

**EFFICIENT ROBUST ALGORITHMS FOR LINEAR DISCRIMINANT  
ANALYSIS AND SEQUENTIAL MATCHING PROBLEMS**

A Dissertation  
Presented to  
The Academic Faculty

By

Yuyang Shi

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy in the  
College of Engineering  
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology

August 2023

© Yuyang Shi 2023

**EFFICIENT ROBUST ALGORITHMS FOR LINEAR DISCRIMINANT  
ANALYSIS AND SEQUENTIAL MATCHING PROBLEMS**

Thesis committee:

Dr. Yajun Mei, Advisor  
School of Industrial and Systems Engineering  
*Georgia Institute of Technology*

Dr. David Goldsman  
School of Industrial and Systems Engineering  
*Georgia Institute of Technology*

Dr. Tuo Zhao  
School of Industrial and Systems Engineering  
*Georgia Institute of Technology*

Dr. Ethan X. Fang  
Department of Biostatistics & Bioinformatics  
*Duke University*

Dr. Enlu Zhou  
School of Industrial and Systems Engineering  
*Georgia Institute of Technology*

Dr. Ofer Sadan  
School of Medicine  
*Emory University*

Date approved: May 9 2023

For my dear family

## ACKNOWLEDGMENTS

First and foremost, I want to express my deep gratitude to my advisor, Dr. Yajun Mei, who has given invaluable support and guidance throughout my research. His mentorship has not only contributed to my academic and professional growth but has also taught me important life skills, such as effective communication and presentation. Moreover, his unwavering mental support has encouraged me to persevere, especially during the difficult time of Covid-19, when I once felt depressed and needed direction. Without him, I could not complete my PhD.

I would also like to thank Dr. Tuo Zhao for his exceptional guidance in the initial years of my PhD, providing me with opportunities to learn new topics in statistics and machine learning in a reading group. Additionally, I am grateful for his support in helping me accomplish my first PhD research paper, which boosted my confidence to further explore interesting topics in the field.

I am also thankful for other members in my thesis committee, Dr. Ethan X. Fang, Dr. David Goldsman, Dr. Ofer Sadan and Dr. Enlu Zhou, for their service and help. Dr. Ethan X. Fang provided me much guidance on my research. Dr. David Goldsman agreed to serve on my thesis committee board at last minute. Dr. Ofer Sadan offered me excellent opportunities to collaborate on biomedical projects, from which I was able learn more about applying statistical methods on real-world problems and data. Dr. Enlu Zhou offered very interesting and useful courses on simulation theory, which extended my research ability.

I have been fortunate to interact with a lot of talented peer graduate students during my PhD journey. Specifically, I would like to thank Zhehui Chen and Haoming Jiang for our collaboration. In addition, I want to thank Minshuo Chen, Shixiang Zhu, Yan Li, Xinyu Liu, Enpeng Yuan, Minghe Zhang, Haoran Sun, Leon Chen among others for sharing broad knowledge and insights that benefitted me in many aspects.

Special thanks are due to people outside of my dissertation committee that have helped

me over the years. I appreciate the support from Dr. Justin Lo, Bala Rangaratnam and Dr. Juwon Lee during my internship at Chevron and Amazon. Also, I want to thank Dr. Feras Akbik, Dr. Venugopal Veeravalli and Aditya Deshmukh for our collaborations.

I gratefully acknowledge the support by M.H. Stewart Fellowship and McLean Fellowship offered from H. Milton Stewart School of Industrial and Systems Engineering that supported me from 2018 to 2019. Also, I appreciate the support of the GT Covid Disruption GRA funding that supported me during the time of Covid-19.

Finally, I am thankful for my family, who have been supporting me throughout my life, and building up my courage and confidence to explore the unknown.

## TABLE OF CONTENTS

<b>Acknowledgments</b> . . . . .	iv
<b>List of Tables</b> . . . . .	x
<b>List of Figures</b> . . . . .	xii
<b>Summary</b> . . . . .	xiii
<b>Chapter 1: Pivotal Estimation of Linear Discriminant Analysis in High Dimensions</b> . . . . .	1
1.1 Introduction . . . . .	1
1.2 Background . . . . .	4
1.2.1 Problem Setup . . . . .	4
1.2.2 The AdaLDA method . . . . .	6
1.2.3 Review of Gautier’s method . . . . .	7
1.3 The PANDA Method . . . . .	8
1.3.1 An ADMM Algorithm for PANDA . . . . .	10
1.4 Statistical Properties . . . . .	12
1.5 Numerical Results . . . . .	16
1.5.1 Simulation . . . . .	16
1.5.2 Leukemia data . . . . .	24

1.6	Extension to multiple-class LDA . . . . .	25
1.7	Proofs of the Main Results . . . . .	27
1.7.1	Proof of Theorem 1 . . . . .	27
1.7.2	Proof of Theorem 2 . . . . .	32
1.7.3	Proof of Theorem 3 . . . . .	34
1.8	Proof of Lemmas . . . . .	36
1.8.1	Proofs . . . . .	36
1.9	Discussions . . . . .	49
<b>Chapter 2: Robust High-Dimensional LDA under Data Contamination . . . . .</b>		<b>51</b>
2.1	Introduction . . . . .	51
2.2	Problem Overview . . . . .	53
2.2.1	LDA . . . . .	53
2.2.2	Data contamination model . . . . .	54
2.2.3	Objective . . . . .	55
2.3	Our Proposed Method . . . . .	55
2.4	Theoretical Results . . . . .	59
2.5	Numerical Study . . . . .	61
2.5.1	Simulation . . . . .	62
2.5.2	MNIST data example . . . . .	65
2.6	Proofs . . . . .	66
2.6.1	Proof of Theorem 4 . . . . .	66
2.6.2	Proof of Theorem 5 . . . . .	70

2.6.3	Proof of lemmas . . . . .	73
<b>Chapter 3:</b>	<b>An Efficient Algorithm for Sequential Matching Problems with Un-</b>	
	<b>known Utility . . . . .</b>	<b>84</b>
3.1	Introduction . . . . .	84
3.2	Problem Formulation and Background . . . . .	86
3.2.1	Optimal assignment with known utilities . . . . .	89
3.3	Our Proposed Method . . . . .	91
3.3.1	Constructing upper confidence bound . . . . .	91
3.3.2	Assignment by maximizing upper confidence bound . . . . .	92
3.4	Theoretical Results . . . . .	94
3.5	Numerical Studies . . . . .	96
3.5.1	Settings . . . . .	97
3.5.2	Performance . . . . .	97
3.5.3	Comparison with $\epsilon$ -greedy method . . . . .	100
3.6	Further Study . . . . .	100
3.6.1	Data format and settings . . . . .	101
3.6.2	Performance . . . . .	102
3.7	Proofs . . . . .	104
3.7.1	Proof of Theorem 7 . . . . .	104
3.7.2	Proof of Theorem 8 . . . . .	108
3.7.3	Proof of Lemma 3.7.4 . . . . .	112
<b>Chapter 4:</b>	<b>Conclusion and Future Research . . . . .</b>	<b>114</b>

4.1	Summary of Contribution . . . . .	114
4.2	Future Research . . . . .	115
	<b>References . . . . .</b>	<b>117</b>

## LIST OF TABLES

1.1	Misclassification rate of the PANDA method under the AR(1) model with $n = 200$ , $p = 400$ , $s = 5$ and different $c$ , averaged over 50 replications. The standard deviations are provided in brackets. . . . .	20
1.2	<i>The <math>\ell_2</math> norm of estimation errors under each setting, averaged over 100 instances. The standard deviations are given in brackets. The lower value at the significance level 0.05 between the AdaLDA and the PANDA method are marked in bold.</i> . . . . .	21
1.3	<i>The classification risk under each setting averaged over 100 instances. The standard deviations are given in brackets. The lower value at the significance level 0.05 between the AdaLDA and the PANDA method are marked in bold.</i> . . . . .	23
1.4	Running time (in seconds) of the PANDA and AdaLDA methods under the Varying Diagonal model using Gurobi. The standard deviations are given in brackets. . . . .	24
1.5	<i>The performance of PANDA, AdaLDA and LPD on the Leukemia dataset. The testing errors are averaged over 100 repetitions. The standard deviation of the testing errors are given in brackets. The difference between PANDA and the other two methods is significant by pair-wise t-test with a p-value less than 0.001.</i> . . . . .	25
2.1	$\ell_2$ error for $\mu_m$ estimate with empirical and robust methods averaged over 100 replications. The standard deviations are presented in brackets. . . . .	63
2.2	$\ell_2$ error for $\beta^*$ estimate averaged over 100 replications. The standard deviations are presented in brackets. . . . .	64
2.3	Misclassification Rate averaged over 100 replications. The standard deviations are presented in brackets. . . . .	65
2.4	Misclassificate rates over the testing data. . . . .	66

3.1	Total regret up to $T = 20$ under setting (a) . . . . .	98
3.2	Total regret up to $T = 20$ under setting (b) . . . . .	98
3.3	Feature variables for mentees . . . . .	101
3.4	Total regret up to $T = 20$ for the simulated data for mentor-mentee matching scenario averaged over 50 repetitions. The standard deviations are presented in brackets. . . . .	103

## LIST OF FIGURES

1.1	<i>The misclassification rate v.s. the parameter <math>\tilde{\lambda}</math>. Results are averaged over 100 instances.</i> . . . . .	18
1.2	<i><math>\ell_2</math> estimation error v.s. values parameter <math>\tilde{\lambda}</math>. Results are averaged over 100 instances.</i> . . . . .	18
1.3	<i>The distribution of the empirically optimal tuning parameter <math>\tilde{\lambda}</math> for LPD (left), AdaLDA (middle) and PANDA (right), approximated with kernel smoothing. The optimal choice of the tuning parameter in our method relies less on the population.</i> . . . . .	19
3.1	Cumulative regret with different $d$ and $\lambda$ under settings (i) and (ii). The curves of cumulative regret are sub-linear in $t$ . With proper choice of $\lambda$ , the performance is improved over the pure-greedy algorithm when $\lambda = 0$ . . . . .	99
3.2	Cumulative regret for $\epsilon$ -greedy method and our UCB-based method with $\lambda = 0.5$ under settings (i) with $d = 10, 100$ . The performance of $\epsilon$ -greedy method is worse than the pure-greedy method under this setting. . . . .	100
3.3	Cumulative regret with different $\lambda$ for the mentor-mentee matching scenario. With a proper $\lambda$ value (e.g. 20), our method has significant improvement in the total utility compared to the greedy method. . . . .	103

## SUMMARY

Data science, machine learning, or statistics are useful to assist making data-driven decisions in many modern applications, and some challenges rise in analyzing real-world data sets such as high-dimensionality, robustness, and computational efficiency. This dissertation investigates three specific topics in statistical machine learning: (1) pivotal method for high-dimensional linear discriminant analysis (LDA); (2) robust algorithm for LDA under data contamination; and (3) efficient algorithm for sequential assignment with unknown utility.

In Chapter 1, we propose a pivotal method for high-dimensional linear discriminant analysis (LDA) that enjoys tuning-insensitive property. We term our method as PivotAl LiNear Discriminant Analysis (PANDA). Our method conducts parameter estimation under a pivotal estimation framework and only needs to solve a single convex optimization problem when both means and variances are unknown for both classes of training data. Theoretically, our method achieves comparable convergence rates as existing methods in terms of both estimation error and misclassification rate.

In Chapter 2, we propose a computationally efficient algorithm for robust LDA under data contamination, where a fraction of sample data might be corrupted by some adversary. Our main ideas are as follows. We first identify the outliers in each class and robustly estimate the mean, and then apply our developed PANDA method for uncontaminated data to estimate the discriminant direction in LDA with data contamination. Theoretical properties of the proposed algorithm are established in terms of both the error in estimating the optimal projection vector and the misclassification rate.

In Chapter 3, we develop an efficient algorithm for sequential assignment with unknown utility, with the objective of nearly maximizing the overall utility for each time. Our proposed algorithm is to use stochastic binary bandit feedback to adaptively estimate the unknown utilities through the logistic regression, and then to combine the Upper Con-

confidence Bound (UCB) algorithm in the multi-armed bandit problem with the Hungarian algorithm in the assignment problem. We derive the theoretical bounds of our algorithm for both the estimation error and the total regret, and numerical studies are also conducted to illustrate the usefulness of our algorithm.

We conclude the dissertation in Chapter 4, where we summarize our contributions, and highlight several potential research topics for future investigation.

**CHAPTER 1**  
**PIVOTAL ESTIMATION OF LINEAR DISCRIMINANT ANALYSIS IN HIGH**  
**DIMENSIONS**

**1.1 Introduction**

We consider the classical problem of linear discriminant analysis (LDA). Suppose that we have samples  $\{X_i^{(\ell)} : \ell = 0, 1; i = 1, 2, \dots, n_\ell\}$  from class  $\ell$ , where  $\ell = 0$  or  $1$ , and each  $X_i^{(\ell)} \in \mathbb{R}^p$  is a sample from class  $\ell$ . We further assume that the samples from each class  $\ell$  are independently drawn from a multivariate Gaussian distribution  $X^{(\ell)} \sim N(\mu^{(\ell)}, \Sigma)$ , where the means of the two classes are different, and the two classes share the same covariance matrix  $\Sigma$ . The goal of LDA is to find a linear discriminant rule  $f$  to classify new samples from these two classes with equal prior weights, and we aim to achieve the optimal misclassification rate.

It is well known in literature [1] that when we have a new sample  $Z$ , the Bayes rule classifies by

$$f^*(Z) = \mathbb{1}\{(Z - \mu_m)^\top \Sigma^{-1} \mu_d > 0\},$$

where  $\mu_m = (\mu^{(0)} + \mu^{(1)})/2$  and  $\mu_d = (\mu^{(1)} - \mu^{(0)})$ . Under the classical low-dimensional setting, where the dimension  $p$  is small, and sample sizes  $n_0, n_1$  are large, we estimate  $\mu^{(0)}, \mu^{(1)}$  and  $\Sigma^{-1}$  by their sample versions, and use the plug-in Bayes rule to classify the new sample. In particular, let  $\hat{\mu}^{(\ell)}$ 's and  $\hat{\Sigma}$  be the the sample means and the pooled sample covariance matrix, and let  $\hat{\mu}_m = (\hat{\mu}^{(0)} + \hat{\mu}^{(1)})/2, \hat{\mu}_d = (\hat{\mu}^{(1)} - \hat{\mu}^{(0)})$ . Given a new sample

$Z$ , the following discriminant rule  $\hat{f}$  achieves the optimal Bayesian risk asymptotically that

$$\hat{f}(Z) = \begin{cases} 0, & \text{if } \hat{\mu}_d^\top \hat{\Sigma}^{-1} (Z - \hat{\mu}_m) \leq 0, \\ 1, & \text{if } \hat{\mu}_d^\top \hat{\Sigma}^{-1} (Z - \hat{\mu}_m) > 0. \end{cases}$$

Unfortunately, this method is inapplicable to high-dimensional settings. This is because it is difficult to estimate  $\Sigma^{-1}$  efficiently when the data dimension  $p$  is high, and the pooled sample covariance matrix  $\hat{\Sigma}$  is even singular when  $p$  is greater than the sample size  $n_0 + n_1$ . On the other hand, such high dimensionality issues unavoidably exist in many critical modern applications such as genomics and fMRI decoding. Thus, it is important to develop efficient methods for LDA in high dimensions.

Several methods have been developed in the literature for high-dimensional LDA with sparsity assumptions imposed. Note that in many real-world applications, such as the fMRI decoding and biomarker identification [2, 3], despite the high dimensionality issue, practitioners believe that the classification problems depend only on a small number of predictors among the many given. Under such motivation, two tracks of work in the literature impose different sparsity assumptions. The first one is to estimate  $\mu_d = \mu^{(1)} - \mu^{(0)}$  and  $\Sigma$  separately, under the assumption that both  $\mu_d$  and  $\Sigma$  are sparse. A simple approach is the naive Bayes rule or independence rule discussed in [4], where the authors assume that  $\Sigma$  is diagonal. [5], and [6] propose the nearest shrunken centroid method and the Features Annealed Independence Rules (FAIR) respectively, for selecting significant features under the assumption that  $\mu_d$  is sparse, and  $\Sigma$  is a diagonal matrix. [7] propose the sparse linear discriminant analysis (SLDA) under the relaxed assumption that both  $\mu_d$  and  $\Sigma$  are sparse. In addition, [8] and [9] propose the sparse discriminant analysis method for LDA with multiple classes, where they assume the sparsity of the discriminant vectors and impose elastic net penalty on them.

Meanwhile, another track of work directly estimates the discriminant direction  $\beta^*$  under the assumption that  $\beta^* = \Sigma^{-1} \mu_d$  is sparse. [10] propose the linear programming discrim-

inant (LPD) rule. [11] propose to estimate  $\beta^*$  by minimizing an  $\ell_1$ -penalized least square loss. [12] propose the regularized optimal affine discriminant (ROAD) method. For better understanding, here we present the LPD rule in more details. The LPD rule provides an estimator  $\hat{\beta}$  for  $\beta^*$  by solving the following linear optimization problem that

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \|\beta\|_1, \quad \text{subject to} \quad \|\hat{\Sigma}\beta - \hat{\mu}_d\|_\infty \leq \lambda \hat{\sigma}_{\max},$$

where  $\lambda$  is a tuning parameter. The authors derive the fast convergence rate of  $\hat{\beta}$  by setting

$$\lambda = C \cdot \Delta \sqrt{\frac{\log p}{n}}, \quad (1.1)$$

where  $\hat{\sigma}_{\max} = \sqrt{\max_j \Sigma_{jj}}$ ,  $\Delta = \sqrt{\beta^{*\top} \Sigma \beta^*}$ , and  $C$  is a generic constant. In practice, this choice of  $\lambda$  relies on the population quantity  $\Delta$ , which takes substantial effort to tune. To reduce the tuning effort, [13] propose the adaptive linear discriminant analysis (AdaLDA) rule, which is a two-stage tuning-insensitive method that achieves the minimax optimal convergence rate in both estimation error and misclassification rate. Specifically, in each stage, the AdaLDA rule solves a linear optimization problem. In the first stage, it gives an initial estimator  $\tilde{\beta}$  to construct an estimator  $\hat{\Delta}$  for  $\Delta$ . In the second stage, plugging  $\hat{\Delta}$  into the LPD framework provides the final estimator for  $\beta^*$ .

In this chapter, we propose a one-stage pivotal method for high-dimensional linear discriminant analysis called **PANDA** (PivotAl liNear Discriminant Analysis), which is tuning-insensitive under the assumption that  $\beta^*$  is sparse. Our method is motivated by the pivotal method in [14] for high-dimensional linear regression, and it is pivotal that our tuning parameters do not depend on any population quantities to guarantee the optimal convergence rates. Unlike the AdaLDA method, the proposed PANDA method simultaneously estimates  $\beta^*$  and  $\Delta$  by solving a single convex optimization problem. Specifically, we impose a quadratic constraint  $\sqrt{\beta^\top \hat{\Sigma} \beta} \leq \tau$ , and introduce a quadratic penalty  $\tau^2$  in the objective. We show that the PANDA method attains the same minimax optimal conver-

gence rate as for the AdaLDA method. Moreover, our detailed numerical results show that the PANDA method achieves similar or more competitive performance than the LPD and AdaLDA methods.

The rest of this chapter is organized as follows. In Section 1.2, we briefly review the LDA problem and the AdaLDA rule. In Section 1.3, we propose the PANDA method. In Section 1.4, we provide theoretical justifications of PANDA. In Section ??, we present the numerical studies. In Section 1.7, we provide proofs of our main results.

**Notations.** Let  $v = (v_1, \dots, v_p)^\top \in \mathbb{R}^p$  be a  $p$ -dimensional real vector. We define the following vector norms:  $\|v\|_1 = \sum_{j=1}^p |v_j|$ ,  $\|v\|_2^2 = \sum_{j=1}^p v_j^2$ , and  $\|v\|_\infty = \max_{1 \leq j \leq p} |v_j|$ . For  $p \in \mathbb{N}$ , we denote by  $[p]$  the set  $\{1, 2, \dots, p\}$ . For  $j \in [p]$ , let  $e_j$  be the  $j$ -th canonical basis in  $\mathbb{R}^p$ . For  $S \subseteq [p]$ , let  $v_S$  denote the subvector of  $v$  confined to  $S$ , and  $S^c$  denotes the complement of  $S$ . For a matrix  $\Sigma \in \mathbb{R}^{p \times p}$ ,  $\Sigma \succ 0$  denotes that  $\Sigma$  is symmetric and positive definite, and  $\lambda_{\min}(\Sigma)$  and  $\lambda_{\max}(\Sigma)$  denote the smallest and the largest eigenvalue of  $\Sigma$ , respectively. We let  $\mathbf{0}$  and  $\mathbf{1}$  denote vectors with all the entries equal to 0 and 1, respectively. We use  $\mathbb{1}(\cdot)$  to denote the indicator function.

