassification risk $R_{\boldsymbol{\theta}}(\hat{C}_{\text{K}-\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta})$ of the $K$-class AdaLDA rule.

THEOREM 10. *Consider the parameter space $\mathcal{G}_K(s, M_{n,p})$ with $M_{n,p} > c_L$ for some $c_L > 0$ and assume the conditions in Theorem 2 hold.*

(a) *If $M_{n,p} \le C_b$ for some $C_b > 0$, then there exists some constant $C > 0$,*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}_K(s, M_{n,p})} \mathbb{P}\left(R_{\boldsymbol{\theta}}(\hat{C}_{\text{K}-\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta}) \le C \cdot \frac{s \log p}{n}\right) \ge 1 - 8p^{-1}.$$

(b) *If $M_{n,p} \to \infty$ as $n \to \infty$, then there exist some constant $C > 0$ and $\delta_n = o(1)$, such that*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}_K(s, M_{n,p})} \mathbb{P}\left(R_{\boldsymbol{\theta}}(\hat{C}_{\text{K}-\text{AdaLDA}}) - R_{\text{opt}}(\boldsymbol{\theta}) \le C \cdot e^{-(\frac{1}{8} + \delta_n)M_{n,p}^2} \cdot \frac{s \log p}{n}\right) \ge 1 - 8p^{-1}.$$

REMARK 5.1. Multi-class sparse discriminant analysis has been considered in the literature [Witten and Tibshirani, 2011, Clemmensen et al., 2011, Mai et al., 2019]. However, there are no theoretical justification for the methods in Witten and Tibshirani [2011] and Clemmensen et al. [2011], and Mai et al. [2019] only shows consistency. In contrast, our result gives an explicit convergence rate.

## 6. Proofs

In this section, we prove the main results, Theorem 2, 3, 4 5, 6 and 7. Theorem 1 follows from Theorems 3 and 7. Since $n_1 \asymp n_2$, without loss of the generality we shall assume $n_1 = n_2 = n$ in the proofs. For reasons of space, the proofs of the technical lemmas are given in the Supplementary Material [Cai and Zhang, 2018].

## 6.1.  Proof of Theorem 2

To prove Theorem 2 we begin by collecting a few important technical lemmas that will be used in the main proofs.

### 6.1.1.  Auxiliary Lemmas

LEMMA 5. *Suppose* $\mathbf{X}_1, ..., \mathbf{X}_n$ *i.i.d.* $\sim N_p(\boldsymbol{\mu}, \Sigma)$, *and assume that* $\hat{\boldsymbol{\mu}}$, $\hat{\Sigma}$ *are the sample mean and sample covariance matrix respectively. Let* $\Gamma(s) = \{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}_{S^C}\|_1 \le \|\boldsymbol{u}_S\|_1$, *for some* $S \subset [p]$ *with* $|S| = s\}$, *then with probability at least* $1 - p^{-1}$,

$$\sup_{\boldsymbol{u} \in \Gamma(s)} \boldsymbol{u}^\top (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \lesssim \sqrt{\frac{s \log p}{n}};$$

$$\sup_{\boldsymbol{u}, \boldsymbol{v} \in \Gamma(s)} \boldsymbol{u}^\top (\hat{\Sigma} - \Sigma) \boldsymbol{v} \lesssim \sqrt{\frac{s \log p}{n}}.$$

LEMMA 6. *Suppose* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$. *Let* $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{y}$ *and* $\mathcal{S} = \operatorname{supp}(\boldsymbol{y})$. *If* $\|\boldsymbol{x}\|_1 \le \|\boldsymbol{y}\|_1$, *then* $h \in \Gamma(s)$ *with* $s = |S|$, *that is,*

$$\|\boldsymbol{h}_{\mathcal{S}^c}\|_1 \le \|\boldsymbol{h}_S\|_1.$$

### 6.1.2.  Main proof of Theorem 2

Recall that $\hat{\boldsymbol{\beta}}_{\mathrm{AdaLDA}}$ is constructed by the following two steps.

**Step 1.** Estimating $\Delta^2$

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ |\boldsymbol{e}_j^\top \left( \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \right)| \le 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}} \cdot (\lambda_0 \boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) + 1),\ j \in [p] \right\}. \tag{21}$$

Then we estimate $\Delta^2$ by $\hat{\Delta}^2 = |\langle \tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1 \rangle|$.

**Step 2.** Adaptive estimation of $\boldsymbol{\beta}$. Given $\hat{\Delta}^2$, the final estimator $\hat{\boldsymbol{\beta}}_{\mathrm{AdaLDA}}$ of $\boldsymbol{\beta}$ is constructed by the following linear optimization problem

$$\hat{\boldsymbol{\beta}}_{\mathrm{AdaLDA}} = \arg\min_{\boldsymbol{\beta}} \left\{ |\boldsymbol{e}_j^\top \left( \hat{\Sigma}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \right)| \le 4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\lambda_0 \hat{\sigma}_{jj} \hat{\Delta}^2 + \hat{\sigma}_{jj}},\ j \in [p] \right\}. \tag{22}$$

Firstly, let's show the consistency of estimating $\Delta^2$. Recall the definition of $\tilde{\boldsymbol{\beta}}$ and using Lemma 5, we have with high probability at least $1 - 3p^{-1}$,

$$
\begin{aligned}
|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| \le{} & |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}})| + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma} - \Sigma)\tilde{\boldsymbol{\beta}})| + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}})| \\
\le{} & \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \|\hat{\Sigma}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}}\|_\infty + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma} - \Sigma)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}))| + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma} - \Sigma)\boldsymbol{\beta})| \\
& + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}})| \\
\lesssim{} & \sqrt{s} \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \|\hat{\Sigma}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}}\|_\infty + \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \sqrt{\frac{s \log p}{n}} \cdot \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2 \\
& + \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2 \sqrt{\frac{s \log p}{n}} \cdot \|\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2 \sqrt{\frac{s \log p}{n}},
\end{aligned}
\tag{23}
$$

where the third inequality uses Lemma 5 and the fact that $\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \in \Gamma(s)$. In fact, $\boldsymbol{\beta}$ is a feasible solution to (8) due to Lemma 1 and thus $\|\tilde{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}\|_1$. Then by Lemma 6, we have $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \in \Gamma(s)$. In addition, $\|\boldsymbol{\beta}\|_0 \leq s$, so we have $\boldsymbol{\beta} \in \Gamma(s)$.

In addition, by standard derivation of the accuracy of sample variance, since $M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M$, by using the union bound technique, we have with probability at least $1 - p^{-1}$,

$$\max_{i \in [p]} |\hat{\sigma}_{ii} - \sigma_{ii}| \lesssim \sqrt{\frac{\log p}{n}},$$

which implies with probability at least $1 - p^{-1}$,

$$\max_{i \in [p]} |\hat{\sigma}_{ii}| \leq 2M.$$

In addition, since $\Delta \geq M_{n,p} \geq c_L > 0$, then with probability at least $1 - 3p^{-1}$,

$$\begin{aligned}
\|\hat{\Sigma}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}}\|_\infty \leq &4\sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}} \cdot (\lambda_0 \tilde{\boldsymbol{\beta}}^\top (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) + 1) \\
\lesssim &\sqrt{\frac{\log p}{n}} \cdot |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) + 1| + \sqrt{\frac{\log p}{n}} \cdot |\boldsymbol{\beta}^\top (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)| \\
\leq &\sqrt{\frac{\log p}{n}} \cdot (|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)| + |(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)| + 1) \\
&+ \sqrt{\frac{\log p}{n}} \cdot (|\boldsymbol{\beta}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)| + |\boldsymbol{\beta}^\top (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)|) \\
\lesssim &\sqrt{\frac{\log p}{n}} \Delta \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \sqrt{s} \cdot \frac{\log p}{n} \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \sqrt{\frac{\log p}{n}} \Delta^2 + \sqrt{s} \cdot \frac{\log p}{n} \Delta,
\end{aligned}$$

where the last inequality uses the fact that $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2, \|\boldsymbol{\beta}\|_2 \lesssim \Delta$, since $\Delta = \sqrt{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \Omega (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)} \geq \frac{1}{\sqrt{M}} \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2$, and $\Delta = \sqrt{\boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta}} \geq \frac{1}{\sqrt{M}} \|\boldsymbol{\beta}\|_2$.

It follows that with probability at least $1 - 6p^{-1}$,

$$\begin{aligned}
|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| \lesssim &\sqrt{\frac{s \log p}{n}} \Delta \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 + \frac{s \log p}{n} \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 + \sqrt{\frac{s \log p}{n}} \Delta^2 \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \\
&+ \frac{s \log p}{n} \Delta \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \sqrt{\frac{s \log p}{n}} \cdot \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \\
&+ \sqrt{\frac{s \log p}{n}} \Delta \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \sqrt{\frac{s \log p}{n}} \\
\lesssim &\sqrt{\frac{s \log p}{n}} \Delta \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 + \sqrt{\frac{s \log p}{n}} \Delta^2 \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2,
\end{aligned}$$

where the last inequality uses the fact that $\Delta \geq M_{n,p} \geq c_L > 0$.

On the other hand, since

$$|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| \geq \lambda_{\min}(\Sigma) \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \geq \frac{1}{M} \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2.$$

We then have, with probability at least $1 - 6p^{-1}$,

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \sqrt{\frac{s \log p}{n}} \left( \Delta \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \Delta^2 \right),$$

Assuming $M_{n,p}\sqrt{\frac{s\log p}{n}} = o(1)$, which implies $\Delta\sqrt{\frac{s\log p}{n}} = o(1)$, then we have

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \frac{\Delta^2\sqrt{\frac{s\log p}{n}}}{1 - \Delta\sqrt{\frac{s\log p}{n}}}.$$

Since $\|\tilde{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}\|_1$ and combining with Lemma 5, we then have with probability at least $1 - 7p^{-1}$,

$$\begin{aligned}
|\frac{\hat{\Delta}^2 - \Delta^2}{\Delta^2}| &\leq \frac{|\tilde{\boldsymbol{\beta}}^\top(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}})| + |\boldsymbol{\delta}^\top(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})|}{\Delta^2} \leq \frac{\|\boldsymbol{\beta}\|_1 \cdot \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|_\infty + \|\boldsymbol{\delta}\|_2 \cdot \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2}{\Delta^2} \\
&\leq \frac{\sqrt{s} \cdot \|\boldsymbol{\beta}\|_2 \cdot \|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|_\infty + \|\boldsymbol{\delta}\|_2 \cdot \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2}{\Delta^2} \\
&\lesssim \frac{\sqrt{s} \cdot \Delta\sqrt{\frac{\log p}{n}} + \Delta \cdot \frac{\Delta^2\sqrt{\frac{s\log p}{n}}}{1 - \Delta\sqrt{\frac{s\log p}{n}}}}{\Delta^2} = o(1),
\end{aligned}$$

given $\Delta \geq c_L$ and $\Delta\sqrt{\frac{s\log p}{n}} = o(1)$.