## 1.2 Background

In this section, we provide necessary mathematical background of LDA. For better presentation, we split this section into three subsections. We review the problem setup of LDA in Section 1.2.1, the AdaLDA method in Section 1.2.2, and Gautier’s method for linear regression in Section 1.2.3.

### 1.2.1 Problem Setup

We consider a binary classification, where the samples are from two classes that follow two Gaussian distributions, respectively. Specifically, we assume that the two Gaussian distributions have different means, denoted by  $\mu^{(0)}$  and  $\mu^{(1)}$ , respectively, and share the same covariance matrix  $\Sigma$ .

Our goal is to find a linear discriminant rule  $f_{\alpha,\beta}(\cdot)$  such that given a new sample  $Z$ , we predict the class label of  $Z$  by

$$f_{\alpha,\beta}(Z) = \begin{cases} 0, & \text{if } \beta^\top(Z - \alpha) \leq 0, \\ 1, & \text{if } \beta^\top(Z - \alpha) > 0, \end{cases} \quad (1.2)$$

where  $\alpha, \beta \in \mathbb{R}^p$ . For simplicity, we assume the two classes have equal prior weights, i.e.,

$$\mathbb{P}(Z \text{ is from Class 0}) = \mathbb{P}(Z \text{ is from Class 1}) = 1/2.$$

Then the misclassification rate of  $f_{\alpha,\beta}(\cdot)$  is

$$\begin{aligned} \mathcal{R}(f_{\alpha,\beta}) &= \frac{1}{2} \mathbb{P}_{Z \sim N(\mu^{(0)}, \Sigma)}(f_{\alpha,\beta}(Z) = 1) + \frac{1}{2} \mathbb{P}_{Z \sim N(\mu^{(1)}, \Sigma)}(f_{\alpha,\beta}(Z) = 0) \\ &= \frac{1}{2} \Phi \left( -\frac{\beta^\top(\alpha - \mu^{(0)})}{\sqrt{\beta^\top \Sigma \beta}} \right) + \frac{1}{2} \Phi \left( -\frac{\beta^\top(\mu^{(1)} - \alpha)}{\sqrt{\beta^\top \Sigma \beta}} \right), \end{aligned} \quad (1.3)$$

where  $\Phi$  is the CDF of the standard Gaussian distribution.

The optimal misclassification rate (also known as the Bayes error) is achieved by the Fisher's discriminant rule  $f_{\alpha^*, \beta^*}(\cdot)$ , where

$$\alpha^* = \frac{\mu^{(0)} + \mu^{(1)}}{2}, \quad \beta^* = \Sigma^{-1}(\mu^{(1)} - \mu^{(0)}).$$

Accordingly, the optimal misclassification rate is  $\mathcal{R}^* = \Phi(-\Delta/2)$ , where  $\Delta = \sqrt{\beta^{*\top} \Sigma \beta^*} = \sqrt{\mu_d^\top \Sigma^{-1} \mu_d}$  can be viewed as the signal-noise ratio of the classification problem.

### 1.2.2 The AdaLDA method

Suppose that we observe  $n_\ell$  independent samples for class  $\ell$ , denoted as  $\{X_i^{(\ell)}\}_{i=1}^{n_\ell}$ , where  $\ell = 0, 1$ . We denote the sample means of and the pooled covariance matrix as

$$\widehat{\mu}^{(\ell)} = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} X_i^{(\ell)} \quad \text{and} \quad \widehat{\Sigma} = \frac{1}{n_0 + n_1} \sum_{\ell=0,1} \sum_{i=1}^{n_\ell} (X_i^{(\ell)} - \widehat{\mu}^{(\ell)})(X_i^{(\ell)} - \widehat{\mu}^{(\ell)})^\top,$$

respectively. For notational simplicity, we assume  $n_0 = n_1 = n$ .

The AdaLDA method estimates  $\beta^*$  through two steps. In the first step, we solve the following linear optimization problem to obtain an initial estimator  $\widetilde{\beta}$ ,

$$\begin{aligned} \widetilde{\beta} \in \arg \min_{\beta} \quad & \|\beta\|_1, \\ \text{subject to} \quad & \|\widehat{\Sigma}\beta - \widehat{\mu}_d\|_\infty \leq 4\widehat{\sigma}_{\max} \cdot \sqrt{\frac{\log p}{n}} \cdot (\lambda\beta^\top \widehat{\mu}_d + 1), \end{aligned}$$

where  $\lambda$  is a tuning parameter,  $\widehat{\mu}_d = \widehat{\mu}^{(1)} - \widehat{\mu}^{(0)}$  is the difference of the sample means, and  $\widehat{\sigma}_{\max} = \sqrt{\max_j \widehat{\Sigma}_{jj}}$ . In the second step, we solve another linear optimization problem to obtain the final estimator  $\widehat{\beta}$  that

$$\begin{aligned} \widehat{\beta} \in \arg \min_{\beta} \quad & \|\beta\|_1, \\ \text{subject to} \quad & |e_j^\top (\widehat{\Sigma}\beta - \widehat{\mu}_d)| \leq 4\widehat{\sigma}_{\max} \cdot \sqrt{\frac{\log p}{n}} \cdot \sqrt{\lambda\widehat{\Delta}^2 + 1}, \quad \text{for all } j \in [p], \end{aligned} \quad (1.4)$$

where  $\widehat{\Delta}^2 = |\widetilde{\beta}^\top \widehat{\mu}_d|$  is essentially an estimator of  $\Delta$ . Consequently, we obtain the linear discriminant rule  $f_{\widehat{\mu}_m, \widehat{\beta}}$ , where  $\widehat{\mu}_m = (\widehat{\mu}^{(0)} + \widehat{\mu}^{(1)})/2$ .

With a slight abuse of the notation, we let  $\mathcal{R}(\widehat{\beta}) = \mathcal{R}(f_{\widehat{\mu}_m, \widehat{\beta}})$ . Since the tuning parameters in the two steps do not depend on any unknown population quantities, the AdaLDA method is tuning-insensitive. Assuming  $\beta^*$  contains at most  $s$  nonzero entries, [13] prove that under some mild assumptions, by choosing  $\lambda$  as a proper constant, both  $\widehat{\beta}$  and  $\mathcal{R}(\widehat{\beta})$

achieve the minimax optimal rates of convergence that

$$\|\widehat{\beta} - \beta^*\|_2 = \mathcal{O}_P \left( \Delta \sqrt{\frac{s \log p}{n}} \right) \quad \text{and} \quad \mathcal{R}(\widehat{\beta}) - \mathcal{R}^* = \mathcal{O}_P \left( \exp \left( -\frac{\Delta^2}{8} \right) \Delta \frac{s \log p}{n} \right).$$

### 1.2.3 Review of Gautier's method

In this subsection, we provide a brief review of Gautier's pivotal method for high-dimensional linear regression in [14] that inspires our work. Note that they consider a more complicated high-dimensional instrumental variables model. Here we discuss the particular case where the regressors and instruments are identical for ease of presentation. Specifically, let  $X \in \mathbb{R}^{n \times p}$  be a design matrix with  $n$  observations and  $p$  variables, and let  $y \in \mathbb{R}^n$  be the response vector. We consider the following linear model that

$$y = X^\top \beta + \varepsilon \quad \text{with} \quad \varepsilon \sim N(0, \sigma^2 I_n),$$

where  $\beta \in \mathbb{R}^p$  is the unknown regression coefficient with  $\|\beta\|_0 = s^* < n \ll p$ , and  $\varepsilon$  is the noise. The Gautier's estimator can be viewed as a variant of the Dantzig selector [15], and is the optimal solution to the following convex optimization problem that

$$\begin{aligned} (\widehat{\beta}, \widehat{\gamma}) &= \underset{\beta, \gamma}{\operatorname{argmin}} \|\beta\|_1 + c\gamma, \\ \text{subject to} \quad &\frac{1}{n} \|X^\top (Y - X\beta)\|_\infty \leq \lambda\gamma, \quad \frac{1}{n} \|Y - X\beta\|_2^2 \leq \gamma^2, \end{aligned} \tag{1.5}$$

where  $c$  and  $\lambda$  are two tuning parameters, and  $\widehat{\gamma}$  is an estimator of  $\sigma$ . The theoretical analysis in [14] suggests that the tuning parameter  $c$  can be set as a constant between 0 and 1, and the tuning parameter  $\lambda$  can be chosen as

$$\lambda = A \cdot \sqrt{\frac{2 \log p}{n}},$$

where  $A$  is a constant independent of  $\sigma$ . In practice,  $c$  can be chosen as 0.5, and  $A$  only needs to be tuned between 0.5 and 2. Therefore, the Gautier's estimator is less sensitive to the parameter tuning than the Dantzig selector, where the tuning parameter depends on  $\sigma$ .

### 1.3 The PANDA Method

We propose the PANDA method for linear discriminant analysis in high dimensions. The PANDA method is a one-stage method. Specifically, we estimate  $\beta^*$  and  $\Delta$  simultaneously by solving the convex optimization problem that

$$\begin{aligned} (\hat{\beta}, \hat{\tau}) \in \arg \min_{\beta \in \mathbb{R}^p, \tau \in \mathbb{R}} \quad & \|\beta\|_1 + c\tau^2, \\ \text{subject to} \quad & \|\hat{\Sigma}\beta - \hat{\mu}_d\|_\infty \leq \lambda \hat{\sigma}_{\max}(\tau + 1), \quad \sqrt{\beta^\top \hat{\Sigma} \beta} \leq \tau, \end{aligned} \quad (1.6)$$

where  $c > 0$  and  $\lambda > 0$  are two tuning parameters.

Our formulation is motivated by the LPD method in [10] and the pivotal method for high-dimensional linear regression in [14], as introduced in the previous 1.2.3. Specifically, the LPD method estimates  $\beta^*$  by solving the following linear optimization problem that

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \|\beta\|_1, \quad \text{subject to} \quad \|\hat{\Sigma}\beta - \hat{\mu}_d\|_\infty \leq \lambda \hat{\sigma}_{\max}. \quad (1.7)$$

However, as discussed earlier, the tuning parameter  $\lambda$  in (1.7) depends on the population quantity  $\Delta = \sqrt{\beta^{*\top} \Sigma \beta^*}$ , which is difficult to tune in practice. To address this issue, we introduce an estimator of  $\Delta$ , and plug it into (1.7). This leads to the following problem that

$$\begin{aligned} (\hat{\beta}, \hat{\tau}) \in \arg \min_{\beta \in \mathbb{R}^p, \tau \in \mathbb{R}} \quad & \|\beta\|_1, \quad \text{subject to} \quad \|\hat{\Sigma}\beta - \hat{\mu}_d\|_\infty \leq \lambda \hat{\sigma}_{\max}(\tau + 1), \quad \sqrt{\beta^\top \hat{\Sigma} \beta} = \tau, \end{aligned} \quad (1.8)$$

where  $\tau$  estimates  $\Delta$  by the constraint  $\sqrt{\beta^\top \hat{\Sigma} \beta} = \tau$ .

The optimization problem in (1.8) is nonconvex due to the quadratic equality constraint  $\sqrt{\beta^\top \widehat{\Sigma} \beta} = \tau$ . Thus, we propose to relax the equality constraint into an inequality constraint, and obtain

$$(\widehat{\beta}, \widehat{\tau}) \in \arg \min_{\beta \in \mathbb{R}^p, \tau \in \mathbb{R}} \|\beta\|_1, \quad \text{subject to} \quad \|\widehat{\Sigma} \beta - \widehat{\mu}_d\|_\infty \leq \lambda \widehat{\sigma}_{\max}(\tau + 1), \quad \sqrt{\beta^\top \widehat{\Sigma} \beta} \leq \tau. \quad (1.9)$$

However, as the objective function in (1.9) is free of  $\tau$ ,  $\tau$  can be arbitrarily large. Therefore, (1.9) admits a trivial solution  $\widehat{\beta} = 0$ , and  $\widehat{\tau}$  can be any value larger than  $\lambda^{-1} \|\widehat{\mu}_d\|_\infty - 1$ .

To address this issue, we introduce a penalty term  $c\tau^2$  to the objective in (1.9), which leads to our proposed formulation in (1.6). Note that different from the linear penalty term used in Gautier's pivotal method, our penalty term is quadratic in  $\tau$ , which is due to some specific modeling structure in LDA.

We remark that the problem in (1.6) is a second order conic optimization problem. By introducing auxiliary variables  $w \in \mathbb{R}^p$  and  $u \in \mathbb{R}$ , the problem in (1.6) is equivalent to the following problem that

$$\begin{aligned} \min_{\beta, \tau, w, u} \quad & \sum_{j=1}^p w_j + cu, & (1.10) \\ \text{subject to} \quad & -w_j \leq \beta_j \leq w_j, \quad -\lambda \widehat{\sigma}_{\max}(\tau + 1) \mathbf{1} \leq \widehat{\Sigma} \beta - \widehat{\mu}_d \leq \lambda \widehat{\sigma}_{\max}(\tau + 1) \mathbf{1}, \\ & \|\widehat{\Sigma}^{1/2} \beta\|_2 \leq \tau, \quad \sqrt{\tau^2 + \frac{1}{4}(1-u)^2} \leq \frac{1}{2}(1+u). \end{aligned}$$

Such a second order conic optimization problem is convex, and can be solved in a polynomial time using the interior point method [16]. We also provide an efficient scheme using the alternating direction method of multipliers (ADMM) following [17] in the next subsection to solve it.

### 1.3.1 An ADMM Algorithm for PANDA

This section present the implementation of the ADMM algorithm for solving (1.6). For this purpose, we first re-write the problem (1.6) as

$$\begin{aligned}
 (\hat{\beta}, \hat{\tau}) \in \arg \min_{\beta, u, v, w, \tau \in \mathbb{R}^p, \tau \in \mathbb{R}} \quad & \|\beta\|_1 + c\tau^2 & (1.11) \\
 \text{subject to} \quad & \hat{\Sigma}\beta - \lambda\hat{\sigma}_{\max}\tau\mathbf{1} + u = \hat{\mu}_d + \lambda\hat{\sigma}_{\max}\mathbf{1}, \\
 & \hat{\Sigma}\beta + \lambda\hat{\sigma}_{\max}\tau\mathbf{1} - v = \hat{\mu}_d - \lambda\hat{\sigma}_{\max}\mathbf{1}, \\
 & w - \hat{\Sigma}^{1/2}\beta = 0, \\
 & u \geq 0, v \geq 0, \\
 & \|w\|_2 \leq \tau.
 \end{aligned}$$

Note that the first three constraints in (1.11) are linear and the last three constraints are conic.

To simplify the notation, we write the first three linear constraints as

$$A_\beta\beta + A_uu + A_vv + A_w w + A_\tau\tau = b$$

for some real matrices  $A_\beta, A_u, A_v, A_w, A_\tau$  and real vector  $b$ . We can further write the problem as

$$\begin{aligned}
 (\hat{\beta}, \hat{\tau}) \in \arg \min_{\beta, u, v, w \in \mathbb{R}^p, \tau \in \mathbb{R}} \quad & \|\beta\|_1 + c\tau^2 \\
 \text{subject to} \quad & A_\beta\beta + A_uu + A_vv + A_w w + A_\tau\tau = b, \\
 & u, v \in \mathcal{C}_1, \\
 & (w, \tau) \in \mathcal{C}_2,
 \end{aligned}$$

where

$$\mathcal{C}_1 = \{x \in \mathbb{R}^p : x_j \geq 0, j \in [p]\},$$

$$\mathcal{C}_2 = \left\{ (x, y) \in \mathbb{R}^p \times \mathbb{R} : y \geq \sqrt{\sum_{j=1}^p x_j^2} \right\}$$

are two convex cones.

The augmented Lagrangian function with scaled dual variables is

$$L_\rho(\beta, u, v, w, \tau, s) = \|\beta\|_1 + c\tau^2 + \frac{\rho}{2} \|A_\beta\beta + A_u u + A_v v + A_w w + A_\tau \tau - b + s\|_2^2 - \frac{\rho}{2} \|s\|_2^2,$$

where  $s$  is the scaled dual variable and  $\rho > 0$  is the penalty parameter.

Algorithm 1 summarizes the ADMM algorithm for solving (1.6).

---

**Algorithm 1** ADMM with proximal method for solving problem (1.6)

---

**Require:** Sample mean difference  $\widehat{\mu} = \widehat{\mu}^{(1)} - \widehat{\mu}^{(0)}$ ; Pooled sample covariance matrix  $\widehat{\Sigma}$ ;  
Tuning parameter  $c, \lambda$ ; Initialization  $\beta^0, \tau^0, u^0, v^0, w^0, s^0$ ; Penalty parameter  $\rho > 0$ ;  
Primal step size  $\eta > 0$ ; Number of iterations  $T$ .

**for**  $t = 1, 2, \dots, T$  **do**

$$\begin{aligned} \beta^t &\leftarrow \beta^{t-1} - \eta \nabla_\beta L_\rho(\beta^{t-1}, u^{t-1}, v^{t-1}, w^{t-1}, \tau^{t-1}, s^{t-1}) \\ u^t &\leftarrow \Pi_{\mathcal{C}_1}[u^{t-1} - \eta \nabla_u L_\rho(\beta^t, u^{t-1}, v^{t-1}, w^{t-1}, \tau^{t-1}, s^{t-1})] \\ v^t &\leftarrow \Pi_{\mathcal{C}_2}[v^{t-1} - \eta \nabla_v L_\rho(\beta^t, u^t, v^{t-1}, w^{t-1}, \tau^{t-1}, s^{t-1})] \\ \tilde{\tau}^t &\leftarrow \tau^{t-1} - \eta \nabla_\tau L_\rho(\beta^t, u^t, v^t, w^{t-1}, \tau^{t-1}, s^{t-1}) \\ \tilde{w}^t &\leftarrow w^{t-1} - \eta \nabla_w L_\rho(\beta^t, u^t, v^t, w^{t-1}, \tau^{t-1}, s^{t-1}) \\ (w^t, \tau^t) &\leftarrow \Pi_{\mathcal{C}_2}(\tilde{w}^t, \tilde{\tau}^t) \\ s^t &\leftarrow s^{t-1} + A_\beta \beta^t + A_u u^t + A_v v^t + A_w w^t + A_\tau \tau^t - b^t \end{aligned}$$

**end for**

---

## 1.4 Statistical Properties

We establish theoretical guarantees for our proposed PANDA method. We assume that  $\beta^*$  belongs to a class of weakly sparse vectors that for some  $q \in [0, 1)$

$$\mathbb{B}_q(R) = \left\{ \beta \in \mathbb{R}^p : \sum_j |\beta_j|^q \leq R \right\}, \quad (1.12)$$

where  $R$  can scale with  $n$  and  $p$ . Note that when  $q = 0$ ,  $\mathbb{B}_q(R)$  is reduced to the class of  $R$ -sparse vectors, i.e.,  $\mathbb{B}_0(R) = \left\{ \beta \in \mathbb{R}^p : \sum_j \mathbb{1}(\beta_j \neq 0) \leq R \right\}$ .

For notational simplicity, we let

$$\begin{aligned} \mu_m &= (\mu^{(0)} + \mu^{(1)})/2, & \hat{\mu}_m &= (\hat{\mu}^{(0)} + \hat{\mu}^{(1)})/2, & \mu_d &= \mu^{(1)} - \mu^{(0)}, \\ \hat{\mu}_d &= \hat{\mu}^{(1)} - \hat{\mu}^{(0)}, & \sigma_{\max} &= \max_j (\Sigma_{jj})^{1/2}, & \hat{\sigma}_{\max} &= \max_j (\hat{\Sigma}_{jj})^{1/2}. \end{aligned}$$

Here we only consider the case where  $n_0 = n_1 = n$ , and our results can be easily extended to the general case where  $n_0 \neq n_1$ .

Before we proceed with our main results, we first impose the following mild assumptions.