Secondly, let's proceed to showing the accuracy of $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$. We use $\hat{\boldsymbol{\beta}}$ to denote $\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}$ in this subsection for simplicity. By Lemma 1, $\boldsymbol{\beta}$ lies in the feasible set of (9), so $\|\hat{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}\|_1$. By a similar argument as in (23), we have that with probability at least $1 - 3p^{-1}$,

$$\begin{aligned}
&|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})| \\
&\leq |(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}})| + |(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\Sigma} - \Sigma)\hat{\boldsymbol{\beta}})| + |(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\boldsymbol{\delta} - \hat{\boldsymbol{\delta}})| \\
&\lesssim \sqrt{s}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \|\hat{\Sigma}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}}\|_\infty + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \sqrt{\frac{s\log p}{n}} \cdot \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2 \quad (24) \\
&\quad + \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2 \sqrt{\frac{s\log p}{n}} \cdot \|\boldsymbol{\beta}\|_2 + \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2 \sqrt{\frac{s\log p}{n}}.
\end{aligned}$$

Now since we have $|\frac{\hat{\Delta}^2 - \Delta^2}{\Delta^2}| = o(1)$ with probability at least $1 - 7p^{-1}$, this implies with probability at least $1 - 10p^{-1}$,

$$\|\hat{\Sigma}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\delta}}\|_\infty \leq \sqrt{\frac{\log p}{n}} \cdot \sqrt{\hat{\sigma}_{jj}\hat{\Delta}^2 + 2\hat{\sigma}_{jj}} \lesssim \Delta\sqrt{\frac{\log p}{n}}.$$

Then using the fact $|(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \Sigma (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})| \geq \lambda_{\min}(\Sigma)\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ again, we have with probability at least $1 - 10p^{-1}$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \lesssim \Delta\sqrt{\frac{s\log p}{n}} \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 + \sqrt{\frac{s\log p}{n}} \cdot \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2.$$

This implies that there exists some constant $C > 0$, such that with probability at least $1 - 10p^{-1}$,

$$\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 \leq C\Delta \cdot \sqrt{\frac{s\log p}{n}}.$$

In addition, since $\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}}\|_1 \leq \|\boldsymbol{\beta}\|_1 \leq \sqrt{p}\|\boldsymbol{\beta}\|_2 \leq \sqrt{pM} \cdot \Delta$, we then have

$$\begin{aligned}
&\mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2] \\
&\leq \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 \cdot 1_{\{\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 > C\Delta \cdot \sqrt{\frac{s\log p}{n}}\}}] + \mathbb{E}[\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 \cdot 1_{\{\|\hat{\boldsymbol{\beta}}_{\text{AdaLDA}} - \boldsymbol{\beta}\|_2 \leq C\Delta \cdot \sqrt{\frac{s\log p}{n}}\}}] \\
&\leq \sqrt{pM} \cdot \Delta \cdot 10p^{-1} + C\Delta \cdot \sqrt{\frac{s\log p}{n}} \lesssim \Delta \cdot \sqrt{\frac{s\log p}{n}} \lesssim M_{n,p} \cdot \sqrt{\frac{s\log p}{n}}.
\end{aligned}$$

### 6.2. Proofs of Theorem 3

For a vector $\boldsymbol{x} \in \mathbb{R}^p$, we define $\|\boldsymbol{x}\|_{2,s} = \sup_{\|\boldsymbol{y}\|_2=1, \boldsymbol{y} \in \Gamma(s)} |\boldsymbol{x}^\top \boldsymbol{y}|$. We start with the following lemma.

LEMMA 7. *For two vectors $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$, if $\|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_2 = o(1)$ as $n \to \infty$, and $\|\boldsymbol{\gamma}\|_2 \geq c$ for some constant $c > 0$, then when $n \to \infty$,*

$$\|\boldsymbol{\gamma}\|_2 \cdot \|\hat{\boldsymbol{\gamma}}\|_2 - \boldsymbol{\gamma}^\top \hat{\boldsymbol{\gamma}} \asymp \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_2^2.$$

We postpone the proof of Lemma 7 to Section A.6 in the supplement, and continue the proof of Theorem 3.

Let $\delta_n = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \vee \|\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\|_{2,s} \vee \|\hat{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2\|_{2,s}$. We are going to show

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-\Delta^2/8} \cdot \Delta \cdot \delta_n^2.$$

Given the estimators $\hat{\omega}, \hat{\boldsymbol{\mu}}_k$, and $\hat{\boldsymbol{\beta}}$, the sample $\boldsymbol{Z}$ is classified as

$$\hat{C}(\boldsymbol{Z}) = \begin{cases} 1, & (\boldsymbol{Z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)^\top \hat{\boldsymbol{\beta}} \geq 0 \\ 2, & (\boldsymbol{Z} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)^\top \hat{\boldsymbol{\beta}} < 0. \end{cases}$$

Let $\hat{\Delta} = \sqrt{\hat{\boldsymbol{\beta}}^\top \Sigma \hat{\boldsymbol{\beta}}}$ and $\hat{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}$. The misclassification error is

$$R_{\boldsymbol{\theta}}(\hat{C}) = \frac{1}{2}\Phi\Big(-\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Big) + \frac{1}{2}\bar{\Phi}\Big(-\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top \hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Big),$$

with $R_{\text{opt}}(\boldsymbol{\theta}) = \frac{1}{2}\Phi\big(-\Delta/2\big) + \frac{1}{2}\bar{\Phi}\big(\Delta/2\big)$. Define an intermediate quantity

$$R^* = \frac{1}{2}\Phi\Big(-\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\Big) + \frac{1}{2}\bar{\Phi}\Big(\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\Big).$$

We first show that $R^* - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-\Delta^2/8} \cdot \Delta \cdot \delta_n^2$. Applying Taylor's expansion to the two terms in $R^*$ at $\frac{\Delta}{2}$ and $-\frac{\Delta}{2}$ respectively, we obtain

$$R^* - R_{\text{opt}}(\boldsymbol{\theta}) = \frac{1}{2}\Big(\frac{\Delta}{2} - \frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}}\Big)\Phi'\Big(\frac{\Delta}{2}\Big) + \frac{1}{2}\Big(-\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} + \frac{\Delta}{2}\Big)\Phi'\Big(-\frac{\Delta}{2}\Big) + O\Big(e^{-\Delta^2/8}\frac{1}{\Delta} \cdot \delta_n^4\Big), \tag{25}$$

In fact, the remaining term can be written as

$$\frac{1}{2}\Big(\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\Delta}{2}\Big)^2 \Phi''(t_{1,n}) + \Big(\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\Delta}{2}\Big)^2 \Phi''(t_{2,n}),$$

where $t_{1,n}, t_{2,n}$ are some constants satisfying $|t_{1,n}|, |t_{2,n}|$ are between $\frac{\Delta}{2}$ and $\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}$.

Therefore, the remaining term can be bounded by using the facts that

$$\Big|\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}}{2\hat{\Delta}} - \frac{\Delta}{2}\Big| = O\big(\frac{1}{\Delta}\delta_n^2\big), \text{ and } \Phi''(t_n) = O(e^{-\Delta^2/8}\Delta),$$

for $|t_n|$ is between $\frac{\Delta}{2}$ and $\frac{\boldsymbol{\delta}^\top \hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}$.

In fact, for the first term, we can obtain this inequality by letting $\boldsymbol{\gamma} = \Sigma^{1/2}\boldsymbol{\beta}$ and $\hat{\boldsymbol{\gamma}} = \Sigma^{1/2}\hat{\boldsymbol{\beta}}$ in Lemma 7. Then

$$\left|\Delta - \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}}\right| = \left|\|\boldsymbol{\gamma}\|_2 - \frac{\boldsymbol{\gamma}^\top\hat{\boldsymbol{\gamma}}}{\|\hat{\boldsymbol{\gamma}}\|_2}\right| = \left|\frac{\|\boldsymbol{\gamma}\|_2\|\hat{\boldsymbol{\gamma}}\|_2 - \boldsymbol{\gamma}^\top\hat{\boldsymbol{\gamma}}}{\|\hat{\boldsymbol{\gamma}}\|_2}\right| \lesssim \frac{1}{\Delta}\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2 \lesssim \frac{1}{\Delta}\delta_n^2.$$

In addition, since as $\delta_n \to 0$, $\frac{(\boldsymbol{\delta}^*)^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}} \to \frac{\Delta}{2}$, we then have $|\Phi''(t_n)| \asymp \Delta \cdot e^{-\frac{(\Delta/2)^2}{2}} = \Delta \cdot e^{-\Delta^2/8}$. Then (25) can be further expanded such that

$$R^* - R_{\text{opt}}(\boldsymbol{\theta}) \asymp \left(-\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}}{2\hat{\Delta}} + \frac{\Delta}{2}\right)e^{-\frac{1}{2}\left(\frac{\Delta}{2}\right)^2} + \left(-\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}}{2\hat{\Delta}} + \frac{\Delta}{2}\right)e^{-\frac{1}{2}\left(-\frac{\Delta}{2}\right)^2} + O\left(e^{-\Delta^2/8}\frac{1}{\Delta} \cdot \delta_n^4\right)$$

$$= \exp\left(-\frac{\Delta^2}{8}\right) \cdot \left(-\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}} + \Delta\right) + O\left(e^{-\Delta^2/8}\frac{1}{\Delta} \cdot \delta_n^4\right)$$

$$\lesssim e^{-\Delta^2/8} \cdot \left|\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}} - \Delta\right| + O\left(e^{-\Delta^2/8}\frac{1}{\Delta} \cdot \delta_n^4\right) \lesssim e^{-\Delta^2/8} \cdot \delta_n^2.$$

Eventually we obtain $R^* - R_{\text{opt}}(\boldsymbol{\theta}) \lesssim e^{-\Delta^2/8}\Delta \cdot \delta_n^2$.

To upper bound $R_{\boldsymbol{\theta}}(\hat{C}) - R^*$, applying Taylor's expansion to $R_{\boldsymbol{\theta}}(\hat{C})$,

$$R_{\boldsymbol{\theta}}(\hat{C}) = \frac{1}{2}\left\{\Phi\left(\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right) + \frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top\hat{\boldsymbol{\beta}} - \boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\Phi'\left(\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right) + O\left(e^{-\Delta^2/8}\Delta \cdot \delta_n^2\right)\right\}$$

$$- \frac{1}{2}\left\{\bar{\Phi}\left(\frac{-\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right) + \frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top\hat{\boldsymbol{\beta}} + \boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\Phi'\left(-\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right) + O\left(e^{-\Delta^2/8}\Delta \cdot \delta_n^2\right)\right\},$$

where the remaining term can be obtained similarly as (25) by using the fact

$$\left|\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top\hat{\boldsymbol{\beta}} - \boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right| = O(\delta_n) \text{ and } |\Phi''(\cdot)| = O(e^{-\Delta^2/8}\Delta).$$

In fact, when $|\hat{\Delta} - \Delta| \leq |\sqrt{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\Sigma(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}| \lesssim \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \lesssim \delta_n = o(1)$, we have

$$\left|\frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top\hat{\boldsymbol{\beta}} - \boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right| \leq \frac{1}{2\Delta}|(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2)\hat{\boldsymbol{\beta}}| \lesssim \delta_n.$$