**Assumption 1.4.1** *There exists a constant  $a$  such that  $\|\mu_d\|_\infty \geq a > 0$ .*

**Assumption 1.4.2** *There exists some  $M$  such that  $M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M$ .*

Essentially, Assumption 1.4.1 requires the two classes to be distinguishable, and Assumption 1.4.2 requires the covariance matrix  $\Sigma$  to be sufficiently well-conditioned, as its condition number is upper bounded by  $M^2$ .

We are now ready to present the theoretical guarantees of the PANDA method in (1.6). Let us begin with the convergence rates of  $\hat{\beta}$  and  $\hat{\tau}$ .

**Theorem 1.4.1 (Parameter Estimation)** *Suppose that Assumption 1.4.2 hold, and  $\beta^* \in \mathbb{B}_q(R)$  for some  $q \in [0, 1)$  and some  $R > 0$ . Let  $(\widehat{\beta}, \widehat{\tau})$  be an optimal solution of (1.6). Given*

$$c = \frac{1}{8 \left( \|\widehat{\mu}_d\|_\infty + 4\widehat{\sigma}_{\max} \sqrt{\frac{\log p}{n}} \right)}, \quad \lambda = 20 \sqrt{\frac{\log p}{n}}, \quad (1.13)$$

for sufficiently large  $n$  such that

$$n \geq C \cdot a^{-2} \Delta^2 \sigma_{\max}^2 M^{2+\frac{1}{1-q}} R^{\frac{2}{1-q}} \log p \quad (1.14)$$

where  $C$  is an absolute constant, we have, with probability goes to 1,

$$\|\widehat{\beta} - \beta^*\|_1 \leq C_1 \cdot (\Delta + 1) (\sigma_{\max} M)^{1-q} R \left( \frac{\log p}{n} \right)^{(1-q)/2}, \quad (1.15a)$$

$$\|\widehat{\beta} - \beta^*\|_2 \leq C_2 \cdot (\Delta + 1) (\sigma_{\max} M)^{1-q/2} \sqrt{R} \left( \frac{\log p}{n} \right)^{1/2-q/4}, \quad (1.15b)$$

$$\frac{|\widehat{\tau}^2 - \Delta^2|}{\Delta^2} \leq C_3 \cdot (1 + \Delta^{-1}) \sigma_{\max}^{1-q/2} M^{3/2-q} R \left( \frac{\log p}{n} \right)^{(1-q)/2}, \quad (1.15c)$$

where  $C_1, C_2$  and  $C_3$  are positive constants.

Note that our proposed PANDA method is tuning-insensitive, as the chosen tuning parameters  $c$  and  $\lambda$  in (1.13) do not depend on any unknown population quantity. In the next theorem, we show that the sample complexity requirement (1.14) can be relaxed under some more restrictive conditions.

**Theorem 1.4.2** *Suppose that Assumption 1.4.2 holds, and  $\beta^* \in \mathbb{B}_q(R)$  for some  $q \in [0, 1)$  and some  $R > 0$ . Let  $(\widehat{\beta}, \widehat{\tau})$  be an optimal solution to problem (1.6). When  $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta}}$ ,*

given

$$c = \frac{1}{8 \left( \|\widehat{\mu}_d\|_\infty + 4\widehat{\sigma}_{\max} \sqrt{\frac{\log p}{n}} \right)}, \quad \lambda = 20 \sqrt{\frac{\log p}{n}},$$

for sufficiently large  $n$  such that

$$n \geq C \cdot a^{-2} \Delta^2 \sigma_{\max}^2 M^{2+\frac{1}{1-q}} R^{\frac{2}{2-q}} \log p, \quad (1.16)$$

where  $C$  is an absolute constant, we have, with probability goes to 1,

$$\|\widehat{\beta} - \beta^*\|_2 \leq C_1 \cdot (\sigma_{\max} M)^{1-q/2} (\Delta + 1) \sqrt{R} \left( \frac{\log p}{n} \right)^{1/2-q/4}, \quad (1.17a)$$

$$\frac{|\widehat{\tau}^2 - \Delta^2|}{\Delta^2} \leq C_2 \cdot (1 + \Delta^{-1}) \sigma_{\max}^{1-q/2} M^{3/2-q} \sqrt{R} \left( \frac{\log p}{n} \right)^{(1-q)/2}, \quad (1.17b)$$

where  $C_1$  and  $C_2$  are positive constants.

Note that in the above theorem, we impose the additional assumption that  $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta}}$ , i.e. the second inequality constraint of PANDA is active at the optimal solution. We point out that in practice, we can numerically verify if this assumption indeed holds. Also, in our later simulations, we find that this assumption holds when the tuning parameters are properly chosen.

We next compare our results with [13] for  $q = 0$ . Note that [13] consider the following parameter space of  $\beta^*$  and  $\Sigma$ ,

$$\Theta_s = \left\{ (\beta^*, \Sigma) : \beta^* \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p}, |\text{supp}(\beta^*)| \leq s, \right. \\ \left. M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M, \Delta \geq c_L > 0 \right\}, \quad (1.18)$$

where  $M$  and  $c_L$  are absolute constants that do not scale with  $n$ ,  $p$  and  $s$ . They then establish

the following minimax lower bound,

$$\inf_{\widehat{\beta}} \sup_{(\beta^*, \Sigma) \in \Theta_s} \mathbb{E} \left[ \|\widehat{\beta} - \beta^*\|_2 \right] \geq C_M \cdot \Delta \sqrt{\frac{s \log p}{n}},$$

where the infimum is taken over any estimator  $\widehat{\beta}$  based on the samples, and  $C_M$  is some constant depending on  $M$ . Under such a setting, both AdaLDA and PANDA are minimax optimal in terms of estimating  $\beta^*$ . When  $M$  is allowed to scale with  $n$ ,  $p$  and  $s$ , the PANDA method still attains the same rates of convergence for parameter estimation as the AdaLDA method. Specifically, we follow the same analysis in [13] and rewrite their results with explicit dependence on  $M$  as follows,

$$\begin{aligned} \|\widehat{\beta} - \beta^*\|_2 &= \mathcal{O}_P \left( \sigma_{\max} M \Delta \sqrt{\frac{s \log p}{n}} \right), \\ \frac{|\widehat{\Delta}^2 - \Delta^2|}{\Delta^2} &= \mathcal{O}_P \left( \sigma_{\max} M^{3/2} \sqrt{\frac{s \log p}{n}} \right). \end{aligned}$$

In addition, to ensure the above rates of convergence with high probability, the sample size  $n$  needs to satisfy that

$$n = \mathcal{O}_P \left( \sigma_{\max}^2 M^3 \Delta^2 s \log p \right).$$

As can be seen, in Theorem 1.4.2, our convergence rates (1.17a) and (1.17b) matches the convergence rates in [13] with the same order of sample complexity.

Next, let us establish an upper bound for the misclassification rate of the obtained estimator  $\widehat{\beta}$  in the PANDA method.

**Theorem 1.4.3 (Misclassification Rate)** *Under the identical conditions as in Theorem 1.4.1 or 1.4.2, we have, with probability goes to 1,*

$$\mathcal{R}(\widehat{\beta}) - \mathcal{R}^* \leq C \cdot \exp \left( -\frac{\Delta^2}{8} \right) \sigma_{\max}^{-q} M^{3-q} \Delta R \left( \frac{\log p}{n} \right)^{1-q/2}$$

where  $C$  is an absolute positive constants.

When  $q = 0$  and  $R = s$ , [13] consider the parameter space of  $\beta^*$  and  $\Sigma$  defined in (1.18), where  $M$  is a constant, and establish the following minimax lower bound

$$\inf_{\hat{f}} \sup_{(\beta^*, \Sigma) \in \Theta_s} \mathcal{R}(\hat{f}) - \mathcal{R}^* \geq C \cdot \exp\left(-\frac{\Delta^2}{8}\right) \Delta^{-1} \frac{s \log p}{n},$$

where the infimum is taken over any linear discriminant rule  $\hat{f}$  based on the samples. Under such a setting, both AdaLDA and PANDA attain the minimax optimal rates of convergence for the misclassification rate that

$$\mathcal{R}(\hat{\beta}) - \mathcal{R}^* = \mathcal{O}_P\left(\exp\left(-\frac{\Delta^2}{8}\right) M^3 \Delta \frac{s \log p}{n}\right).$$

## 1.5 Numerical Results

In this section, we thoroughly compare our proposed PANDA method with the LPD method and AdaLDA method through numerical experiments using both simulated and real data.

### 1.5.1 Simulation

To make a fair comparison of the three methods' performances, we fine-tune the parameters for each method on a validation dataset independent from the training data, and we provide both the estimation error of  $\beta^*$  (in  $\ell_2$  norm) and the population risk (2.4) of each method.

**Settings:** We follow the settings in [13] to generate  $\Sigma$  and  $\beta^*$ .

- (a) **AR(1).** We let  $\Omega_{j,k} = 0.9^{|j-k|}$ ,  $\Sigma = \Omega^{-1}$  and  $\beta^* = (2/\sqrt{s}, \dots, 2/\sqrt{s}, 0, \dots, 0)^\top$ , where the first  $s$  entries are non-zero.
- (b) **Varying diagonal.** We let the diagonal entries of  $\Sigma$  as  $\Sigma_{j,j} = 11$  for  $j = 1, 2, \dots, 5$ , and  $\Sigma_{j,j} = 1 + U_j$  for  $j = 6, 7, \dots, p$ , where  $U_i$ 's are independently drawn from the

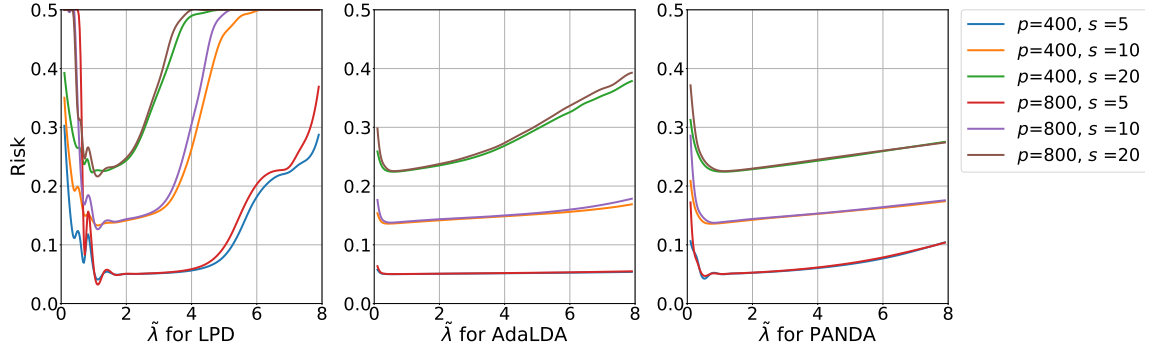
uniform distribution  $U(0, 1)$ , and we let the off-diagonal entries be  $\Sigma_{j,k} = 0.9^{|j-k|}$ . We let  $\beta^* = (1/\sqrt{s}, \dots, 1/\sqrt{s}, 0, \dots, 0)^\top$ , where only the first  $s$  entries are non-zero.

(c) **Erdős-Rényi random graph.** We let  $\tilde{\Omega}_{j,k} = u_{j,k}v_{j,k}$ , where  $v_{j,k}$ 's are i.i.d. Bernoulli random variables with success rate 0.2, and  $u_{j,k}$ 's are i.i.d. uniform random variables over  $[0.5, 1] \cup [-1, -0.5]$ , and  $v_{j,k}$ 's and  $u_{j,k}$ 's are independent. Then we let  $\tilde{\Omega}_s = (\tilde{\Omega} + \tilde{\Omega}^\top)/2$  and  $\Omega_0 = \tilde{\Omega}_s + \left[ \max(-\lambda_{\min}(\tilde{\Omega}_s), 0) + 0.05 \right] I_p$ . Let  $D_0$  be a diagonal matrix with diagonal elements same as  $\Omega_0$ 's. We let  $\Omega = D_0^{-1/2} \Omega_0 D_0^{-1/2}$  and  $\Sigma = \Omega^{-1}$ , and let  $\beta^* = (1/\sqrt{s}, \dots, 1/\sqrt{s}, 0, \dots, 0)^\top$  where only the first  $s$  entries are non-zero.

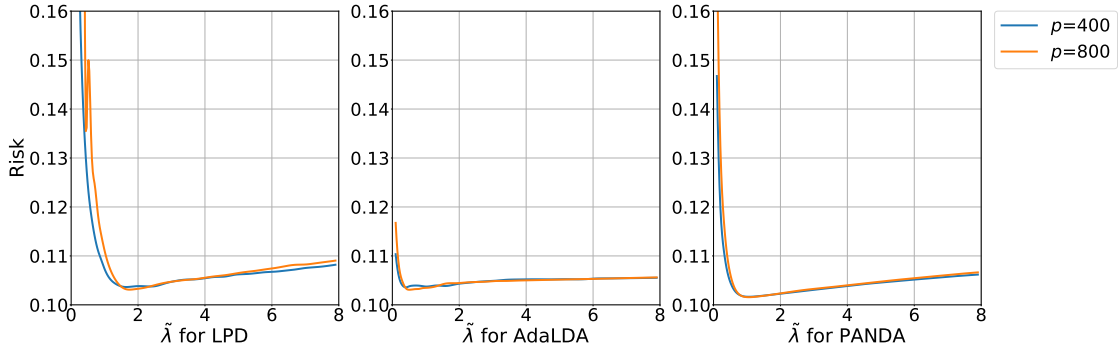
(d) **Block sparse model.** We first construct a matrix  $B$  of size  $p \times p$  as follows. For  $1 \leq j \leq p/2$  and  $j < k \leq p$ , we let  $B_{j,k} = B_{k,j} = 10b_{j,k}$ , where  $b_{j,k}$ 's are i.i.d. Bernoulli variables with success rate 0.5. For  $p/2 < j < k \leq p$ , we let  $B_{j,k} = B_{k,j} = 10$ . For the diagonal elements, we let  $B_{j,j} = 1$  for  $1 \leq j \leq p$ . Then we let  $w = \max(-\lambda_{\min}(B), 0) + 0.05$  and let  $\Omega = (B + wI_p)/(1 + w)$  and  $\Sigma = \Omega^{-1}$ . We let  $\beta^* = (\frac{1}{2\sqrt{s}}, \dots, \frac{1}{2\sqrt{s}}, 0, \dots, 0)^\top$ , where only the first  $s$  entries are non-zero.

(e) **Approximately sparse setting.** We let  $\Sigma_{j,k} = 0.9^{|j-k|}$  and  $\beta_j^* = 0.75^j$ , which are approximately sparse.

**Parameter Tuning:** While both the AdaLDA method and PANDA method achieve guaranteed theoretical properties with specific tuning parameters, we observe in our experiments that tuning these parameters via a validation set yields better empirical results. In our experiments, under each setting, we randomly sample a validation dataset with  $n = 200$  data points from each class. Motivated by the choice of  $\lambda$  in (1.13), we let  $\lambda = \tilde{\lambda} \cdot \sqrt{\log p/n}$ , and we tune the parameter  $\tilde{\lambda}$ , as equivalent to tuning  $\lambda$ . For a fair comparison, for all the three methods (LPD, AdaLDA, and PANDA) we tune  $\tilde{\lambda}$  by a grid search over a range from 0.1 to 8.0, with a grid size 0.1. Figures 1.1 and 1.2 show the results of the misclassification risks and the estimation errors  $\|\hat{\beta} - \beta^*\|_2$  versus the  $\tilde{\lambda}$  value in the three methods, averaged over 100 repetitions under each setting of different  $p$  and  $s$ . For the parameter  $c$  in

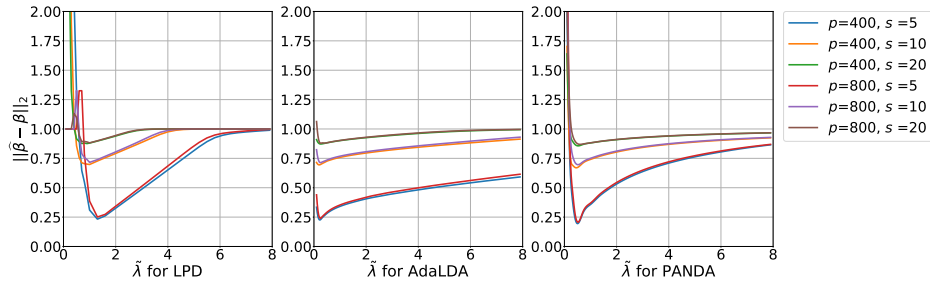


(a) Varying diagonal

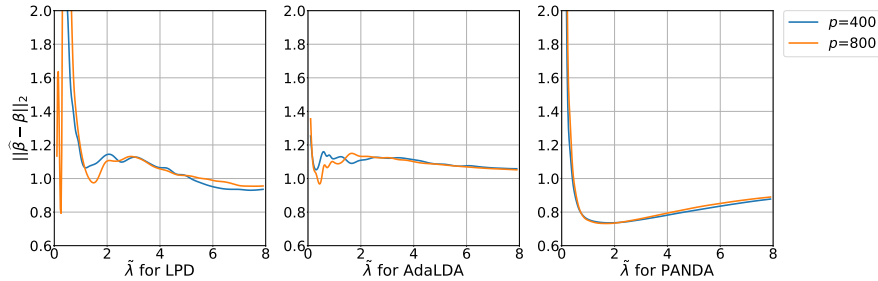


(b) Approximately sparse

Figure 1.1: The misclassification rate v.s. the parameter  $\tilde{\lambda}$ . Results are averaged over 100 instances.



(a) Varying diagonal



(b) Approximately Sparse

Figure 1.2:  $\ell_2$  estimation error v.s. values parameter  $\tilde{\lambda}$ . Results are averaged over 100 instances.

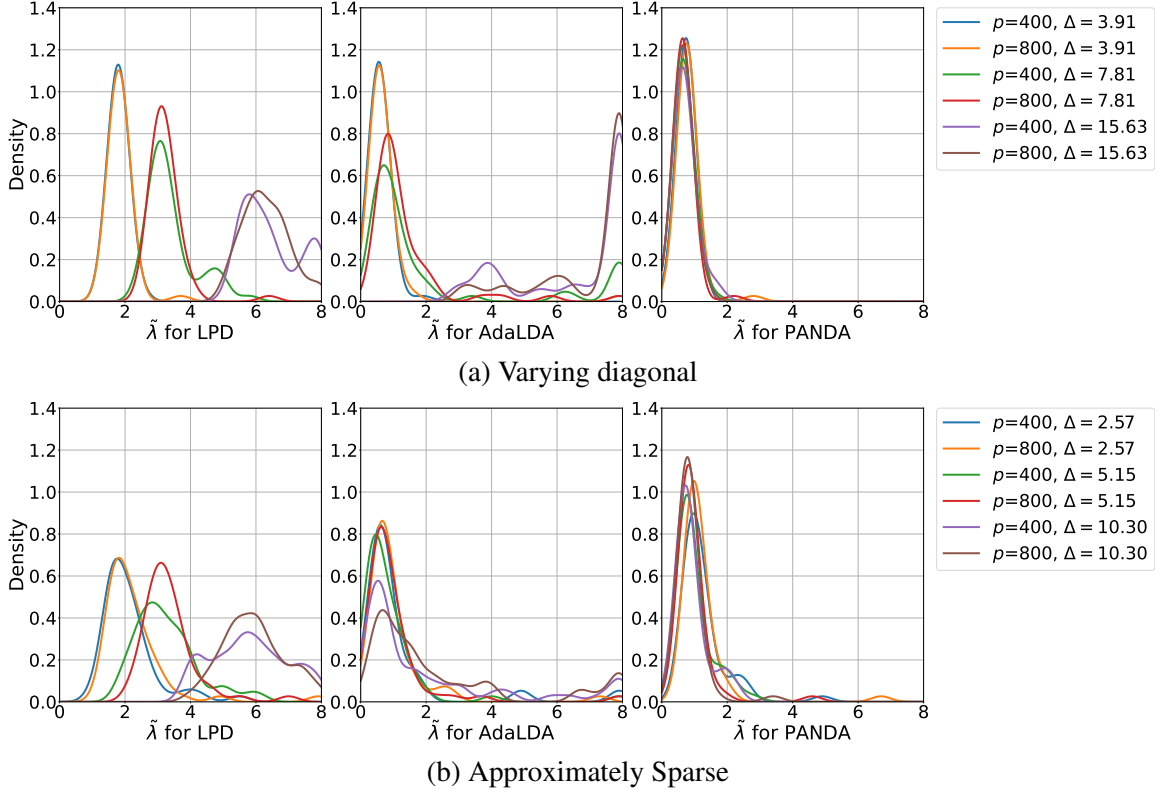


Figure 1.3: The distribution of the empirically optimal tuning parameter  $\tilde{\lambda}$  for LPD (left), AdaLDA (middle) and PANDA (right), approximated with kernel smoothing. The optimal choice of the tuning parameter in our method relies less on the population.

the PANDA method, we observe that the results are insensitive to the value of  $c$  as long as  $c$  is not too small, see Table 1.1 for the result of the classification risk with different choices of  $c$  under the AR(1) model as an example. Hence we set  $c = 20$  for all settings.