This leads to

$$|R_{\boldsymbol{\theta}}(\hat{C}) - R^*| \lesssim \left|\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Phi'\left(\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right)\right.$$

$$\left. + \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2 + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}}\Phi'\left(-\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right) + O\left(e^{-\Delta^2/8}\Delta \cdot \delta_n^2\right)\right|$$

$$= \left|\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}}e^{-\frac{1}{2}\left\{\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right\}^2}\right.$$

$$\left. + \frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2 + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2)^\top\hat{\boldsymbol{\beta}}}{\hat{\Delta}}e^{-\frac{1}{2}\left\{\frac{\boldsymbol{\delta}^\top\hat{\boldsymbol{\beta}}/2}{\hat{\Delta}}\right\}^2} + O\left(e^{-\Delta^2/8}\Delta \cdot \delta_n^2\right)\right|.$$

Since

$$\boldsymbol{\delta}/2 - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_1) + \boldsymbol{\delta}/2 + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_2) = \boldsymbol{\delta} - (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = 0,$$

then it follows that

$$|R_{\boldsymbol{\theta}}(\hat{C}) - R^*| \lesssim e^{-\Delta^2/8} \Delta \cdot \delta_n^2.$$

Combining the pieces, we obtain

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \lesssim e^{-\Delta^2/8} \cdot \Delta \cdot \delta_n^2.$$

Finally, by Lemma 5 and the derivation in Theorem 2, with probability at least $1 - 12p^{-1}$, $\delta_n \lesssim M_{n,p}\sqrt{\frac{s\log p}{n}}$. In addition, $\Delta \in [M_{n,p}, 3M_{n,p}]$, we then have with probability at least $1 - 12p^{-1}$,

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \lesssim e^{-M_{n,p}^2/8} \cdot M_{n,p}^3 \cdot \frac{s\log p}{n}.$$

Now we consider the two cases. On the one hand, when $M_{n,p}$ is bounded by $C_b$, we have

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \lesssim e^{-M_{n,p}^2/8} \cdot \frac{s\log p}{n}.$$

On the other hand, when $M_{n,p} \to \infty$ as $n$ grows,

$$R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \lesssim e^{-(\frac{1}{8} - \frac{3\log M_{n,p}}{M_{n,p}^2})M_{n,p}^2} \cdot M_{n,p}^3 \cdot \frac{s\log p}{n},$$

where $\frac{3\log M_{n,p}}{M_{n,p}^2}$ is an $o(1)$ term as $n \to \infty$.

### 6.3. Proofs of Theorems 4 and 5

We proceed to proving Theorems 4 and 5 under the event $\{c_1 n_0 \leq n_{\min}^*(S) \leq c_2 n_0\}$ that happens with probability at least $1 - p^{-1}$. The results then rely on the following lemma.

LEMMA 8. *Consider the MCR model and assume that* $\hat{\boldsymbol{\mu}}, \hat{\Sigma}$ *are the generalized sample mean and sample covariance matrix respectively. If* $c_1 n_0 \leq n_{\min}^*(S) \leq c_2 n_0$. *then with probability at least* $1 - p^{-1}$,

$$\sup_{\boldsymbol{u} \in \Gamma(s)} \boldsymbol{u}^\top (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \lesssim \sqrt{\frac{s\log p}{n_0}};$$

$$\sup_{\boldsymbol{u}, \boldsymbol{v} \in \Gamma(s)} \boldsymbol{u}^\top (\hat{\Sigma} - \Sigma)\boldsymbol{v} \lesssim \sqrt{\frac{s\log p}{n_0}}.$$

Given Lemma 8, the derivation of Theorems 4 is very similar to the case with AdaLDA in Section 6.1, and 5 can be derived from Theorem 4 by using the same logic as in Section 6.2, and thus are omitted.

### 6.4. Proofs of the minimax lower bound results (Theorems 6 and 7)

In this section we are going to prove Theorems 6 and 7. We start with providing lemmas that will be used in the proof.

*6.4.1.  Auxiliary lemmas*

The proof of Theorem 6 relies on the following Fano's Lemma.

LEMMA 9 (TSYBAKOV [2009]). *Suppose $\Theta_p$ is a parameter space consisting of $M$ parameters $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M \in \Theta_p$ for some $M > 0$, and $d(\cdot, \cdot) : \Theta_p \times \Theta_p \to \mathbb{R}^+$ is some distance. Denote $\mathbb{P}_{\boldsymbol{\theta}}$ to be some probability measure parametrized by $\boldsymbol{\theta}$. If for some constants $\alpha \in (0, 1/8), \gamma > 0$, $KL(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_{\boldsymbol{\theta}_0}) \leq \alpha \log M / n$ for all $1 \leq i \leq M$, and $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) \geq \gamma$ for all $0 \leq i \neq j \leq M$, then*

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{i \in [M]} \mathbb{E}_{\boldsymbol{\theta}_i}[d_{\boldsymbol{\theta}_i}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_i)] \gtrsim \gamma.$$

The proof of Theorem 7, however, is not straightforward, since the excess risk $R_{\boldsymbol{\theta}}(\hat{C}) - R_{opt}(\boldsymbol{\theta})$ is not a distance as required in Lemma 9. The key step in our proof of Theorem 7 is to reduce the excess risk $R_{\boldsymbol{\theta}}(\hat{C}) - R_{opt}(\boldsymbol{\theta})$ to $L_{\boldsymbol{\theta}}(\hat{C})$, defined in (16). The following lemma suggests that it suffices to provide a lower bound for $L_{\boldsymbol{\theta}}(\hat{C})$, and $L_{\boldsymbol{\theta}}(\hat{C})$ satisfies an approximate triangle inequality (Lemma 4).

Although $L_{\boldsymbol{\theta}}(\hat{C})$ is not a distance function and does not satisfy an exact triangle inequality, the following lemma provides a variant of Fano's Lemma.