**Tuning Sensitivity:** We thoroughly investigate the sensitivity of the tuning parameters under different settings. Since the choice of  $\lambda$  in the LPD method relies on the unknown population quantity  $\Delta$ , so does the optimal value of  $\lambda$  (or  $\tilde{\lambda}$ , equivalently) in practice. We consider following settings to see how the population distribution, especially the scale of  $\Delta$ , changes the empirically optimal tuning parameters of the LPD, AdaLDA and PANDA methods. For the varying diagonal model, we set  $p = 400, 800$ ,  $s = 5$ , and  $\beta^* = \eta \cdot (1/\sqrt{s}, \dots, 1/\sqrt{s}, 0, \dots, 0)^\top$  for  $\eta = 1, 2, 4$ , where the first  $s$  entries are non-zero. For the approximately sparse  $\beta$  model, we set  $p = 400, 800$ , and  $\beta_j^* = \eta \cdot 0.75^j$  for  $\eta = 1, 2, 4$ .

Table 1.1: Misclassification rate of the PANDA method under the AR(1) model with  $n = 200$ ,  $p = 400$ ,  $s = 5$  and different  $c$ , averaged over 50 replications. The standard deviations are provided in brackets.

$c$	1e-4	1e-3	1e-2	0.1
Misclassification rate	0.4271 (0.1677)	0.3729 (0.1489)	0.2155 (0.0035)	0.2106 (0.0050)
$c$	1	10	100	
Misclassification rate	0.2044 (0.0049)	0.2036 (0.0053)	0.2035 (0.0054)	

During the tuning process, we observe that the empirically optimal tuning parameter  $\tilde{\lambda}$  for the PANDA method is less sensitive to the change of unknown population quantities among different settings, in comparison with the LPD method and AdaLDA method. In particular, Figure 1.3 shows the distribution of the empirically optimal tuning parameter over 50 repetitions under each setting as specified above. The results show that for the PANDA method, the optimal tuning parameter is always close to 1, and does not change much across the different settings.

**Parameter Estimation:** Table 1.2 summarizes the estimation error of  $\beta^*$ ,  $\|\hat{\beta} - \beta^*\|_2$ , averaged over 100 random instances under each setting. It is seen that our proposed PANDA method achieves equal or better performance than the LPD and AdaLDA methods for most settings.

**Risk Evaluation:** Table 1.3 summarizes the classification risk under each setting averaged over 100 random instances. It is seen that our proposed PANDA method achieves similar or better performances than the LPD method and AdaLDA method in most settings.

**Running Time Comparison:** Table 1.4 summarizes the running time of our PANDA method and the AdaLDA method under the Varying Diagonal model on a regular computer (Intel Core i5, 2.3GHz). For both methods we use Gurobi, a commercial software that provides state-of-art solver for linear programming and second order cone programming, to solve the optimization problems. As can be seen, our PANDA method requires less running time than the AdaLDA method.

Table 1.2: The  $\ell_2$  norm of estimation errors under each setting, averaged over 100 instances. The standard deviations are given in brackets. The lower value at the significance level 0.05 between the AdaLDA and the PANDA method are marked in bold.

Model	Specification						
	$(s, p)$	(5, 400)	(10, 400)	(20, 400)	(5, 800)	(10, 800)	(20, 800)
AR(1)							
	LPD	3.5650 (0.1859)	3.8453 (0.1251)	3.9386 (0.0402)	3.5964 (0.1568)	3.8692 (0.0782)	3.9474 (0.0374)
	AdaLDA	3.5573 (0.1862)	3.8206 (0.0778)	3.9289 (0.0386)	3.5936 (0.1555)	3.8391 (0.0718)	3.9402 (0.0333)
	PANDA	<b>3.4896</b> (0.2005)	<b>3.8112</b> (0.0891)	3.9260 (0.0443)	<b>3.5577</b> (0.1730)	3.8304 (0.0742)	<b>3.9324</b> (0.0412)
Varying Diagonal							
	LPD	0.1100 (0.0800)	0.5313 (0.0733)	0.7820 (0.0317)	0.1123 (0.0858)	0.5412 (0.0594)	0.8012 (0.0344)
	AdaLDA	<b>0.0808</b> (0.0423)	0.5197 (0.0564)	0.7812 (0.0304)	<b>0.0923</b> (0.0492)	0.5447 (0.0556)	0.7986 (0.0263)
	PANDA	0.1091 (0.0824)	0.5174 (0.0693)	<b>0.7742</b> (0.0302)	0.1157 (0.0917)	<b>0.5350</b> (0.0562)	<b>0.7926</b> (0.0313)
Erdős-Rényi Random Graph							
	LPD	0.3387 (0.1309)	0.5092 (0.1396)	1.1104 (0.4096)	0.3655 (0.1614)	0.5966 (0.1316)	0.8806 (0.1658)
	AdaLDA	0.3363 (0.1355)	0.4811 (0.1054)	1.0151 (0.1278)	0.3634 (0.1209)	0.5922 (0.1466)	0.8790 (0.2318)
	PANDA	<b>0.2991</b> (0.1170)	0.5455 (0.5919)	<b>0.9061</b> (0.1768)	0.3402 (0.1164)	<b>0.5477</b> (0.1716)	0.8857 (0.8482)
Block Sparse							
	LPD	0.2705 (0.1602)	0.3370 (0.2594)	0.4867 (0.5840)	0.2247 (0.0970)	0.3056 (0.0808)	0.3094 (0.1280)
	AdaLDA	0.2657 (0.0321)	0.3003 (0.0090)	0.3354 (0.0127)	0.2307 (0.0220)	0.2908 (0.0154)	<b>0.2537</b> (0.0044)
	PANDA	<b>0.1902</b>	<b>0.2493</b>	<b>0.2934</b>	0.2446	0.2778	0.3053

Continued on next page

**Table 1.2 – continued from previous page**

Setting	Specification					
		(0.0435)	(0.0272)	(0.0305)	(0.1565)	(0.0788) (0.1416)
	<i>p</i>	400	800	1200		
Approximately	LPD	1.2885	1.1354	1.0565		
		(0.9126)	(0.6862)	(0.7033)		
Sparse	AdaLDA	1.2190	1.1768	1.1083		
		(0.6756)	(0.7271)	(0.6613)		
	PANDA	<b>0.8117</b>	<b>0.8482</b>	<b>0.8152</b>		
		(0.5405)	(0.4735)	(0.5004)		

Table 1.3: The classification risk under each setting averaged over 100 instances. The standard deviations are given in brackets. The lower value at the significance level 0.05 between the AdaLDA and the PANDA method are marked in bold.

Model	Specification						
	$(s, p)$	(5, 400)	(10, 400)	(20, 400)	(5, 800)	(10, 800)	(20, 800)
AR(1)	LPD	0.2086	0.2900	0.3535	0.2112	0.2908	0.3532
		(0.0074)	(0.0109)	(0.0099)	(0.0074)	(0.0066)	(0.0080)
	AdaLDA	0.2082	0.2890	0.3522	0.2120	0.2913	<b>0.3525</b>
		(0.0068)	(0.0080)	(0.0075)	(0.0088)	(0.0072)	(0.0082)
PANDA	<b>0.2068</b>	0.2886	0.3542	0.2114	0.2910	0.3571	
		(0.0069)	(0.0087)	(0.0104)	(0.0084)	(0.0079)	(0.01206)
Varying Diagonal	LPD	0.0515	0.1382	0.2269	0.0520	0.1390	0.2289
		(0.0028)	(0.0054)	(0.0065)	(0.0038)	(0.0056)	(0.0087)
	AdaLDA	0.0508	0.1376	0.2266	0.0513	0.1386	0.2284
		(0.0018)	(0.0046)	(0.0063)	(0.0032)	(0.0054)	(0.0081)
PANDA	0.0512	0.1374	0.2266	0.0514	0.1384	0.2292	
		(0.0026)	(0.0040)	(0.0064)	(0.0025)	(0.0048)	(0.0088)
Erdős-Rényi Random Graph	LPD	0.2857	0.2424	0.1150	0.2757	0.3256	0.3289
		(0.0138)	(0.0099)	(0.0054)	(0.0148)	(0.0182)	(0.0145)
	AdaLDA	0.2849	0.2414	0.1162	0.2758	0.3246	0.3281
		(0.0129)	(0.090)	(0.0058)	(0.0138)	(0.0185)	(0.0152)
PANDA	<b>0.2823</b>	0.2403	<b>0.1114</b>	<b>0.2721</b>	<b>0.3183</b>	<b>0.3209</b>	
		(0.0117)	(0.0106)	(0.0044)	(0.0129)	(0.0166)	(0.0161)
Block Sparse	LPD	0.1643	0.0954	0.0451	0.4184	0.1724	0.3776
		(0.0056)	(0.0038)	(0.0028)	(0.0170)	(0.0029)	(0.0077)
	AdaLDA	0.1745	0.1002	0.0451	0.4378	0.1739	0.3811
		(0.0061)	(0.0009)	(0.0003)	(0.0156)	(0.0007)	(0.0020)
PANDA	<b>0.1614</b>	<b>0.0938</b>	<b>0.0437</b>	<b>0.4168</b>	<b>0.1706</b>	<b>0.3753</b>	
		(0.0061)	(0.0009)	(0.0003)	(0.0156)	(0.0007)	(0.0020)

Continued on next page

**Table 1.3 – continued from previous page**

Setting	Specification						
		(0.0047)	(0.0018)	(0.0007)	(0.0159)	(0.0026)	(0.0072)
	$p$	400	800	1200			
Approximately Sparse	LPD	0.1054	0.1047	0.1053			
		(0.0046)	(0.0030)	(0.0040)			
	AdaLDA	0.1042	0.1043	0.1042			
		(0.0029)	(0.0035)	(0.0038)			
	PANDA	0.1034	0.1039	0.1040			
		(0.0033)	(0.0038)	(0.0045)			

Table 1.4: Running time (in seconds) of the PANDA and AdaLDA methods under the Varying Diagonal model using Gurobi. The standard deviations are given in brackets.

$(s, p)$	(5, 400)	(10, 400)	(20, 400)
AdaLDA	106.739 (2.653)	107.743 (2.588)	107.017 (2.782)
PANDA	70.202 (4.751)	71.312 (4.389)	72.112 (4.965)
$(s, p)$	(5, 800)	(10, 800)	(20, 800)
AdaLDA	413.262 (13.209)	413.876 (12.708)	416.793 (12.383)
PANDA	325.486 (16.372)	326.125 (16.504)	333.427 (13.554)

### 1.5.2 Leukemia data

We investigate the performance of the PANDA, LPD, and AdaLDA methods on a Leukemia dataset from high-density oligonucleotide microarrays. This dataset was first analyzed by [18], and it contains 72 samples of two categories: 47 of acute lymphoblastic leukemia (ALL), and 25 of acute myeloid leukemia (AML). Each sample contains the quantitative expression levels of 7129 genes.

**Preprocessing:** We follow the preprocessing steps in [13]. First, we combine the data from both categories and compute the sample variance of each gene. Then, we drop the genes with sample variance beyond the lower and upper 6-quantiles of the total 7129 genes.

**Result:** To provide a fair comparison among the LPD, AdaLDA, and PANDA methods, we tune the parameters using a validation set. After preprocessing the raw data, we randomly split the data into training, validation, and testing sets. Specifically, the training set contains 29 ALL and 15 AML samples, the validation set contains 9 ALL and 5 AML samples, and the testing set contains 9 ALL and 5 AML samples. For the computational efficiency, we only use 2000 genes with the largest absolute values of the two-sample  $t$ -test in the training set, as suggested by [13]. We repeat the process 100 times, and provide the three methods' average misclassification rates on the testing set (testing error) and their standard deviations in Table 1.5. As can be seen, the PANDA method achieves a lower misclassification rate than both the LPD and AdaLDA methods.

Table 1.5: *The performance of PANDA, AdaLDA and LPD on the Leukemia dataset. The testing errors are averaged over 100 repetitions. The standard deviation of the testing errors are given in brackets. The difference between PANDA and the other two methods is significant by pair-wise  $t$ -test with a  $p$ -value less than 0.001.*

	LPD	AdaLDA	PANDA
Testing Error	9.28% (6.87%)	10.64% (7.92%)	6.93% (6.74%)

## 1.6 Extension to multiple-class LDA

In this section, we discuss the extension of PANDA method to  $K$ -class LDA in high dimensions. To be more specific, we consider the following data setting. Suppose we have samples  $\left\{ X_i^{(k)} : k = 1, 2, \dots, K, i = 1, 2, \dots, n_k \right\}$  from  $K$  classes denoted by  $k = 1, 2, \dots, K$ , such that  $X_i^{(k)}$ 's are i.i.d. from  $N(\mu^{(k)}, \Sigma)$ . Also, we suppose that the prior probabilities  $\pi_1, \pi_2, \dots, \pi_K$  for the  $K$  classes are known. Then the oracle classification rule for future data  $Z$  is given by  $f(Z) = \operatorname{argmax}_k D_k$ , where  $D_1 = 0$ ,  $D_k = \left( Z - \frac{\mu^{(1)} + \mu^{(k)}}{2} \right)^\top \beta^{(k)} + \log \left( \frac{\pi_k}{\pi_1} \right)$ , with  $\beta^{(k)} = \Sigma^{-1}(\mu^{(k)} - \mu^{(1)})$ . In addition, we define  $\Delta_k = \sqrt{\beta^{(k)\top} \Sigma \beta^{(k)}}$ . Let  $\hat{\mu}^{(k)}$  be the sample mean of data in class  $k$ , and let  $\hat{\Sigma}$  be the pooled sample covariance matrix over the  $K$  classes. Then, one can construct the classifier

by using the  $K$ -class PANDA method, which simultaneously estimate  $\beta^{(k)}$ 's and  $\Delta_k$ 's via the following optimization problems.

$$\begin{aligned} (\widehat{\beta}^{(k)}, \widehat{\tau}^k) \in \arg \min_{\beta, \tau} \quad & \|\beta\|_1 + c_k \tau^2, \\ \text{subject to} \quad & \|\widehat{\Sigma}\beta - (\widehat{\mu}^{(k)} - \widehat{\mu}^{(1)})\|_\infty \leq \lambda \widehat{\sigma}_{\max}(\tau + 1), \quad \sqrt{\beta^\top \widehat{\Sigma} \beta} \leq \tau. \end{aligned} \quad (1.19)$$

Based on  $\widehat{\beta}^{(k)}$ 's, one can construct the classifier by  $\widehat{f}(Z) = \arg \max_k \widehat{D}_k$  with  $\widehat{D}_1 = 0$  and  $\widehat{D}_k = (Z - \frac{\widehat{\mu}^{(1)} + \widehat{\mu}^{(k)}}{2})^\top \widehat{\beta}^{(k)}$ .

Following the similar technical argument as for Theorems 1.4.1, 1.4.2 and 1.4.3, we can establish the following theoretical properties for  $K$ -class PANDA method.

**Theorem 1.6.1** *Suppose that Assumption 1.4.2 hold, and  $\beta^{(k)*} \in \mathbb{B}_q(R)$  for some  $q \in [0, 1)$  and some  $R > 0$  for all  $k = 2, 3, \dots, K$ . Let  $(\widehat{\beta}^{(k)}, \widehat{\tau}_k)$  be an optimal solution of (1.19). Given*

$$c_k = \frac{1}{8 \left( \|\widehat{\mu}^{(k)} - \widehat{\mu}^{(1)}\|_\infty + 4\widehat{\sigma}_{\max} \sqrt{\frac{\log p}{n}} \right)}, \quad \lambda = 20 \sqrt{\frac{\log p}{n}}, \quad (1.20)$$

for sufficiently large  $n$  such that

$$n \geq C \cdot a^{-2} \Delta^2 \sigma_{\max}^2 M^{2 + \frac{1}{1-q}} R^{\frac{2}{1-q}} \log p \quad (1.21)$$

where  $C$  is an absolute constant, we have, with probability goes to 1,

$$\|\widehat{\beta}^{(k)} - \beta^{(k)*}\|_1 \leq C_1 \cdot (\Delta_k + 1) (\sigma_{\max} M)^{1-q} R \left( \frac{\log p}{n} \right)^{(1-q)/2}, \quad (1.22a)$$

$$\|\widehat{\beta}^{(k)} - \beta^{(k)*}\|_2 \leq C_2 \cdot (\Delta_k + 1) (\sigma_{\max} M)^{1-q/2} \sqrt{R} \left( \frac{\log p}{n} \right)^{1/2-q/4}, \quad (1.22b)$$

$$\frac{|\widehat{\tau}_k^2 - \Delta_k^2|}{\Delta_k^2} \leq C_3 \cdot (1 + \Delta_k^{-1}) \sigma_{\max}^{1-q/2} M^{3/2-q} R \left( \frac{\log p}{n} \right)^{(1-q)/2}, \quad (1.22c)$$

where  $C_1, C_2$  and  $C_3$  are positive constants.

## 1.7 Proofs of the Main Results

In this section, we provide the proof for Theorem 1.4.1 in Section 1.7.1 and Theorem 1.4.3 in Section 1.7.3. The proofs of lemmas can be found in the supplementary material.

### 1.7.1 Proof of Theorem 1

**Proof** We denote by  $\delta = \widehat{\beta} - \beta^*$  and  $\tau^* = \sqrt{\beta^{*\top} \widehat{\Sigma} \beta^*}$ . We first derive the upper bound for  $\|\delta\|_1$ . Based on this upper bound, we then derive the upper bounds for  $\|\delta\|_2$  and  $\widehat{\tau}$ .

For ease of presentation, we first define the following events,

$$\mathcal{E}_\tau = \left\{ |\beta^{*\top} (\widehat{\Sigma} - \Sigma) \beta^*| \leq \frac{1}{2} \beta^{*\top} \Sigma \beta^* \right\} = \left\{ \frac{1}{2} \Delta^2 \leq \tau^{*2} \leq \frac{3}{2} \Delta^2 \right\}, \quad (1.23)$$

$$\mathcal{E}_{\sigma_{\max}} = \left\{ |\widehat{\sigma}_{\max}^2 - \sigma_{\max}^2| \leq \frac{1}{2} \sigma_{\max}^2 \right\}, \quad (1.24)$$

$$\mathcal{E}_{\mu_d} = \left\{ \|\mu_d\|_\infty - 2\sqrt{2}\sigma_{\max} \sqrt{\frac{\log p}{n}} \leq \|\widehat{\mu}_d\|_\infty \leq \|\mu_d\|_\infty + 2\sqrt{2}\sigma_{\max} \sqrt{\frac{\log p}{n}} \right\}, \quad (1.25)$$

$$\mathcal{E}_1 = \left\{ \|(\widehat{\Sigma} - \Sigma) \beta^*\|_\infty \leq 10\sigma_{\max} \Delta \sqrt{\frac{\log p}{n}} \right\}, \quad (1.26)$$

$$\mathcal{E}_2 = \left\{ \|\widehat{\Sigma} \beta^* - \widehat{\mu}\|_\infty \leq 20\widehat{\sigma}_{\max} \sqrt{\frac{\log p}{n}} (\tau^* + 1) \right\}. \quad (1.27)$$

Before we proceed, we introduce the following lemma.

**Lemma 1.7.1** *For any  $\beta^* \in \mathbb{R}^p$ , we have*

$$\mathbb{P}(\mathcal{E}_\tau) \geq 1 - 2 \exp\left(-\frac{n-1}{16}\right), \quad \mathbb{P}(\mathcal{E}_{\sigma_{\max}}) \geq 1 - 2p \exp\left(-\frac{n-1}{16}\right),$$

$$\mathbb{P}(\mathcal{E}_{\mu_d}) \geq 1 - 2p^{-1}, \quad \mathbb{P}(\mathcal{E}_1) \geq 1 - 2p^{-1}.$$

Moreover, we have

$$\mathcal{E}_2 \supseteq \left( \mathcal{E}_\tau \cap \mathcal{E}_{\sigma_{\max}} \cap \mathcal{E}_1 \right).$$

**Upper bound for  $\|\delta\|_1$ .** We first provide an upper bound for  $\delta^\top \widehat{\Sigma} \delta$  in terms of  $\|\delta\|_1$ , which is essential for deriving an upper bound of  $\|\delta\|_1$ .

**Lemma 1.7.2** *Suppose that the events  $\mathcal{E}_\tau$ ,  $\mathcal{E}_{\sigma_{\max}}$ ,  $\mathcal{E}_1$  and  $\mathcal{E}_2$  hold. Then we have*

$$\delta^\top \widehat{\Sigma} \delta \leq 2\lambda\sigma_{\max}\|\delta\|_1 \left( 3\Delta + 2 + \sqrt{\frac{\|\delta\|_1}{c}} \right). \quad (1.28)$$

Our next step is to derive a lower bound for  $\delta^\top \widehat{\Sigma} \delta$  in terms of  $\|\delta\|_1$ , based on the restricted eigenvalue condition of  $\widehat{\Sigma}$  on certain restricted subset of  $\mathbb{R}^p$ . We first introduce the eigenvalue condition of  $\widehat{\Sigma}$  that holds with high probability.

**Lemma 1.7.3** *Suppose that Assumption 1.4.2 holds, and  $n \geq 2$ . There exist absolute positive constants  $c_1$  and  $c_2$  such that*

$$\delta^\top \widehat{\Sigma} \delta \geq \frac{1}{32M} \|\delta\|_2^2 - 81\sigma_{\max}^2 \frac{\log p}{n} \|\delta\|_1^2 \quad \text{for all } \delta \in \mathbb{R}^p, \quad (1.29)$$

with probability at least  $1 - c_1 \exp(-c_2 n)$ .

Based on the above result, we derive the restricted eigenvalue condition of  $\widehat{\Sigma}$  over a restricted subset. In particular, for  $S \subseteq [p]$  and  $\beta^* \in \mathbb{R}^p$ , we let

$$\mathcal{C}_{S,\beta^*} := \{\delta \in \mathbb{R}^p : \|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1 + 4\|\beta_{S^c}^*\|_1\}. \quad (1.30)$$

The next lemma shows that  $\delta \in \mathcal{C}_{S,\beta^*}$  for any  $S \subseteq [p]$ .

**Lemma 1.7.4** *Suppose that Assumption 1.4.1 and events  $\mathcal{E}_{\mu_d}$ ,  $\mathcal{E}_{\sigma_{\max}}$ ,  $\mathcal{E}_1$  and  $\mathcal{E}_2$  hold. Let  $S \subseteq [p]$ . Given  $c$  and  $\lambda$  in (1.13), we have  $\delta \in \mathcal{C}_{S,\beta^*}$  when  $n$  satisfies*

$$n \geq 100a^{-2}\sigma_{\max}^2\Delta^2 \log p.$$

Now, we choose a subset  $S_\eta$  that

$$S_\eta = \{j \in [p] : |\beta_j^*| \geq \eta\}, \quad (1.31)$$

$$\text{where } \eta = \sigma_{\max} M \sqrt{\frac{\log p}{n}}. \quad (1.32)$$

We further show the upper bounds for  $|S_\eta|$  and  $\|\beta_{S_\eta^c}^*\|_1$  in the next lemma.

**Lemma 1.7.5** *When  $\beta^* \in \mathbb{B}_q(R)$ , we have that*

$$|S_\eta| \leq \eta^{-q} R, \quad (1.33)$$

$$\|\beta_{S_\eta^c}^*\|_1 \leq \eta^{1-q} R. \quad (1.34)$$

Note that if  $S_\eta$  is empty, we immediately have that

$$\|\delta\|_1 \leq 4\|\beta_{S_\eta^c}^*\|_1 \leq 4\eta^{1-q} R = 4(\sigma_{\max} M)^{1-q} R \left(\frac{\log p}{n}\right)^{\frac{1-q}{2}},$$

which matches the upper bound in (1.22a).

When  $S_\eta$  is non-empty and  $\delta \in \mathcal{C}_{S_\eta, \beta^*}$ , we have that

$$\|\delta\|_1 \leq 4\|\delta_{S_\eta}\|_1 + 4\|\beta_{S_\eta^c}^*\|_1 \leq 4\sqrt{|S_\eta|}\|\delta\|_2 + 4\|\beta_{S_\eta^c}^*\|_1. \quad (1.35)$$

Plugging the above inequality into (1.29) yields that

$$\begin{aligned} \delta^\top \widehat{\Sigma} \delta &\geq \frac{1}{512M|S_\eta|} \left( \|\delta\|_1 - 4\|\beta_{S_\eta^c}^*\|_1 \right)^2 - 81\sigma_{\max}^2 \frac{\log p}{n} \|\delta\|_1^2 \\ &\geq \left( \frac{1}{512M|S_\eta|} - 81\sigma_{\max}^2 \frac{\log p}{n} \right) \|\delta\|_1^2 - \frac{\|\beta_{S_\eta^c}^*\|_1}{64M|S_\eta|} \|\delta\|_1. \end{aligned}$$

When  $n$  satisfies that

$$n \geq C \cdot \sigma_{\max}^2 M |S_\eta| \log p$$

for some constant  $C$ , we have that

$$\delta^\top \widehat{\Sigma} \delta \geq \frac{1}{1024M|S_\eta|} \|\delta\|_1^2 - \frac{\|\beta_{S_\eta^c}^*\|_1}{64M|S_\eta|} \|\delta\|_1. \quad (1.36)$$

Combining (1.28) with (1.36), we have that

$$\frac{1}{1024M|S_\eta|} \|\delta\|_1^2 - \frac{\|\beta_{S_\eta^c}^*\|_1}{64M|S_\eta|} \|\delta\|_1 \leq 2\lambda\sigma_{\max} \|\delta\|_1 \left( 3\Delta + 2 + \sqrt{\frac{\|\delta\|_1}{c}} \right).$$

Solving the above inequality with our chosen  $c$ ,  $\lambda$  and  $\eta$  as in (1.13) and (1.32), and using the upper bounds (1.33) and (1.34), we have the upper bound for  $\|\delta\|_1$  that

$$\|\delta\|_1 \leq C \cdot (\sigma_{\max} M)^{1-q} (\Delta + 1) R \left( \frac{\log p}{n} \right)^{\frac{1-q}{2}} \quad (1.37)$$

for some constant  $C$ , given  $n$  satisfies that

$$n \geq C \cdot a^{-2} \Delta^2 \sigma_{\max}^2 M^{2+\frac{1}{1-q}} R^{\frac{2}{1-q}} \log p$$

for some constant  $C$ .

**Upper bound for  $\|\delta\|_2$ .** We prove (1.22b) based on the previous upper bound for  $\|\delta\|_1$ .

Following Lemma 1.7.3, there exist some absolute positive constants  $c_1$  and  $c_2$  such that, with probability at least  $1 - c_1 \exp(-c_2 n)$ , we have

$$\delta^\top \widehat{\Sigma} \delta \geq \frac{1}{32M} \|\delta\|_2^2 - 81\sigma_{\max}^2 \frac{\log p}{n} \|\delta\|_1^2. \quad (1.38)$$

The above inequality gives an upper bound of  $\|\delta\|_2^2$  in terms of  $\delta^\top \widehat{\Sigma} \delta$  and  $\|\delta\|_1$ , whereas the latter two terms can be further upper bounded using Lemma 1.7.2 and (1.37), respectively.

To bound  $\delta^\top \widehat{\Sigma} \delta$ , following Lemma 1.7.2, we have that

$$\delta^\top \widehat{\Sigma} \delta \leq \lambda \widehat{\sigma}_{\max} \|\delta\|_1 \left( \sqrt{\frac{\|\delta\|_1}{c}} + 3\Delta + 2 \right). \quad (1.39)$$

Note that  $\Delta = \sqrt{\mu_d^\top \Sigma^{-1} \mu_d} \geq M^{-1/2} \|\mu_d\|_\infty$ , and thus  $\|\mu_d\|_\infty / \Delta \leq M^{1/2}$ . Hence (1.38) and (1.39) together imply that

$$\begin{aligned} \|\delta\|_2^2 &\leq C \cdot \left[ M \delta^\top \widehat{\Sigma} \delta + \sigma_{\max}^2 M \frac{\log p}{n} \|\delta\|_1^2 \right] \\ &\leq C \cdot \left[ \lambda \widehat{\sigma}_{\max} M (\Delta + 1) \|\delta\|_1 + \frac{\lambda \widehat{\sigma}_{\max}}{\sqrt{c}} M \|\delta\|_1^{3/2} + \sigma_{\max}^2 M \frac{\log p}{n} \|\delta\|_1^2 \right] \end{aligned} \quad (1.40)$$

for some constant  $C$ . By our choice of  $c$  and  $\lambda$  in (1.13) and the upper bound of  $\|\delta\|_1$  in (1.22a), when  $n$  satisfies that

$$n \geq C \cdot \sigma_{\max}^2 M^{2+\frac{1}{1-q}} R^{\frac{2}{1-q}} \log p$$

for some absolute constant  $C$ , (1.40) reduces to

$$\|\delta\|_2^2 \leq C \cdot \lambda \widehat{\sigma}_{\max} M (\Delta + 1) \|\delta\|_1 \leq C \cdot (\sigma_{\max} M)^{2-q} (\Delta + 1)^2 R \left( \frac{\log p}{n} \right)^{1-q/2},$$

which shows (1.22b) holds.

**Upper bound of  $|\widehat{\tau}^2 - \Delta^2|/\Delta^2$ .** Note that  $|\widehat{\tau}^2 - \Delta^2| \leq |\widehat{\tau}^2 - \tau^{*2}| + |\tau^{*2} - \Delta^2|$ . We upper bound the two terms on the right-hand side respectively in the next lemma.

**Lemma 1.7.6** *Suppose that Assumption 1.4.2, events  $\mathcal{E}_\tau, \mathcal{E}_{\sigma_{\max}}, \mathcal{E}_{\mu_d}, \mathcal{E}_1$  and (1.22a) hold.*

*When  $n$  satisfies (1.14) for some absolute constant  $C$ , we have that*

$$|\widehat{\tau}^2 - \tau^{*2}| \leq C \cdot \Delta (\Delta + 1) \sigma_{\max}^{1-q/2} M^{(3-q)/2} R \left( \frac{\log p}{n} \right)^{(1-q)/2} \quad (1.41a)$$

$$|\tau^{*2} - \Delta^2| \leq C \cdot \Delta^2 \sigma_{\max}^{1-q/2} M^{(1-q)/2} \sqrt{R} \left( \frac{\log p}{n} \right)^{(2-q)/4} \quad (1.41b)$$

for some absolute constant  $C$ .

Combining (1.41a) and (1.41b), we obtain that

$$\frac{|\widehat{\tau}^2 - \Delta^2|}{\Delta^2} \leq C \cdot (1 + \Delta^{-1}) \sigma_{\max}^{1-q/2} M^{\frac{3-q}{2}} R \left( \frac{\log p}{n} \right)^{\frac{1-q}{2}}$$

for some absolute constant  $C$ , and our claim (1.22c) follows as desired.  $\blacksquare$

### 1.7.2 Proof of Theorem 2

**Proof** We first introduce the following lemma that gives a different upper bound of  $\delta^\top \widehat{\Sigma} \delta$  as in Lemma 1.7.2, with the additional condition that  $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta}}$ .

**Lemma 1.7.7** *Suppose that the events  $\mathcal{E}_\tau$ ,  $\mathcal{E}_1$  and  $\mathcal{E}_2$  hold, and  $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta}}$ . Then we have*

$$\begin{aligned} \delta^\top \widehat{\Sigma} \delta \leq & C \cdot \lambda \sigma_{\max} \|\delta\|_1 \left\{ \lambda \sigma_{\max} \|\delta\|_1 + \tau^* + 1 + \left( 20 \sigma_{\max} \Delta \sqrt{\frac{\log p}{n}} \|\delta\|_1 \right)^{1/2} \right. \\ & \left. + (2 \|\mu_d\|_2 \|\delta\|_2)^{1/2} \right\}, \end{aligned} \quad (1.42)$$

for some constant  $C$ .

**Upper bound of  $\|\delta\|_2$ .** Based on Lemma 1.7.3 in the previous part, with probability goes to 1 we have that

$$\delta^\top \widehat{\Sigma} \delta \geq \frac{1}{32M} \|\delta\|_2^2 - 81 \sigma_{\max}^2 \frac{\log p}{n} \|\delta\|_1^2 \quad \text{for all } \delta \in \mathbb{R}^p.$$

When  $\delta \in \mathcal{C}_{S_\eta, \beta^*}$ , combining the above equation with (1.42), and using (1.35), we have that

$$\frac{1}{M} \|\delta\|_2^2 \leq C \cdot \left[ \sigma_{\max}^2 \frac{\log p}{n} \|\beta_{S_\eta^c}^*\|_1^2 + \lambda \sigma_{\max} \left( \sqrt{|S_\eta|} \|\delta\|_2 + \|\beta_{S_\eta^c}^*\|_1 \right) \left( \Delta + 1 + \sqrt{\|\mu_d\|_2 \|\delta\|_2} \right) \right]$$

for some constant  $C$ , when  $n$  satisfies that

$$n \geq C \cdot \sigma_{\max}^2 M^{\frac{2-2q}{2-q}} R^{\frac{2}{2-q}} \log p$$

for some constant  $C$ . By setting  $\eta$  as in (1.32), and using (1.33) and (1.34), we finally obtain

$$\|\delta\|_2 \leq C \cdot (\sigma_{\max} M)^{1-q/2} (\Delta + 1) \sqrt{R} \left( \frac{\log p}{n} \right)^{1/2-q/4} \quad (1.43)$$

for some constant  $C$ .

**Upper bound of  $|\widehat{\tau}^2 - \Delta^2|/\Delta^2$ .** Note that  $|\widehat{\tau}^2 - \Delta^2| \leq |\widehat{\tau}^2 - \tau^{*2}| + |\tau^{*2} - \Delta^2|$ . In Lemma 1.7.6, we have already shown the upper bound for the term  $|\tau^{*2} - \Delta^2|$  as (1.41b), which we also adopt here. With the additional condition that  $\sqrt{\widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta}} = \widehat{\tau}$ , the upper bound of the term  $|\widehat{\tau}^2 - \tau^{*2}|$  can be tighter than (1.41a), as shown in the following lemma.

**Lemma 1.7.8** *Suppose that Assumption 1.4.2, events  $\mathcal{E}_\tau, \mathcal{E}_{\sigma_{\max}}, \mathcal{E}_{\mu_d}, \mathcal{E}_1$  and (1.22a) hold. Also, suppose that  $\sqrt{\widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta}} = \widehat{\tau}$ . When  $n$  satisfies (1.16), we have that*

$$|\widehat{\tau}^2 - \tau^{*2}| \leq C \cdot \Delta (\Delta + 1) \sigma_{\max}^{1-q/2} M^{(3-q)/2} \sqrt{R} \left( \frac{\log p}{n} \right)^{1/2-q/4}. \quad (1.44)$$

for some absolute constant  $C$ .

Combining (1.41b) and (1.44), we have that

$$\frac{|\widehat{\tau}^2 - \Delta^2|}{\Delta^2} \leq C \cdot (1 + \Delta^{-1}) \sigma_{\max}^{1-q/2} M^{(3-q)/2} \sqrt{R} \left( \frac{\log p}{n} \right)^{1/2-q/4}.$$

for some constant  $C$ . ■

### 1.7.3 Proof of Theorem 3

**Proof** Let  $\widehat{\Delta} = \sqrt{\widehat{\beta}^\top \Sigma \widehat{\beta}}$ . The misclassification risk of  $\widehat{\beta}$  is

$$\mathcal{R}(\widehat{\beta}) = \frac{1}{2} \Phi \left( -\frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) + \frac{1}{2} \Phi \left( \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right), \quad (1.45)$$

where  $\Phi(\cdot)$  is the CDF of the standard Gaussian distribution. Recall that the optimal risk achieved by Fisher's rule is  $\mathcal{R}^* = \Phi(-\frac{\Delta}{2})$ . For the first term on the right-hand side of (2.31), its second order Taylor's expansion is

$$\begin{aligned} \Phi \left( -\frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) &= \Phi \left( -\frac{\Delta}{2} \right) + \Phi' \left( -\frac{\Delta}{2} \right) \left( \frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) \\ &\quad + \frac{\Phi''(t_1)}{2} \left( \frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2, \end{aligned} \quad (1.46)$$

where  $t_1 \in \left( -\frac{\Delta}{2}, -\frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)$ . Similarly, for the second term in (2.31), we have

$$\begin{aligned} \Phi \left( \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) &= \Phi \left( -\frac{\Delta}{2} \right) + \Phi' \left( -\frac{\Delta}{2} \right) \left( \frac{\Delta}{2} + \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) \\ &\quad + \frac{\Phi''(t_2)}{2} \left( \frac{\Delta}{2} + \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2, \end{aligned} \quad (1.47)$$

where  $t_2 \in \left( -\frac{\Delta}{2}, \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)$ . Combining (2.32) and (2.33), we have

$$\begin{aligned} \mathcal{R}(\widehat{\beta}) - \mathcal{R}^* &= \Phi' \left( -\frac{\Delta}{2} \right) \left( \frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \right) + \frac{\Phi''(t_1)}{2} \left( \frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2 \\ &\quad + \frac{\Phi''(t_2)}{2} \left( \frac{\Delta}{2} + \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2. \end{aligned} \quad (1.48)$$

We now introduce a lemma that upper bounds the first term on the right-hand side of (2.34).

**Lemma 1.7.9** *Suppose (1.22b) holds, and  $n$  satisfies that*

$$n \geq C \cdot \sigma_{\max}^2 M^{2+2/(2-q)} R^{2/(2-q)} \log p$$

for some constant  $C$ . Then we have

$$\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \leq \frac{M}{2\Delta} \|\delta\|_2^2, \quad (1.49)$$

Note that  $\Phi'(-\Delta/2) = (2\pi)^{-1/2} \exp(-\Delta^2/8)$ . Following Lemma 2.6.7, we have

$$\Phi' \left( -\frac{\Delta}{2} \right) \left( \frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \right) \leq \frac{M}{2\sqrt{2\pi}\Delta} \exp \left( -\frac{\Delta^2}{8} \right) \|\delta\|_2^2. \quad (1.50)$$

Now we consider the second-order term in (2.34). First, using Lemma 2.6.7, we have

$$\begin{aligned} \frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} &= \frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} + \frac{\widehat{\beta}^\top (\mu_m - \widehat{\mu}_m)}{\widehat{\Delta}} \\ &\leq \frac{M}{2\Delta} \|\delta\|_2^2 + \frac{\widehat{\beta}^\top (\mu^{(0)} - \widehat{\mu}^{(0)}) + \widehat{\beta}^\top (\mu^{(1)} - \widehat{\mu}^{(1)})}{2\widehat{\Delta}}. \end{aligned} \quad (1.51)$$

After taking square, the first term on the right-hand side gives  $\frac{M}{4\Delta^2} \|\delta\|_2^4$ , which is negligible compared to the first-order term. Hence it suffices to bound the second term on the right-hand side of (2.39). For this aim we introduce the next lemma.

**Lemma 1.7.10** *Under the identical conditions as in Theorem 1.4.1 or 1.4.2, with probability at least  $1 - 4p^{-1}$  we have*

$$\left( \frac{\widehat{\beta}^\top (\mu^{(0)} - \widehat{\mu}^{(0)}) + \widehat{\beta}^\top (\mu^{(1)} - \widehat{\mu}^{(1)})}{2\widehat{\Delta}} \right)^2 \leq C \cdot \sigma_{\max}^{-q} M^{1-q} R \left( \frac{\log p}{n} \right)^{1-q/2} \quad (1.52)$$

for some constant  $C$ .

Since  $t_1 > -\Delta/2$ , we have  $|\Phi''(t_1)| \leq C \cdot \Delta \exp(-\Delta^2/8)$ . Combining this with (2.40),

we bound the second term in (2.34) by

$$\frac{|\Phi''(t_1)|}{2} \left( \frac{\Delta}{2} - \frac{(\hat{\mu}_m - \mu^{(0)})^\top \hat{\beta}}{\hat{\Delta}} \right)^2 \leq C \cdot \Delta \exp \left( -\frac{\Delta^2}{8} \right) \sigma_{\max}^{-q} M^{1-q} R \left( \frac{\log p}{n} \right)^{1-q/2} \quad (1.53)$$

for some constant  $C$ . Likewise, the third term in (2.34) is also subject to this bound.