LEMMA 10 (TSYBAKOV [2009]). *Let $M \geq 0$ and $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M \in \Theta_p$. For some constants $\alpha_0 \in (0, 1/8], \gamma > 0$, and any classifier $\hat{C}$, if $KL(\mathbb{P}_{\boldsymbol{\theta}_i}, \mathbb{P}_{\boldsymbol{\theta}_0}) \leq \alpha_0 \log M / n$ for all $1 \leq i \leq M$, and $L_{\boldsymbol{\theta}_i}(\hat{C}) < \gamma$ implies $L_{\boldsymbol{\theta}_j}(\hat{C}) \geq \gamma$ for all $0 \leq i \neq j \leq M$, then*

$$\inf_{\hat{C}} \sup_{i \in [M]} \mathbb{P}_{\boldsymbol{\theta}_i}(L_{\boldsymbol{\theta}_i}(\hat{C}) \geq \gamma) \geq \frac{\sqrt{M}}{\sqrt{M}+1}\left(1 - 2\alpha_0 - \sqrt{\frac{2\alpha_0}{\log M}}\right).$$

LEMMA 11 (TSYBAKOV [2009]). *Define $\mathcal{A}_{p,s} = \{\boldsymbol{u} : \boldsymbol{u} \in \{0, 1\}^p, \|\boldsymbol{u}\|_0 \leq s\}$. If $p \geq 4s$, then there exists a subset $\{\boldsymbol{u}_0, \boldsymbol{u}_1, ..., \boldsymbol{u}_M\} \subset \mathcal{A}_{p,s}$ such that $\boldsymbol{u}_0 = \{0, ..., 0\}^\top$, $\rho_H(\boldsymbol{u}_i, \boldsymbol{u}_j) \geq s/2$ and $\log(M+1) \geq \frac{s}{5}\log(\frac{p}{s})$, where $\rho_H$ denotes the Hamming distance.*

*6.4.2.  Proof of Theorem 6*

In this section we prove the lower bound of estimation of $\boldsymbol{\beta}$. First we construct a subset of the parameter space $\Theta$ that characterizes the hardness of the problem. By Lemma 11, there exist $\boldsymbol{u}_0, \boldsymbol{u}_1, ..., \boldsymbol{u}_M \in \mathcal{A}_{p,s} = \{\boldsymbol{u} \in \{0, 1\}^p : \|\boldsymbol{u}\|_0 \leq s\}$, such that $\rho_H(\boldsymbol{u}_i, \boldsymbol{u}_j) > s/2$ and $\log(M+1) \geq \frac{s}{5}\log(\frac{p}{s})$, denote this collection of $\boldsymbol{u}_i$ by $\tilde{\mathcal{A}}_{p,s}$. In addition, denote $\boldsymbol{u}_0 = \boldsymbol{0}_p$.

Since $\frac{\log p}{\log(p/s)} = O(1)$, so for sufficiently large $p$, we have $s < p/2$. Define $\boldsymbol{b}_0$ be the $p$-dimensional vector with the last $s$ entries being $\frac{M_{n,p}}{\sqrt{s}}$ and the rest being 0, so we have $\|\boldsymbol{b}_0\|_2 = M_{n,p}$. Let $r = \lceil p/2 \rceil$. For $\boldsymbol{u} \in \tilde{\mathcal{A}}_{p,s} = \{\boldsymbol{u}_0, \boldsymbol{u}_1, ..., \boldsymbol{u}_M\}$, let $B_{\boldsymbol{u}}$ be the $p \times p$ symmetric matrix whose $i$-th row and column are both $\epsilon \cdot u_i \cdot \frac{\boldsymbol{b}_0}{M_{n,p}}$ for $i \in \{1, ..., r\}$, where $\epsilon$ is to be determined later. The parameter set we considered is

$$\Theta_0 = \{\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1 = \boldsymbol{b}_0, \boldsymbol{\mu}_2 = -\boldsymbol{b}_0, \Sigma = (I_p + B_{\boldsymbol{u}})^{-1}; \boldsymbol{u} \in \tilde{\mathcal{A}}_{p,s} \cup \{\boldsymbol{0}_p\}\}.$$

For a given $\boldsymbol{u}$, the corresponding discriminating direction is $\boldsymbol{\beta}_{\boldsymbol{u}} = -2(I_p + B_{\boldsymbol{u}})\boldsymbol{b}_0$, which implies

$$\|\boldsymbol{\beta}_{\boldsymbol{u}} - \boldsymbol{\beta}_{\tilde{\boldsymbol{u}}}\|_2^2 = 4\|(B_{\boldsymbol{u}} - B_{\tilde{\boldsymbol{u}}})\boldsymbol{b}_0\|_2^2 \geq 4\rho_H(\boldsymbol{u}, \tilde{\boldsymbol{u}})\epsilon^2 \|\boldsymbol{b}_0\|_2^2 \geq 2s M_{n,p}^2 \epsilon^2.$$

In addition, when $\|B_{\boldsymbol{u}}\|_2 = o(1)$, for sufficiently large $n$, we have $\Delta = \sqrt{4\boldsymbol{b}_0^\top (I_p + B_{\boldsymbol{u}})\boldsymbol{b}_0} \in (M_{n,p}, 3M_{n,p})$, which implies that $\Theta_0 \subset \mathcal{G}(s, M_{n,p})$.