Finally, plugging (2.38) and (2.41) into (2.34), and using (1.22b), we achieve that

$$\mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta^*) \leq C \cdot \exp \left( -\frac{\Delta^2}{8} \right) \sigma_{\max}^{-q} M^{3-q} \Delta R \left( \frac{\log p}{n} \right)^{1-q/2}$$

for some constant  $C$ , which completes the proof. ■

## 1.8 Proof of Lemmas

In this section, we provide additional detailed proof to the lemmas in the previous section of this chapter.

### 1.8.1 Proofs

This section provides the detailed proofs to the lemmas in the main body of the paper, and is split into eight subsections, one subsection for the proof of each lemma.

#### *Proof of Lemma 1.7.1*

**Proof** There are four main statements in Lemma 1.7.1, and let us prove them one by one.

(i) It suffices to show that

$$\mathbb{P}(\mathcal{E}_\tau) = \mathbb{P} \left( |\beta^{*\top} (\hat{\Sigma} - \Sigma) \beta^*| \leq \frac{1}{2} \beta^{*\top} \Sigma \beta^* \right) \geq 1 - 2e^{-(n-1)/16}. \quad (1.54)$$

Let  $\{Y_i\}_{i=1}^{2n-2}$  be i.i.d. random vectors following the multivariate normal distribution  $N(\mathbf{0}, \Sigma)$ . Then

$$\widehat{\Sigma} \stackrel{d}{=} \frac{1}{2n-2} \sum_{i=1}^{2n-2} Y_i Y_i^\top, \quad \text{and} \quad \beta^{*\top} \widehat{\Sigma} \beta^* \stackrel{d}{=} \frac{1}{2n-2} \sum_{i=1}^{2n-2} (\beta^{*\top} Y_i)^2,$$

where  $\stackrel{d}{=}$  denotes equal in distribution. Note that  $\{\beta^{*\top} Y_i\}$  are i.i.d Gaussian r.v.s following distribution  $N(0, \beta^{*\top} \Sigma \beta^*)$ , thus  $\{(\beta^{*\top} Y_i)^2\}$  are i.i.d. sub-exponential r.v.s, so for any  $t \in (0, \beta^{*\top} \Sigma \beta^*)$ , we have

$$\mathbb{P} \left( \left| \frac{1}{2n-2} \sum_i (\beta^{*\top} Y_i)^2 - \beta^{*\top} \Sigma \beta^* \right| \geq t \right) \leq 2 \exp \left\{ -\frac{(2n-2)t^2}{8(\beta^{*\top} \Sigma \beta^*)^2} \right\}.$$

Relation (2.42) follows directly by taking  $t = \frac{1}{2} \beta^{*\top} \widehat{\Sigma} \beta^*$ , and thus part (i) of Lemma 1.7.1 holds.

(ii) Now we need to show that

$$\mathbb{P}(\mathcal{E}_{\sigma_{\max}}) = \mathbb{P} \left( |\widehat{\sigma}_{\max}^2 - \sigma_{\max}^2| \leq \frac{1}{2} \sigma_{\max}^2 \right) \geq 1 - 2pe^{-(n-1)/16}.$$

To prove this, we set  $\beta^* = e_j$  for  $j \in [p]$  and use (2.42) with a union bound argument to obtain that

$$\mathbb{P} \left( |\widehat{\Sigma}_{j,j} - \Sigma_{j,j}| \leq \frac{1}{2} \Sigma_{j,j}, \forall j \in [p] \right) \geq 1 - 2pe^{-(n-1)/16}, \quad (1.55)$$

where the event on the left-hand side implies that  $|\widehat{\sigma}_{\max}^2 - \sigma_{\max}^2| \leq \frac{1}{2} \sigma_{\max}^2$ .

(iii) Here it suffices to show that

$$\mathbb{P} \left( \|\mu_d\|_\infty - 2\sqrt{2}\sigma_{\max} \sqrt{\frac{\log p}{n}} \leq \|\widehat{\mu}_d\|_\infty \leq \|\mu_d\|_\infty + 2\sqrt{2}\sigma_{\max} \sqrt{\frac{\log p}{n}} \right) \geq 1 - 2p^{-1}.$$

Notice that  $\widehat{\mu}_d \sim N(\mu_d, \frac{2}{n}\Sigma)$ . Let  $\mu_{d,j}$  and  $\widehat{\mu}_{d,j}$  denote the  $j$ -th coordinate of  $\mu_d$  and

$\widehat{\mu}_d$ , respectively. We have  $\widehat{\mu}_{d,j} \sim N(\mu_{d,j}, \frac{2}{n}\Sigma_{j,j})$ . Therefore, for any  $j \in [p]$  we have that

$$\mathbb{P}(|\widehat{\mu}_{d,j} - \mu_{d,j}| > t) \leq 2 \exp \left\{ \frac{-nt^2}{4(\Sigma_{j,j})^2} \right\} \leq 2 \exp \left\{ -\frac{nt^2}{4\sigma_{\max}^2} \right\}.$$

Taking  $t = \sigma_{\max} \sqrt{\frac{8 \log p}{n}}$  and applying the union bound for all  $j \in [p]$ , we have with probability at least  $1 - 2p^{-1}$  that

$$|\widehat{\mu}_{d,j} - \mu_{d,j}| \leq \sigma_{\max} \sqrt{\frac{8 \log p}{n}}, \quad \forall j \in [p],$$

which implies that  $|\|\widehat{\mu}_d\|_{\infty} - \|\mu_d\|_{\infty}| \leq 2\sqrt{2}\sigma_{\max} \sqrt{\log p/n}$ .

(iv) The lower bound of  $\mathbb{P}(\mathcal{E}_1)$  follows an argument in [13]. Since  $\beta^* = \Sigma^{-1}\mu_d$ , we have that  $\widehat{\Sigma}\beta^* - \widehat{\mu}_d = (\widehat{\Sigma} - \Sigma)\beta^* - (\widehat{\mu}_d - \mu_d)$ . By A.5.1 in the supplement of [13], we have that

$$\mathbb{P} \left( |e_j^{\top} (\widehat{\Sigma} - \Sigma)\beta^*| \leq 10\sigma_{\max} \Delta \sqrt{\frac{\log p}{n}}, \quad \forall j \in [p] \right) \geq 1 - 2p^{-1}, \quad (1.56)$$

where the event on the left-hand side is equivalent to event  $\mathcal{E}_1$ . Furthermore, recall that  $\Delta^2 = \beta^{*\top} \Sigma \beta^*$ . Therefore, under events  $\mathcal{E}_{\tau}$  and  $\mathcal{E}_{\sigma_{\max}}$ , we have  $\Delta \leq \sqrt{2}\tau^*$  and  $\sigma_{\max} \leq \sqrt{2}\widehat{\sigma}_{\max}$ . These two conditions and event  $\mathcal{E}_1$  together imply  $\mathcal{E}_2$ . ■

### *Proof of Lemma 1.7.2*

**Proof** When  $(\beta^*, \tau^*)$  is feasible to (1.6), from the first constraint of (1.6) we have

$$\|\widehat{\Sigma}\delta\|_{\infty} = \|\widehat{\Sigma}(\widehat{\beta} - \beta^*)\|_{\infty} \leq \|\widehat{\Sigma}\widehat{\beta} - \widehat{\mu}_d\|_{\infty} + \|\widehat{\Sigma}\beta^* - \widehat{\mu}_d\|_{\infty} \leq \lambda\widehat{\sigma}_{\max}(\widehat{\tau} + \tau^*) + 2\lambda\widehat{\sigma}_{\max}. \quad (1.57)$$

In addition, due to the optimality of  $(\widehat{\beta}, \widehat{\tau})$ , we have

$$\|\widehat{\beta}\|_1 + c\widehat{\tau}^2 \leq \|\beta^*\|_1 + c\tau^{*2},$$

which implies that

$$\widehat{\tau} \leq \tau^* + \sqrt{\frac{\|\delta\|_1}{c}}.$$

Plugging the above inequality into (1.57), we obtain that

$$\|\widehat{\Sigma}\delta\|_\infty \leq 2\lambda\widehat{\sigma}_{\max}(\tau^* + 1) + \lambda\widehat{\sigma}_{\max}\sqrt{\frac{\|\delta\|_1}{c}}.$$

Under the events  $\mathcal{E}_\tau$  and  $\mathcal{E}_{\sigma_{\max}}$ , we have  $\tau^* \leq \sqrt{\frac{3}{2}}\Delta$  and  $\widehat{\sigma}_{\max} \leq 2\sigma_{\max}$ , so we further have that

$$\|\widehat{\Sigma}\delta\|_\infty \leq 2\lambda\sigma_{\max} \left( 3\Delta + 2 + \sqrt{\frac{\|\delta\|_1}{c}} \right).$$

Finally, applying Hölder's inequality, we obtain that

$$\delta^\top \widehat{\Sigma}\delta \leq \|\delta\|_1 \|\widehat{\Sigma}\delta\|_\infty \leq 2\lambda\sigma_{\max} \|\delta\|_1 \left( 3\Delta + 2 + \sqrt{\frac{\|\delta\|_1}{c}} \right).$$

Thus Lemma 1.7.2 holds. ■

*Proof of Lemma 1.7.3*

**Proof** Lemma 1.7.3 is an application of a theorem in [19], which is given by the following lemma.

**Lemma 1.8.1 (Theorem 1 of [19])** *For any Gaussian random design  $Z \in \mathbb{R}^{n \times p}$  with i.i.d.*

$N(\mathbf{0}, \Sigma)$  laws, there exist absolute positive constants  $c_1, c_2$  such that

$$\frac{\|Z\delta\|_2}{\sqrt{n}} \geq \frac{1}{4} \|\Sigma^{1/2}\delta\|_2 - 9\sigma_{\max} \sqrt{\frac{\log p}{n}} \|\delta\|_1, \quad \forall \delta \in \mathbb{R}^p,$$

with probability at least  $1 - c_1 \exp(-c_2 n)$ .

Now we ready to prove Lemma 1.7.3. Suppose  $n \geq 2$ . Then the pooled covariance matrix  $\widehat{\Sigma}$  is obtained by

$$\widehat{\Sigma} = \frac{1}{2n-2} \left[ \sum_{i=1}^n (X_i^0 - \widehat{\mu}^{(0)}) (X_i^0 - \widehat{\mu}^{(0)})^\top + \sum_{i=1}^n (X_i^1 - \widehat{\mu}^{(1)}) (X_i^1 - \widehat{\mu}^{(1)})^\top \right],$$

and  $\widehat{\Sigma}$  has the same distribution as

$$\widetilde{\Sigma} = \frac{1}{2n-2} \sum_{i=1}^{2n-2} Z_i Z_i^\top,$$

where  $Z_j$ 's are i.i.d. samples from  $N(0, \Sigma)$ . Hence  $\widehat{\Sigma}$  can be viewed as the sample covariance matrix of a Gaussian random design with 0 mean.

By Lemma 2.6.9 (i.e., Theorem 1 of [19]), there exist absolute positive constants  $c_1$  and  $c_2$  such that with probability at least  $1 - c_1 \exp(-c_2 n)$ ,

$$\|\widehat{\Sigma}^{1/2}\delta\|_2 \geq \frac{1}{4} \|\Sigma^{1/2}\delta\|_2 - 9\sigma_{\max} \sqrt{\frac{\log p}{2n-2}} \|\delta\|_1.$$

When  $n \geq 2$  and  $\lambda_{\min} \geq M^{-1}$ , we have

$$\|\widehat{\Sigma}^{1/2}\delta\|_2 \geq \frac{1}{4\sqrt{M}} \|\delta\|_2 - 9\sigma_{\max} \sqrt{\frac{\log p}{n}} \|\delta\|_1,$$

and thus

$$\begin{aligned}\delta^\top \widehat{\Sigma} \delta &\geq \left( \frac{1}{4\sqrt{M}} \|\delta\|_2 - 9\sigma_{\max} \sqrt{\frac{\log p}{n}} \|\delta\|_1 \right)^2 \\ &\geq \frac{1}{32M} \|\delta\|_2^2 - 81\sigma_{\max}^2 \frac{\log p}{n} \|\delta\|_1^2.\end{aligned}$$

Here the last inequality follows from the fact that

$$(a - b)^2 = \left( \frac{1}{2}a^2 - 2ab + 2b^2 \right) + \frac{1}{2}a^2 - b^2 \geq \frac{1}{2}a^2 - b^2$$

for any number  $a, b \geq 0$ . Thus Lemma 1.7.3 holds. ■

*Proof of Lemma 1.7.4*

**Proof** For any  $S \subseteq [p]$ , we have

$$\|\widehat{\beta}\|_1 = \|\beta^* + \delta\|_1 \geq \|\beta_S^*\|_1 + \|\delta_{S^c}\|_1 - \|\beta_{S^c}^*\|_1 - \|\delta_S\|_1.$$

Combining the above inequality with  $\|\beta^*\|_1 \leq \|\beta_S^*\|_1 + \|\beta_{S^c}^*\|_1$ , we have

$$\|\widehat{\beta}\|_1 - \|\beta^*\|_1 \geq \|\delta_{S^c}\|_1 - \|\delta_S\|_1 - 2\|\beta_{S^c}^*\|_1. \quad (1.58)$$

When  $(\beta^*, \tau^*)$  is feasible to (1.6), by optimality we have

$$\|\widehat{\beta}\|_1 + c\widehat{\tau}^2 \leq \|\beta^*\|_1 + c\tau^{*2}. \quad (1.59)$$

Combining (1.58) and (1.59) yields

$$\|\delta_{S^c}\|_1 - \|\delta_S\|_1 - 2\|\beta_{S^c}^*\|_1 \leq \|\widehat{\beta}\|_1 - \|\beta^*\|_1 \leq c(\tau^{*2} - \widehat{\tau}^2). \quad (1.60)$$

Since  $\tau^{*2} = \beta^{*\top} \widehat{\Sigma} \beta^*$  and  $\widehat{\tau}^2 \geq \widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta}$ , it follows that

$$\begin{aligned} c(\tau^{*2} - \widehat{\tau}^2) &\leq -2c\delta^\top (\widehat{\Sigma} \beta^*) \\ &= -2c\delta^\top (\widehat{\Sigma} - \Sigma) \beta^* - 2c\delta^\top \mu_d \\ &\leq 2c\|(\widehat{\Sigma} - \Sigma) \beta^*\|_\infty \|\delta\|_1 + 2c\|\mu_d\|_\infty \|\delta\|_1. \end{aligned}$$

Under event  $\mathcal{E}_1$ , we have

$$\|(\widehat{\Sigma} - \Sigma) \beta^*\|_\infty \leq 10\sigma_{\max} \Delta \sqrt{\frac{\log p}{n}}.$$

When  $n$  satisfies that

$$n \geq 100a^{-2} \sigma_{\max}^2 \Delta^2 \log p,$$

we have

$$c(\tau^{*2} - \widehat{\tau}^2) \leq 4c\|\mu_d\|_\infty \|\delta\|_1.$$

By setting  $c$  as in (1.13), we have that

$$\frac{1}{2}\|\delta_{S^c}\|_1 \leq \frac{3}{2}\|\delta_S\|_1 + 2\|\beta_{S^c}^*\|_1.$$

Thus  $\|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1 + 4\|\beta^*\|_1$ , which completes the proof of Lemma 1.7.4. ■

*Proof of Lemma 1.7.5*

**Proof** From the definitions of  $\mathbb{B}_q(R)$  and  $S_\eta$ , we have that

$$R \geq \sum_j |\beta_j^*|^q \geq \eta^q |S_\eta|,$$

and

$$R \geq \sum_j |\beta_j^*|^q = \sum_j |\beta_j^*| \cdot |\beta_j^*|^{q-1} \geq \eta^{q-1} \|\beta_{S_\eta}^*\|_1.$$

Lemma 1.7.5 follows immediately from these two inequalities, and thus holds. ■

*Proof of Lemma 1.7.6*

**Proof** Let us first prove relation (1.41a). Under the optimality condition, we have  $\|\widehat{\beta}\|_1 + c\widehat{\tau}^2 \leq \|\beta^*\|_1 + c\tau^{*2}$ , and thus

$$\widehat{\tau}^2 - \tau^{*2} \leq \frac{1}{c} \|\delta\|_1 \leq C \cdot \Delta(\Delta + 1) \sigma_{\max}^{1-q} M^{3/2-q} R \left( \frac{\log p}{n} \right)^{\frac{1-q}{2}} \quad (1.61)$$

for some positive constant  $C$ . Here the last inequality uses (1.22a) and  $\|\mu_d\|_\infty \leq M^{1/2}\Delta$ .

Note that the second constraint in (1.6) implies that

$$\widehat{\tau}^2 \geq \widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta} = (\beta^* + \delta)^\top \widehat{\Sigma} (\beta^* + \delta) \geq \tau^{*2} + 2\delta^\top \widehat{\Sigma} \beta^*,$$

hence

$$\widehat{\tau}^2 - \tau^{*2} \geq -2|\delta^\top \widehat{\Sigma} \beta^*| \quad (1.62)$$

$$\geq -2 \left| \delta^\top \left[ (\widehat{\Sigma} - \Sigma) \beta^* + \mu_d \right] \right| \quad (1.63)$$

$$\geq -2\|\delta\|_2 \|\mu_d\|_2 - 2\|\delta\|_1 \|(\widehat{\Sigma} - \Sigma) \beta^*\|_\infty. \quad (1.64)$$

Note that under the event  $\mathcal{E}_1$ , we have

$$\|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty \leq 10\sigma_{\max}\Delta\sqrt{\frac{\log p}{n}}. \quad (1.65)$$

Plugging (1.22a), (1.22b), (1.65) and  $\|\mu_d\|_2 \leq M^{1/2}\Delta$  into (1.64), we obtain that

$$\widehat{\tau}^2 - \tau^{*2} \geq -C \cdot \Delta(\Delta + 1)\sigma_{\max}^{1-q/2}M^{(3-q)/2}\sqrt{R}\left(\frac{\log p}{n}\right)^{1/2-q/4}.$$

Combining the above equation and (1.61) yields (1.41a).

Next, let us prove the result (1.41b) in Lemma 1.7.6. Note that the gap between  $\tau^{*2}$  and  $\Delta^2$  can be written as  $|\tau^{*2} - \Delta^2| = |\beta^{*\top}(\widehat{\Sigma} - \Sigma)\beta^*|$ . To bound this gap, we first apply Hölder's inequality that

$$|\beta^{*\top}(\widehat{\Sigma} - \Sigma)\beta^*| \leq \|\beta^*\|_1\|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty.$$

Under event  $\mathcal{E}_1$ , the term  $\|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty$  can be again bounded by (1.65). To bound the term  $\|\beta^*\|_1$ , we note that

$$\|\beta^*\|_1 = \|\beta_{S_\eta}^*\|_1 + \|\beta_{S_\eta^c}^*\|_1 \leq \sqrt{|S_\eta|}\|\beta^*\|_2 + \|\beta_{S_\eta^c}^*\|_1 \leq \eta^{-q/2}\sqrt{R}M^{1/2}\Delta + \eta^{1-q}R.$$

The last inequality above uses equations (1.33) and (1.34). By our choice of  $\eta$  in (1.32), when  $n$  satisfies that

$$n \geq C \cdot \Delta^2\sigma_{\max}^2MR\log p$$

for some absolute constant  $C$ , we have that

$$\|\beta^*\|_1 \leq C \cdot \eta^{-q/2}\sqrt{R}M^{1/2}\Delta \leq C \cdot \Delta^2\sigma_{\max}^{-q/2}M^{(1-q)/2}\Delta\sqrt{R}\left(\frac{\log p}{n}\right)^{-q/4}.$$

Hence we have

$$|\tau^{*2} - \Delta^2| = |\beta^{*\top}(\widehat{\Sigma} - \Sigma)\beta^*| \leq \|\beta^*\|_1 \|(\widehat{\Sigma} - \Sigma)\beta^*\|_\infty \leq C \cdot \sigma_{\max}^{1-q/2} M^{(1-q)/2} \sqrt{R} \left(\frac{\log p}{n}\right)^{\frac{2-q}{4}},$$

and thus (1.41b) holds. ■

*Proof of Lemma 1.7.7*

**Proof** When  $(\beta^*, \tau^*)$  is feasible to (1.6), from the first constraint of (1.6) we have

$$\|\widehat{\Sigma}\delta\|_\infty = \|\widehat{\Sigma}(\widehat{\beta} - \beta^*)\|_\infty \leq \|\widehat{\Sigma}\widehat{\beta} - \widehat{\mu}_d\|_\infty + \|\widehat{\Sigma}\beta^* - \widehat{\mu}_d\|_\infty \leq \lambda\widehat{\sigma}_{\max}(\widehat{\tau} + \tau^*) + 2\lambda\widehat{\sigma}_{\max}. \quad (1.66)$$

When  $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta}}$ , we have

$$\begin{aligned} \widehat{\tau}^2 &= \widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta} = (\beta^* + \delta)^\top \widehat{\Sigma} (\beta^* + \delta) = \tau^{*2} + 2\delta^\top \widehat{\Sigma} \beta^* + \delta^\top \widehat{\Sigma} \delta \\ &= \tau^{*2} + \delta^\top \widehat{\Sigma} \delta + 2\delta^\top (\widehat{\Sigma} - \Sigma) \beta^* + 2\delta^\top \mu_d \\ &\leq \tau^{*2} + \delta^\top \widehat{\Sigma} \delta + 20\sigma_{\max} \Delta \sqrt{\frac{\log p}{n}} \|\delta\|_1 + 2\|\mu_d\|_2 \|\delta\|_2. \end{aligned}$$

Plugging the above inequality into (1.66), we have

$$\|\widehat{\Sigma}\delta\|_\infty \leq \lambda\widehat{\sigma}_{\max} \left[ 2\tau^* + 2 + \sqrt{\delta^\top \widehat{\Sigma} \delta} + \left( 20\sigma_{\max} \Delta \sqrt{\frac{\log p}{n}} \|\delta\|_1 \right)^{1/2} + (2\|\mu_d\|_2 \|\delta\|_2)^{1/2} \right].$$

Applying Hölder's inequality, we obtain that

$$\delta^\top \widehat{\Sigma} \delta \leq \lambda\widehat{\sigma}_{\max} \|\delta\|_1 \left[ 2\tau^* + 2 + \sqrt{\delta^\top \widehat{\Sigma} \delta} + \left( 20\sigma_{\max} \Delta \sqrt{\frac{\log p}{n}} \|\delta\|_1 \right)^{1/2} + (2\|\mu_d\|_2 \|\delta\|_2)^{1/2} \right].$$

From the above inequality, we may derive that

$$\begin{aligned} \delta^\top \widehat{\Sigma} \delta \leq & C \cdot \lambda \sigma_{\max} \|\delta\|_1 \left\{ \lambda \sigma_{\max} \|\delta\|_1 + \tau^* + 1 + \left( 20 \sigma_{\max} \Delta \sqrt{\frac{\log p}{n}} \|\delta\|_1 \right)^{1/2} \right. \\ & \left. + (2 \|\mu_d\|_2 \|\delta\|_2)^{1/2} \right\}, \end{aligned}$$

where  $C$  is a constant. ■

*Proof of Lemma 1.7.8*

**Proof** When  $\widehat{\tau} = \sqrt{\widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta}}$ , we have

$$\begin{aligned} \widehat{\tau}^2 &= \widehat{\beta}^\top \widehat{\Sigma} \widehat{\beta} = (\beta^* + \delta)^\top \widehat{\Sigma} (\beta^* + \delta) = \tau^{*2} + 2\delta^\top \widehat{\Sigma} \beta^* + \delta^\top \widehat{\Sigma} \delta \\ &= \tau^{*2} + \delta^\top \widehat{\Sigma} \delta + 2\delta^\top (\widehat{\Sigma} - \Sigma) \beta^* + 2\delta^\top \mu_d. \end{aligned}$$

With event  $\mathcal{E}_1$ , we have that

$$\begin{aligned} |\widehat{\tau}^2 - \tau^{*2}| &= |\delta^\top \widehat{\Sigma} \delta + 2\delta^\top (\widehat{\Sigma} - \Sigma) \beta^* + 2\delta^\top \mu_d| \\ &\leq \delta^\top \widehat{\Sigma} \delta + 20 \sigma_{\max} \Delta \sqrt{\frac{\log p}{n}} \|\delta\|_1 + 2 \|\mu_d\|_2 \|\delta\|_2. \end{aligned}$$

Then, using the previous results (1.35), (1.42) and (1.43), we obtain that

$$|\widehat{\tau}^2 - \tau^{*2}| \leq C \cdot \Delta (\Delta + 1) \sigma_{\max}^{1-q/2} M^{(3-q)/2} \sqrt{R} \left( \frac{\log p}{n} \right)^{1/2-q/4}$$

for some constant  $C$ . ■

*Proof of Lemma 2.6.7*

**Proof** Note that

$$\begin{aligned}\widehat{\Delta} &= \sqrt{\widehat{\beta}^\top \Sigma \widehat{\beta}} = \sqrt{\beta^\top \Sigma \beta + 2\beta^\top \Sigma \delta + \delta^\top \Sigma \delta} \\ &\leq \sqrt{\beta^\top \Sigma \beta} \left( 1 + \frac{2\beta^\top \Sigma \delta + \delta^\top \Sigma \delta}{2\beta^\top \Sigma \beta} \right) \\ &= \Delta + \frac{2\mu_d^\top \delta + \delta^\top \Sigma \delta}{2\Delta}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} &= \frac{1}{2\widehat{\Delta}} (\Delta \widehat{\Delta} - \mu_d^\top \widehat{\beta}) \\ &\leq \frac{1}{2\widehat{\Delta}} \left( \Delta^2 + \mu_d^\top (\delta - \widehat{\beta}) + \frac{1}{2} \delta^\top \Sigma \delta \right) \\ &= \frac{1}{4\widehat{\Delta}} \delta^\top \Sigma \delta \leq \frac{\delta^\top \Sigma \delta}{4(\Delta + \frac{\mu_d^\top \delta}{\Delta})}.\end{aligned}\tag{1.67}$$

Note that  $|\mu_d^\top \delta| \leq \|\mu_d\|_2 \|\delta\|_2 \leq M^{1/2} \Delta \|\delta\|_2$ . Using the convergence rate of  $\|\delta\|_2$  in (1.22b) from Theorem 1.4.1, when  $n$  satisfies that

$$n \geq C \cdot \sigma_{\max}^2 M^{2+2/(2-q)} R^{2/(2-q)} \log p$$

for some constant  $C$ , we have that  $|\mu_d^\top \delta| \leq \Delta^2/2$ , and thus it follows from (2.53) that

$$\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \leq \frac{\delta^\top \Sigma \delta}{2\Delta} \leq \frac{M}{2\Delta} \|\delta\|_2^2.$$

■

*Proof of Lemma 2.6.8*

**Proof** We first show that

$$\mathbb{P} \left( \|\widehat{\mu}^{(\ell)} - \mu^{(\ell)}\|_{\infty} \leq \sigma_{\max} \sqrt{\frac{2 \log p}{n}}, \ell = 0, 1 \right) \geq 1 - 4p^{-1}. \quad (1.68)$$

Note that  $\widehat{\mu}^{(\ell)} \sim N(\mu^{(\ell)}, \Sigma/n)$  for  $\ell = 0, 1$ , and thus  $\widehat{\mu}_j^{(\ell)} \sim N(\mu_j^{(\ell)}, \Sigma_{j,j}/n)$ , for  $j \in [p]$ .

Hence,

$$\mathbb{P} \left( |\widehat{\mu}_j^{(\ell)} - \mu_j^{(\ell)}| \geq t \right) \leq 2 \exp \left( -\frac{nt^2}{\Sigma_{j,j}} \right) \leq 2 \exp \left( -\frac{nt^2}{\sigma_{\max}^2} \right) \quad \text{for all } \ell \in \{0, 1\}, j \in [p].$$

Taking  $t = \sigma_{\max} \sqrt{2 \log p/n}$  and applying the union bound for all  $j \in [p]$ , we have

$$\mathbb{P} \left( \|\widehat{\mu}^{(\ell)} - \mu^{(\ell)}\|_{\infty} \leq \sigma_{\max} \sqrt{\frac{2 \log p}{n}}, \ell = 0, 1 \right) \geq 1 - 4p \exp(-2 \log p) = 1 - 4p^{-1}.$$

We next bound the term  $\widehat{\beta}^{\top}(\mu^{(\ell)} - \widehat{\mu}^{(\ell)})$  for  $\ell = 0, 1$ . Note that

$$\begin{aligned} \widehat{\beta}^{\top}(\mu^{(\ell)} - \widehat{\mu}^{(\ell)}) &= (\beta^* + \delta)^{\top}(\mu^{(\ell)} - \widehat{\mu}^{(\ell)}) \\ &\leq (\|\beta_{S_{\eta}}^*\|_1 + \|\beta_{S_{\eta}^c}^*\|_1 + \|\delta\|_1) \|\mu^{(\ell)} - \widehat{\mu}^{(\ell)}\|_{\infty} \\ &\leq \left( \sqrt{|S_{\eta}|} \|\beta^*\|_2 + 5 \|\beta_{S_{\eta}^c}^*\|_1 + 4 \sqrt{|S_{\eta}|} \|\delta\|_2 \right) \|\mu^{(\ell)} - \widehat{\mu}^{(\ell)}\|_{\infty}. \end{aligned}$$

Here the last inequality uses (1.35). Also, note that  $\|\beta^*\| \leq M^{1/2} \Delta$ . With our choice of  $\eta$  in (1.32) and the upper bound for  $\|\delta\|_2$ , when  $n$  satisfies that

$$n \geq C \cdot \sigma_{\max}^2 M \Delta^{-\frac{4}{2-q}} R^{\frac{2}{2-q}} \log p$$

for some constant  $C$ , we have that

$$\widehat{\beta}^\top (\mu^{(\ell)} - \widehat{\mu}^{(\ell)}) \leq \sigma_{\max}^{-q/2} M^{\frac{1-q}{2}} \Delta \sqrt{R} \left( \frac{\log p}{n} \right)^{\frac{2-q}{4}}. \quad (1.69)$$

We then consider the term  $\widehat{\Delta} = \sqrt{\widehat{\beta}^\top \Sigma \widehat{\beta}}$ . Note that

$$\widehat{\Delta}^2 = \widehat{\beta}^\top \Sigma \widehat{\beta} = \Delta^2 + 2\mu_d^\top \delta + \delta^\top \Sigma \delta,$$

Hence we have

$$|\widehat{\Delta}^2 - \Delta^2| \leq 2\|\mu_d\| \|\delta\|_2 + M\|\delta\|_2^2.$$

When  $n$  is sufficiently large, we have that  $|\widehat{\Delta}^2 - \Delta^2| \leq \frac{1}{2}\Delta^2$ . Combining this with (2.56), we have that

$$\left( \frac{\widehat{\beta}^\top (\mu^{(0)} - \widehat{\mu}^{(0)}) + \widehat{\beta}^\top (\mu^{(1)} - \widehat{\mu}^{(1)})}{2\widehat{\Delta}} \right)^2 \leq C \cdot \sigma_{\max}^{-q} M^{1-q} R \left( \frac{\log p}{n} \right)^{1-q/2}$$

for some constant  $C$ . Therefore, Lemma 2.6.8 holds true. ■

## 1.9 Discussions

Besides [14], there are other pivotal methods for regression and inverse covariance estimation problems. For examples, [20] and [21] propose the scaled Lasso method (also known as square-root Lasso) for sparse linear regression, which enjoys a similar tuning-insensitive property to [14]; [22] extend the scaled Lasso to nonparametric regression; [23] extend the scaled Lasso to sparse multivariate regression with inhomogeneous noise; [24] extend the scaled Lasso to sparse linear regression with group structures; [25] and [26] extend the scaled Lasso to inverse covariance matrix estimation; [27] extend [14] to inverse covari-

ance matrix estimation for heavy tail elliptical distributions; [28] and [29] show that the sparse quantile regression and LAD Lasso are also pivotal methods, which enjoy similar tuning-insensitive properties, respectively.

## CHAPTER 2

### ROBUST HIGH-DIMENSIONAL LDA UNDER DATA CONTAMINATION

In this chapter, we introduce a method for enhancing the robustness of high-dimensional LDA against data contamination, where a fraction of the observed sample may be corrupted. Theoretical results and numerical studies are provided to show the usefulness of our proposed method.

#### 2.1 Introduction

In the standard binary classification problem, we observe two classes of  $p$ -dimensional training data,  $\{x_j^{(i)} : j = 1, 2, \dots, n_i; i = 0, 1\}$ , where  $i = 0$  or  $1$  denotes the class label. Linear discriminant analysis (LDA) is a specialization of Fisher's linear discriminant, a statistical method to find a linear combination of the features that is used in decision-making.

We study the high-dimensional and sparse setting of LDA, motivated by many modern real-world applications such as the fMRI decoding and biomarker identification (see, e.g. [2, 3]). Practitioners in these domains suggest that the classification problems in high-dimensions might actually depend only on a small fraction of predictors or features. Several methods have been developed in the literature on high-dimensional LDA with different sparsity assumptions. For instance, [4, 5, 6] assume that the mean difference  $\mu_d = (\mu^{(1)} - \mu^{(0)})$  is sparse, and the common covariance matrix  $\Sigma$  is diagonal, whereas [7] proposes the sparse linear discriminant analysis (SLDA) under the assumption that both  $\mu_d$  and  $\Sigma$  are sparse. Recently, [10, 11, 13] impose a sparsity assumption on the direction  $\beta^* = \Sigma^{-1}\mu_d$ , which will also be adopted in our paper.

Furthermore, we consider an adversarial setup where an adversary can corrupt a fraction of the feature vectors in the training data. It is well known that LDA and other classification

algorithms can be sensitive to outliers, especially in high dimensions. Robust estimates of the class means and the inverse of common covariance matrix would suffice to make LDA robust. However, joint robust estimation of the mean and covariance matrix in high dimensions is known to have high computational complexity and high sample complexity. This beckons the search for methods for robust LDA which are computationally efficient and have low sample complexity.

Note that classical robust mean estimation methods such as coordinate-wise median and geometric median suffer in high dimensions since their error bound scale with the dimension. There are a number of studies that investigate robust mean estimation for contaminated data in high dimensions, e.g., Tukey’s median [30] and Iterative Filtering [31, 32, 33]. More recent studies propose to minimize the spectral norm of weighted sample covariance, and estimate the mean by weighted average, (see, e.g., [34, 35, 36]). Inspired by this approach, the work in [37] proposed to minimize the  $\ell_1$  norm of an outlier indicator vector under the constraint on the spectral norm of weighted sample covariance, thereby developing an order-optimal robust estimate of the mean. Besides the estimation of robust mean, there are recent studies on robust estimation in other statistical problems, such as covariance or precision matrix estimation and regression. For example, [38] investigate the estimation of mixtures of spherical Gaussians in high-dimensions; [39] propose robust covariance and precision matrix estimators for distribution with bounded fourth moment, [40] develop a high dimensional robust sparse regression algorithm based on iterative filtering. [41, 42] develop ad hoc methods for robust sparse linear discriminant analysis.

In our work, we develop a general  $\ell_1$  norm-based optimization framework for robust high-dimensional LDA by combining the robust mean estimation in [37] with the formulation in [14] to bypass the estimation of the inverse of the covariance matrix. We further develop an approximation algorithm for solving the optimization problem. We show that our method yields a solution for which we establish statistical guarantees. Extensive numerical studies using both synthetic data and real data (MNIST) illustrate the usefulness of

our method.

The rest of this chapter is organized as below. In Section 2.2, we briefly review the LDA problem, and introduce the data contamination model. In Section 2.3, we introduce our method. In Section 2.4, we discuss the theoretical properties of our method. In Section 2.5, we present the numerical studies. The proofs of our theoretical results are presented in Section 2.6.

## 2.2 Problem Overview

In this section, we provide necessary background on three topics: (1) LDA in subsection 2.2.1; (2) the data contamination model in subsection 2.2.2, and (3) objective in subsection 2.2.3.

### 2.2.1 LDA

The widely used LDA classifier adopts the structure of the optimal Bayes classifier for the scenario wherein the conditional distributions of the datapoints (conditioned on the class) are Gaussian distributions with a common covariance matrix  $\Sigma$  but different means  $\mu^{(0)}$  and  $\mu^{(1)}$ . The decision criterion of the LDA classifier for a datapoint  $z$  is given by

$$f(z) = 1\{\hat{\beta}^\top z \geq \hat{\gamma}\}, \quad (2.1)$$

where

$$\hat{\beta} = \hat{\Sigma}^{-1} (\hat{\mu}^{(1)} - \hat{\mu}^{(0)}), \quad \hat{\gamma} = \hat{\beta}^\top \left( \frac{\hat{\mu}^{(0)} + \hat{\mu}^{(1)}}{2} \right), \quad (2.2)$$

$\hat{\Sigma}$  is the sample covariance matrix under the common covariance assumption, and  $\hat{\mu}^{(i)}$  is the sample mean of class  $i$ .

We represent the LDA classifier in a general form which is characterized by the  $(p+1)$ -

dimensional quantities  $(\beta, \gamma)$ :

$$f_{\beta, \gamma}(z) = 1\{\beta^\top z \geq \gamma\} \quad (2.3)$$

The misclassification rate of the LDA classifier given by  $f_{\beta, \gamma}(\cdot)$ , under the assumption that the conditional distributions are Gaussian distributions, is given by

$$\begin{aligned} \mathcal{R}(\beta, \gamma) &= \frac{1}{2} \mathbb{P}_{Z \sim N(\mu^{(0)}, \Sigma)}(f_{\beta, \gamma}(Z) = 1) + \frac{1}{2} \mathbb{P}_{Z \sim N(\mu^{(1)}, \Sigma)}(f_{\beta, \gamma}(Z) = 0) \quad (2.4) \\ &= \frac{1}{2} \Phi \left( \frac{\beta^\top \mu^{(0)} - \gamma}{\sqrt{\beta^\top \Sigma \beta}} \right) + \frac{1}{2} \Phi \left( -\frac{\beta^\top \mu^{(1)} - \gamma}{\sqrt{\beta^\top \Sigma \beta}} \right), \end{aligned}$$

where  $\Phi$  is the CDF of the standard Gaussian distribution. The optimal misclassification rate (also known as the Bayes error) is achieved by Fisher's LDA  $f_{\beta^*, \gamma^*}(\cdot)$ , where

$$\beta^* = \Sigma^{-1} (\mu^{(1)} - \mu^{(0)}), \gamma^* = \beta^{*\top} \left( \frac{\mu^{(0)} + \mu^{(1)}}{2} \right).$$

The corresponding optimal misclassification rate is  $\mathcal{R}^* = \Phi(-\Delta/2)$ , where  $\Delta = \sqrt{\beta^{*\top} \Sigma \beta^*} = \sqrt{\mu_d^\top \Sigma^{-1} \mu_d}$  can be viewed as the signal-noise ratio.

### 2.2.2 Data contamination model

We consider the following data contamination model. Let  $\{\tilde{x}_j^{(i)} : i = 0, 1, j \in [n_i]\}$  be an i.i.d. sample drawn from the two class distributions (indexed by  $i = 0, 1$ ) before contamination. With certain contamination level  $0 < \epsilon < 1/2$ , the adversary can view all data and replace at most  $\epsilon n_i$  datapoints in class  $i$  with any arbitrary datapoints that deviate from the original datapoints by an  $\ell_2$ -distance of at most  $R < \infty$ . We denote the resulting contaminated sample by  $\{x_j^{(i)} : i = 0, 1, j \in [n_i]\}$ .

We acknowledge that our contamination model with bounded  $R$  is weaker than the strong contamination model considered in the literature of robust estimation (see, e.g., [31, 33, 37, 32, 36, 34, 35]) wherein  $R$  is unbounded. This assumption on the boundedness

of  $R$  is needed for the theoretical analysis of guarantees on the robust estimation of both covariance matrix and the class means with limited sample sizes.