We then proceed to bound $KL(\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_0}})$ for $i \in [M]$, where $\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_0}}$ denote the distributions $N_p(\boldsymbol{b}_0, (I_p + B_{\boldsymbol{u}_i})^{-1})$ and $N_p(\boldsymbol{b}_0, I_p)$ respectively. We then have

$$KL(\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_0}}) = \frac{1}{2}\left[-\log|I_p + B_{\boldsymbol{u}_i}| - p + \mathrm{tr}(I_p + B_{\boldsymbol{u}_i}))\right].$$

Note that $\frac{\boldsymbol{b}_0}{M_{n,p}}$ is a unit vector. If we take $\epsilon$ such that $\|B_{\boldsymbol{u}_i}\|_2 \le \|B_{\boldsymbol{u}}\|_F \le \sqrt{2s \cdot \epsilon^2} = o(1)$, and denote the eigenvalues of $I_p + B_{\boldsymbol{u}_i}$ by $1 + \Delta_{\lambda_1},...,1 + \Delta_{\lambda_p}$ with $\Delta_{\lambda_j} = o(1)$. We then have

$$\begin{aligned}
KL(\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_0}}) =& \frac{1}{2}\left[-\sum_{j=1}^p \log(1 + \Delta_{\lambda_j}) - p + \sum_{j=1}^p (1 + \Delta_{\lambda_j})\right] \\
\asymp& \frac{1}{4}\sum_{j=1}^p \Delta_{\lambda_j}^2 = \frac{1}{4}\|B_{\boldsymbol{u}}\|_F^2 \le \frac{1}{2}s\epsilon^2
\end{aligned}$$

where we use the fact that $\log(1 + x) \asymp x - \frac{x^2}{2}$ when $x = o(1)$. Now let $\epsilon = \frac{1}{5\sqrt{2}}\sqrt{\frac{\log p}{n}}$, then $KL(\mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_i}}, \mathbb{P}_{\boldsymbol{\theta}_{\boldsymbol{u}_0}}) \le \alpha \log M/n$ for $\alpha = 1/8$.

In addition, let $\gamma = \frac{1}{10}M_{n,p}\sqrt{\frac{s\log p}{n}}$, then for $0 \le i \ne j \le M$ and any $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$, such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\boldsymbol{u}_i}\|_2 \le \gamma$, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\boldsymbol{u}_j}\|_2 \ge \|\boldsymbol{\beta}_{\boldsymbol{u}_j} - \boldsymbol{\beta}_{\boldsymbol{u}_j}\|_2 - \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\boldsymbol{u}_i}\|_2 \ge \frac{1}{5}M_{n,p}\sqrt{\frac{s\log p}{n}} - \frac{1}{10}M_{n,p}\sqrt{\frac{s\log p}{n}} = \frac{1}{10}M_{n,p}\sqrt{\frac{s\log p}{n}} = \gamma.$$

Then by Fano's lemma (Lemma 9), we have $\inf_{\hat{\boldsymbol{\beta}}} \sup_{i \in [M]} \mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\boldsymbol{u}_i}\|_2 \gtrsim M_{n,p}\sqrt{\frac{s\log p}{n}}$.

For the incomplete data case with $n_0 \ge 1$, we consider a special pattern of missingness $S_0$:

$$(S_0)_{ij} = 1_{\{1 \le i \le n_0, 1 \le j \le p\}} \quad \text{with probability 1.}$$

Under this missingness pattern, $n_{\min}^* = n_0$ with probability 1, and the problem essentially becomes complete data problem with $n_0$ samples, which implies

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \boldsymbol{S} \in \boldsymbol{\Psi}(n_0;n,p)}} \mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2] \gtrsim M_{n,p}\sqrt{\frac{s\log p}{n_0}}.$$

### 6.4.3. Proof of Theorem 7

We proceed by applying Lemma 10 to obtain the minimax lower bound for the excess misclassification error. We first construct a subset of the parameter space $\Theta$ that characterizes the hardness of the problem. Let $\boldsymbol{e}_1$ be the basis vector in the standard Euclidean space whose first entry is 1 and zero elsewhere. By Lemma 11, there exist $\boldsymbol{u}_1, ..., \boldsymbol{u}_M \in \check{\mathcal{A}}_{p,s} = \{\boldsymbol{u} \in \{0,1\}^p : , \boldsymbol{u}^\top \boldsymbol{e}_1 = 0, \|\boldsymbol{u}\|_0 = s\}$, such that $\rho_H(\boldsymbol{u}_i, \boldsymbol{u}_j) > s/2$ and $\log(M + 1) \ge \frac{s}{5}\log(\frac{p-1}{s})$. Note the first entry in $\boldsymbol{u}_j$ is 0 for all $j = 1, \ldots, M$.

Define the parameter space

$$\Theta_1 = \{\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) : \boldsymbol{\mu}_1 = \epsilon\boldsymbol{u} + \lambda\boldsymbol{e}_1, \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1, \Sigma = \sigma^2 I_p; \boldsymbol{u} \in \check{\mathcal{A}}_{p,s}\},$$

where $\epsilon = \sigma\sqrt{\log p/n}$, $\sigma^2 = O(1)$ and $\lambda$ is chosen to ensure $\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})$ such that

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{4\|\epsilon\boldsymbol{u} + \lambda\boldsymbol{e}_1\|_2^2}{\sigma^2} = M_{n,p}.$$

To apply Lemma 10, we need to verify two conditions: (i) the upper bound on the KL divergence between $\mathbb{P}_{\boldsymbol{\theta}_u}$ and $\mathbb{P}_{\boldsymbol{\theta}_v}$, and (ii) the lower bound of $L_{\boldsymbol{\theta}_u}(\hat{C}) + L_{\boldsymbol{\theta}_v}(\hat{C})$ for $\boldsymbol{u} \neq \boldsymbol{v} \in \check{\mathcal{A}}_{p,s}$.

We calculate the KL divergence first. For $\boldsymbol{u} \in \check{\mathcal{A}}_{p,s}$, denote $\boldsymbol{\mu}_u = \epsilon \boldsymbol{u} + \lambda \boldsymbol{e}_1$. For $\boldsymbol{\theta}_u = (\boldsymbol{\mu}_u, -\boldsymbol{\mu}_u, \sigma^2 I_p) \in \Theta_1$, we consider the distribution $N_p(\boldsymbol{\mu}_u, \sigma^2 I_p)$.

Then, the KL divergence between $\mathbb{P}_{\boldsymbol{\theta}_u}$ and $\mathbb{P}_{\boldsymbol{\theta}_v}$ can be bounded by

$$\mathrm{KL}(\mathbb{P}_{\boldsymbol{\theta}_u}, \mathbb{P}_{\boldsymbol{\theta}_v}) \leq \frac{1}{2}\|\boldsymbol{\mu}_u - \boldsymbol{\mu}_v\|_2^2 \leq \sigma^2 \cdot \frac{s \log p}{n}. \tag{26}$$

In addition, by applying Lemma 4, we have that for any $\boldsymbol{u}, \boldsymbol{v} \in \check{\mathcal{A}}_{p,s}$,

$$L_{\boldsymbol{\theta}_u}(\hat{C}) + L_{\boldsymbol{\theta}_v}(\hat{C}) \gtrsim \frac{1}{M_{n,p}} e^{-M_{n,p}^2/8} \sqrt{\frac{s \log p}{n}}.$$

So far we have verified the aforementioned conditions (i) and (ii). Lemma 10 immediately implies that, there is some $C_\alpha \geq 0$, such that

$$\inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}(L_{\boldsymbol{\theta}}(\hat{C}) \geq C_\alpha \frac{1}{M_{n,p}} e^{-M_{n,p}^2/8} \sqrt{\frac{s \log p}{n}}) \geq 1 - \alpha. \tag{27}$$

Finally combining (27) with Lemma 3, we obtain the desired lower bound for the excess mis-classficiation error

$$\inf_{\hat{C}} \sup_{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p})} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \geq C_\alpha \frac{1}{M_{n,p}} e^{-M_{n,p}^2/8} \frac{s \log p}{n}) \geq 1 - \alpha.$$

Under this missingness data case, we consider the same missingness pattern $S_0$ as described in Section 6.4.2 with $n_{\min}^* = n_0$. Then we have

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \boldsymbol{S} \in \boldsymbol{\Psi}(n_0; n, p)}} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \geq C_\alpha \frac{1}{M_{n,p}} e^{-M_{n,p}^2/8} \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

This implies that

(a) If $M_{n,p} \leq C_b$ for some $C_b > 0$, then

$$\inf_{\hat{C}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \boldsymbol{\Psi}(n_0; n, p)}} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \geq C_\alpha e^{-\frac{1}{8}M_{n,p}^2} \cdot \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

(b) If $M_{n,p} \to \infty$ as $n \to \infty$, then for any $\delta > 0$,

$$\inf_{\hat{C}} \sup_{\substack{\boldsymbol{\theta} \in \mathcal{G}(s, M_{n,p}) \\ \mathcal{F} \in \boldsymbol{\Psi}(n_0; n, p)}} \mathbb{P}(R_{\boldsymbol{\theta}}(\hat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \geq C_\alpha e^{-(\frac{1}{8}+\delta)M_{n,p}^2} \cdot \frac{s \log p}{n_0}) \geq 1 - \alpha.$$

# References

Theodore Wilbur Anderson. *An Introduction To Multivariate Statistical Analysis*. Wiley-Intersceince, 3rd ed, New York, 2003.

Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, pages 2139–2147, 2013.

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

T Tony Cai and Ming Yuan. Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40(4):2014–2042, 2012.

T. Tony Cai and Anru Zhang. Estimation of high-dimensional covariance matrices with incomplete data. *Journal of Multivariate Analysis*, 150:55–74, 2016.

T. Tony Cai and Linjun Zhang. Supplement to "High-dimensional linear discriminant analysis: Optimality, adaptive algorithm, and missing data". 2018.

T Tony Cai, Weidong Liu, and Harrison H Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016.

T Tony Cai, Jing Ma, and Linjun Zhang. CHIME: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, to appear, 2018.

Richard Y Chen, Alex Gittens, and Joel A Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20, 2012.

Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Ann. Statist.*, 36(6):2605–2637, 2008.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

John W Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576, 2009.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning 2nd edition*. New York: Springer, 2009.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Elizaveta Levina and Roman Vershynin. Partial estimation of covariance matrices. *Probab. Theory Relat. Fields*, 153(3):405–419, 2012.

Tianyang Li, Xinyang Yi, Constantine Carmanis, and Pradeep Ravikumar. Minimax gaussian classification & clustering. In *Artificial Intelligence and Statistics*, pages 1–9, 2017.

Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, 2015.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Ann. Statist.*, 40:1637–1664, 2012.

Karim Lounici. Sparse principal component analysis with missing observations. *High dimensional probability VI*, 66 of Prog. Proba., IMS Collections:327–356, 2013.

Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

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

Qing Mai, Yi Yang, and Hui Zou. Multiclass sparse discriminant analysis. *Statistica Sinica*, 29: 97–111, 2019.

Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516): 1673–1683, 2016.

Milind Rao, Tara Javidi, Yonina C Eldar, and Andrea Goldsmith. Fundamental estimation limits in autoregressive processes with compressive measurements. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 2895–2899. IEEE, 2017.

James M Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997.

Andrea Rotnitzky and James Robins. Analysis of semi-parametric regression models with nonignorable non-response. *Statistics in Medicine*, 16(1):81–102, 1997.

Tapio Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14(5):853–871, 2001.

Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265, 2011.

BaoLuo Sun, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric Tchetgen Tchetgen. Semiparametric estimation with data missing not at random using an instrumental variable. *arXiv preprint arXiv:1607.03197*, 2016.

Eric J Tchetgen Tchetgen and Kathleen E Wirth. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73(4):1123–1131, 2017.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.

Daniela M Witten and Robert Tibshirani. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):615–636, 2009.

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

Michael C Wu, Lingsong Zhang, Zhaoxi Wang, David C Christiani, and Xihong Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.