### 2.2.3 Objective

Our main goal is to robustly estimate the parameters of the LDA classifier under the presence of outliers in the training data with the objective of minimizing the misclassification rate  $\mathcal{R}$  with respect to clean testing data without outliers. The breakdown point  $\epsilon^*$  is defined as the minimum fraction of contamination in the training data that can result in 50% misclassification rate. To be more concrete, denote by  $\tilde{\mathbf{x}}$  and  $\mathbf{x}_\epsilon$  the observed training data before and after  $\epsilon$ -contamination, respectively. With slight abuse of notation, let  $\mathcal{R}(\mathbf{x}_\epsilon)$  be the misclassification rate of LDA learned from  $\mathbf{x}_\epsilon$ . We define the breakdown point  $\epsilon^*$  (in the regime  $\min_i n_i \rightarrow \infty$ ) as

$$\epsilon^* = \inf \left\{ \epsilon : \sup_{\mathbf{x}_\epsilon} \mathcal{R}(\mathbf{x}_\epsilon) = 0.5 \right\}. \quad (2.5)$$

The breakdown point  $\epsilon^*$  depends on the algorithm for discriminant analysis, and reflects the robustness of the algorithm to  $\epsilon$ -contamination. An algorithm with higher  $\epsilon^*$  means one needs to contaminate more datapoints in order to fool the algorithm to perform as poorly as random guessing.

## **2.3 Our Proposed Method**

For the rest of the paper, we assume that the conditional distributions of the training data follow Gaussian distributions with common unknown covariance matrix  $\Sigma$  and unknown class means  $\mu^{(i)}$ ,  $i = 0, 1$ . Let us first provide some high-level ideas of our proposed framework and method. Firstly, we define indicator vector  $h^{(i)}$  such that  $h_j^{(i)} = 1$  represents that  $j$ th observation in the  $i$ th class is an outlier, and ideally we would to estimate the unknown Gaussian means and the common covariance based only on those datapoints with  $h_j^{(i)} = 0$ .

Since the indicators  $h_j^{(i)}$  are unobservable, we would use  $w_j^{(i)}$ , which can be thought of as an estimate of  $1 - h_j^{(i)}$  as well as the weight of  $x_j^{(i)}$  for the purpose of computing weighted mean and covariance matrix.

Joint robust estimation of  $\mu^{(0)}, \mu^{(1)}$  and  $\Sigma$  has high computational complexity as well as sample complexity, under the strong contamination model considered in the literature. Working with a weaker contamination model helps us guarantee robust estimation of the class means and common covariance matrix with low computational and sample complexities. Secondly, we impose the sparsity assumption on  $\beta^* = \Sigma^{-1} (\mu^{(1)} - \mu^{(0)})$ . This inspires us to minimize the  $\ell_1$  norm of  $\beta$  among those vectors satisfying  $\Sigma\beta = \mu^{(1)} - \mu^{(0)}$ . Towards this end, we adapt our Pivotal Linear Discriminant Analysis (PANDA) method in Chapter 1 for estimating sparse  $\beta^*$  in high dimensions, which introduces a new nuisance parameter  $\tau$  that controls the error between  $\Sigma\beta$  and  $\mu^{(1)} - \mu^{(0)}$ .

More specifically, in order to estimate  $\beta^*$  while being robust to outliers, we propose to develop robust LDA via the following optimization problem adapted from our PANDA method:

$$\min_{h^{(0)}, h^{(1)}, \mu^{(0)}, \mu^{(1)}, \beta, \tau} \|\beta\|_1 + c_1 \tau^2, \quad (2.6)$$

subject to the constraints

$$0 \leq h_j^{(i)} \leq 1, \forall i \in \{0, 1\} \text{ and } j \in [n], \quad (2.7)$$

$$\|h^{(i)}\|_1 \leq c_2 n, \quad i \in \{0, 1\} \quad (2.8)$$

$$\lambda_{\max}(\Sigma^{(i)}(h^{(i)}, \mu^{(i)})) \leq c_3 \sigma^2, \quad i \in \{0, 1\} \quad (2.9)$$

$$\|\Sigma(h, \mu)\beta - (\mu^{(1)} - \mu^{(0)})\|_{\infty} \leq c_4 \tau, \quad (2.10)$$

$$\sqrt{\beta^{\top} \Sigma(h, \mu) \beta} \leq \tau. \quad (2.11)$$

Here  $c_1$ ,  $c_2$ ,  $c_3$  and  $c_4$  are tuning parameters, and

$$\begin{aligned}\Sigma^{(i)}(h^{(i)}, \mu^{(i)}) &:= \frac{\sum_{j=1}^{n_i} (1 - h_j^{(i)})(x_j^{(i)} - \mu^{(i)})(x_j^{(i)} - \mu^{(i)})^\top}{\sum_{j=1}^{n_i} (1 - h_j^{(i)})}, \\ \Sigma(h, \mu) &:= \frac{\sum_{i=0}^1 \sum_{j=1}^{n_i} (1 - h_j^{(i)})(x_j^{(i)} - \mu^{(i)})(x_j^{(i)} - \mu^{(i)})^\top}{\sum_{i=0}^1 \sum_{j=1}^{n_i} (1 - h_j^{(i)})},\end{aligned}$$

where  $h = (h^{(0)}, h^{(1)})$  and  $\mu = (\mu^{(0)}, \mu^{(1)})$ .

It is useful to add some remarks to the constraints of this optimization problem. The parameter  $\tau$  plays a crucial role and is used for calibration through constraints (2.10) and (2.11), such that one can simultaneously estimate  $\Delta$  and  $\beta^*$ . In addition, by (2.11),  $\tau$  can also be thought of as the estimator for  $\Delta = \sqrt{\beta^{*\top} \Sigma \beta^*}$ , which is an important quantity in LDA. The constraints (2.7), (2.8) and (2.9) are for robust mean estimation of the two classes, as inspired by reference [37]. We take  $h^{(i)}$  as the outlier indicator vector for class  $i$ , and we relax the binary constraint ( $h_j^{(i)} \in \{0, 1\}$ ) to the continuous one for computational tractability. Furthermore, we impose constraints on  $\ell_1$  norm of  $h^{(i)}$ , as we also expect sparsity in  $h^{(i)}$ , since the outliers only take a small fraction of the sample. Constraint (2.9) guarantees that the mean estimate is not drawn too far away by those outliers whose  $h_j^{(i)}$ 's are close to 1.

On one hand, the optimization problem (2.6)-(2.11) is non-convex and hence it is difficult to find a global optimum in general. On the other hand, fortunately, as we will show in next section, any locally optimal solution enjoys nice properties both in terms of estimation error and misclassification rate for LDA. With this motivation, we develop an efficient (approximation) algorithm for robust LDA, which is outlined in Algorithm 2. Note that the subroutine (2.12)-(2.14) can be viewed as a packing semi-definite programming, and can be solved through approximate algorithms such as in [43]. Moreover, the optimization problem (2.15)-(2.17) is a second-order cone programming, which is convex and can be efficiently solved by off-the-shelf software packages.

---

**Algorithm 2** An Approximation Algorithm for Robust LDA
 

---

**Input:** Specify (contaminated) data  $\{x_j^{(i)} : i \in \{0, 1\}, j \in [n_i]\}$ . Specify  $\epsilon$  as the upper bound of the fraction of contamination, and  $\sigma$  such that  $\Sigma \preceq \sigma^2 I_p$ . Specify the parameters  $c_1, c_2, c_3, c_4$ .

Initialize  $\mu^{(i),0}$  as the median of all samples in class  $i$ .

**for**  $t = 1$  to  $T$  **do**

Solve  $h^{(i),t}$  through the following sub-routine:

$$\min_{h^{(i)}} \|h^{(i)}\|_1 \quad (2.12)$$

$$\text{subject to } 0 \leq h_j^{(i)} \leq 1, \forall j \in [n_i], \quad (2.13)$$

$$\lambda_{\max}(V^{(i)}(h^{(i)}, \mu^{(i),t-1})) \leq c_3 \sigma^2 n_i, \quad (2.14)$$

where  $w_j^{(i)} = 1 - h_j^{(i)}$  and

$$\begin{aligned} & V^{(i)}(h^{(i)}, \mu^{(i),t-1}) \\ &= \sum_j w_j^{(i)} (x_j^{(i)} - \mu^{(i),t-1})(x_j^{(i)} - \mu^{(i),t-1})^\top. \end{aligned}$$

Update  $\mu^{(i),t}$  by  $\mu^{(i),t} = \frac{\sum_j w_j^{(i),t} x_j^{(i)}}{\sum_j w_j^{(i),t}}$ .

**end for**

Let  $\hat{\mu}^{(i)} = \mu^{(i),T}$  and  $\hat{h}^{(i)} = h^{(i),T}$ . Then compute

$$\begin{aligned} \hat{\mu}_m &= \frac{1}{2} (\hat{\mu}^{(1)} + \hat{\mu}^{(0)}), \quad \hat{\mu}_d = \hat{\mu}^{(1)} - \hat{\mu}^{(0)}, \\ \hat{\Sigma} &= \frac{\sum_{i,j} w^{(i),t} (x_j^{(i)} - \hat{\mu}^{(i)})(x_j^{(i)} - \hat{\mu}^{(i)})^\top}{\sum_{i,j} w^{(i),t}}. \end{aligned}$$

Solve  $(\hat{\beta}, \hat{\tau})$  from the following optimization problem:

$$(\hat{\beta}, \hat{\tau}) \in \arg \min_{\beta, \tau} \|\beta\|_1 + c_1 \tau^2 \quad (2.15)$$

$$\text{subject to } \|\hat{\Sigma} \beta - \hat{\mu}_d\|_\infty \leq c_4 \tau, \quad (2.16)$$

$$\sqrt{\beta^\top \hat{\Sigma} \beta} \leq \tau. \quad (2.17)$$

**Output:**  $(\hat{\beta}, \hat{\mu}_m, \hat{h}^{(0)}, \hat{h}^{(1)})$ .

---

## 2.4 Theoretical Results

In this section, we discuss the theoretical properties of our method. We first state some assumptions for our results to hold, which are standard in the literature of robust estimation and LDA in high dimensions, see [37] and [13] for example.

**Assumption 2.4.1** *We assume that  $\beta^*$  is sparse that  $\|\beta\|_0 \leq s$ .*

**Assumption 2.4.2** *We assume  $\frac{1}{M}I_p \preceq \Sigma \preceq \sigma^2 I_p$  for some  $M$  and  $\sigma^2$ . Also, we assume that  $\|\mu_d\|_2 \geq a > 0$ .*

For convenience, we define the following notation, which will be used throughout the rest of the paper.

$$\begin{aligned}\tilde{h}_j^{(i)} &= \begin{cases} 0 & \text{if } x_j^{(i)} = \tilde{x}_j^{(i)} \\ 1 & \text{otherwise} \end{cases}, \quad \tilde{w}_j^{(i)} = 1 - \tilde{h}_j^{(i)}, \\ \tilde{\mu}^{(i)} &= \frac{1}{\sum_j \tilde{w}_j^{(i)}} \sum_j \tilde{w}_j^{(i)} x_j^{(i)}, \\ \tilde{\Sigma}^{(i)} &= \frac{1}{\sum_j \tilde{w}_j^{(i)}} \sum_j \tilde{w}_j^{(i)} (x_j^{(i)} - \tilde{\mu}^{(i)})(x_j^{(i)} - \tilde{\mu}^{(i)})^\top, \\ \tilde{\Sigma} &= \frac{1}{\sum_{i,j} \tilde{w}_j^{(i)}} \sum_{i,j} \tilde{w}_j^{(i)} (x_j^{(i)} - \tilde{\mu}^{(i)})(x_j^{(i)} - \tilde{\mu}^{(i)})^\top.\end{aligned}$$

We first present an upper bound for the  $\ell_2$  norm of the estimation error of  $\beta^*$ .

**Theorem 2.4.1** *Suppose that Assumptions 2.4.1 and 2.4.2 hold. Let  $(\hat{\beta}, \hat{\tau}, \hat{h}^{(0)}, \hat{h}^{(1)}, \hat{\mu}^{(0)}, \hat{\mu}^{(1)})$*

be a local optimal solution for (2.6). When  $\epsilon < 1/2$ ,  $c_2 < 1/2$ ,  $c_3 > 1$ ,

$$c_1 = \frac{1}{4} \left( \|\mu_d\|_\infty + 10\sigma\Delta\sqrt{\frac{\log p}{(1-\epsilon)n}} + \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}(3(1+c_3)\sigma + 2R)\|\beta^*\|_\infty \right)^{-1} \quad (2.18)$$

$$c_4(\sigma, M, \epsilon, R, \Delta) = 40\sigma\sqrt{\frac{\log p}{(1-\epsilon)n}} + 2\sqrt{M}\sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}(3(1+c_3)\sigma + 2R) + \frac{2c_3\sigma}{\Delta} \left( \sqrt{\frac{1}{1-\epsilon}} + \sqrt{\frac{1}{1-c_2}} \right) \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}, \quad (2.19)$$

and the following conditions are satisfied that

$$n \geq \max \left( \frac{C \cdot p \log p}{c_3 - 1}, \frac{C}{1-\epsilon} M^2 \sigma^4 s \log p \right), \quad (2.20)$$

$$\sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}(3(1+c_3)\sigma + 2R) \leq \frac{1}{C' M^{3/2} \sigma \sqrt{s}} \quad (2.21)$$

for some universal constants  $C$  and  $C'$ , we have, with probability going 1, as  $n \rightarrow \infty$ ,

$$\|\widehat{\beta} - \beta^*\|_2 \leq C \cdot M\Delta\sqrt{s}c_4(\sigma, M, \epsilon, R, \Delta) \quad (2.22)$$

for some universal constant  $C$ .

**Remark 2.4.1** From (2.19) and (2.22), we can see that the estimation error for  $\beta^*$  depends on the choice of tuning parameters, especially  $c_2$ . Ideally, we want to set  $c_2$  close to  $\epsilon$ , which gives smaller error while allowing us to identify the outliers. The choice of  $c_3$  is also essential: if  $c_3$  is too small, the optimization problem might be infeasible, whereas a  $c_3$  too large can also hurt performance, as more outliers might be taken as inliers. Therefore, one needs to fine-tune the parameters for more desirable performance. In addition, the quantity  $\sigma^2$  might be also unknown in practice, in which case one can fine-tune  $c_3\sigma^2$  directly as one

tuning parameter. Finally, for the case of clean data ( $\epsilon = 0$ ), with sufficiently small  $c_2$ , we can achieve a convergence rate of  $O\left(\sigma M \Delta \sqrt{\frac{s \log p}{n}}\right)$  for  $\|\hat{\beta} - \beta^*\|_2$ , which matches the minimax optimal rate of convergence [13].

Following the result in Theorem 2.4.1, we next introduce an upper bound for the misclassification rate of the discriminant rule based on  $(\hat{\beta}, \hat{\mu}_m)$ .

**Theorem 2.4.2** *Under the same condition as in Theorem 2.4.1, we have, with probability going to 1, as  $n \rightarrow \infty$ , the following bound for the misclassification rate  $\mathcal{R}(\hat{\beta}, \hat{\mu}_m)$ :*

$$\mathcal{R}(\hat{\beta}, \hat{\mu}_m) - \mathcal{R}^* \leq C \cdot \sigma^2 M^2 \Delta s \exp\left(-\frac{\Delta^2}{8}\right) c_4^2. \quad (2.23)$$

Here  $\mathcal{R}(\hat{\beta}, \hat{\mu}_m)$  denotes the misclassification rate under the discriminant rule  $f(\hat{\beta}, \hat{\beta}^\top \hat{\mu}_m)$  as defined in (2.1), and  $\mathcal{R}^*$  is the minimal misclassification rate under the oracle Fisher's discriminant rule using  $\beta^*$  and  $\mu_m$ .

Note that when  $\epsilon = 0$ , with sufficiently small  $c_2$  we can achieve the convergence rate of  $O\left(\sigma^2 M^3 \Delta \exp\left(-\frac{\Delta^2}{8}\right) \frac{s \log p}{n}\right)$  for the misclassification rate, which also matches the minimax optimal convergence rate in [13]. Also, note that with our method for high-dimensional LDA, the breakdown point for LDA depends on the data distribution, as implied by (2.21). In particular the breakdown point of the proposed algorithm can be lower bounded as:

$$\epsilon^* \geq \frac{1}{CR^2 M^3 \sigma^2 s}$$

for some constant  $C$ .

## 2.5 Numerical Study

In this section, we present experimental results over both simulation data and MNIST data to show the usefulness of our proposed method.

### 2.5.1 Simulation

**Data settings.** Given  $n$  for sample size,  $p$  for data dimension and  $s$  for  $\beta^*$  sparsity, we consider the following three models for  $\beta^*$  and  $\Sigma$ :

- (a) **AR(1).** We let  $\Omega_{j,k} = 0.9^{|j-k|}$ ,  $\Sigma = \Omega^{-1}$  and  $\beta^* = (2/\sqrt{s}, \dots, 2/\sqrt{s}, 0, \dots, 0)^\top$ , where the first  $s$  entries are non-zero.
- (b) **Varying diagonal.** We let the diagonal entries of  $\Sigma$  as  $\Sigma_{j,j} = 11$  for  $j = 1, 2, \dots, 5$ , and  $\Sigma_{j,j} = 1 + U_j$  for  $j = 6, 7, \dots, p$ , where  $U_i$ 's are independently drawn from the uniform distribution  $U(0, 1)$ , and we let the off-diagonal entries be  $\Sigma_{j,k} = 0.9^{|j-k|}$ . We let  $\beta^* = (1/\sqrt{s}, \dots, 1/\sqrt{s}, 0, \dots, 0)^\top$ , where only the first  $s$  entries are non-zero.
- (c) **Approximately sparse setting.** We let  $\Sigma_{j,k} = 0.9^{|j-k|}$  and  $\beta_j^* = 0.75^j$ , which is approximately sparse.

We then set  $\mu^{(0)} = \mathbf{0}_p$ ,  $\mu^{(1)} = \Sigma\beta^*$ , and we draw random pre-contamination sample from  $N(\mu^{(i)}, \Sigma)$  for two classes  $i = 0, 1$ , each of size  $n = 100$ . Next, we apply the data contamination. We randomly select  $\epsilon = 0.2$  fraction of the sample in each class, and replace them with random data generated by  $N(\mu^{(i)} + 10\mu_d, I_p)$ , where  $\mu_d = \mu^{(1)} - \mu^{(0)}$ . Finally, we repeat the process above to get 100 independent datasets for comparing the performance.

**Performance.** We compare the performance of our proposed method with classical LDA which does not consider the outliers, by looking at the estimation error for  $\beta^*$  and the misclassification rate over the data. For LDA without considering robust mean estimation, we also use the similar calibration framework and minimize the  $\ell_1$  norm of the  $\beta$  estimator. We set  $n = 100$ ,  $p = 100, 200$ , and  $s = 5$  for all the data generation models. For the tuning parameters, we set  $c_1 = 20$ ,  $c_2 = 0.3$ ,  $c_3 = 1.1$ ,  $c_4 = 0.5$  based on cross-validation. In the followings, we present the performance of different methods in the estimation error of  $\mu_m, \beta^*$ , and the misclassification rate over the testing set, which is independently drawn from the underlying distribution without contamination.

Table 2.1:  $\ell_2$  error for  $\mu_m$  estimate with empirical and robust methods averaged over 100 replications. The standard deviations are presented in brackets.

AR(1)	$(s, p)$	(5, 100)	(5, 200)
	Empirical	12.144 (0.293)	12.412 (0.137)
	Robust	2.304 (0.192)	3.454 (0.149)
Varying diagonal	$(s, p)$	(5, 100)	(5, 200)
	Empirical	22.910 (0.258)	23.123 (0.254)
	Robust	2.421 (0.198)	3.586 (0.157)
Approximately sparse	$p$	100	200
	Empirical	15.093 (0.233)	15.074 (0.179)
	Robust	0.756 (0.199)	1.075 (0.185)

**$\mu_m$  estimation error.** Table 2.1 presents the estimation error of  $\mu_d$  based on the empirical method (i.e. difference of sample mean of two classes) and our method using robust mean estimators. The results are averaged over 100 replications. As can be seen, the robust approach yields much smaller error for  $\mu_m$  by identifying the outliers.

**$\beta^*$  estimation error.** Table 2.2 presents the estimation error of  $\mu_d$  with our method compared with three other methods: The empirical method directly applies Fisher’s discriminant rule with plug-in estimators; The Calibrated LDA method estimates  $\hat{\mu}^{(i)}$  and  $\hat{\Sigma}$  by empirical estimators, and solves  $\hat{\beta}$  using the calibrated framework in (2.15)-(2.17); The Robust Mean method robustly estimates  $\mu^{(i)}$  and  $\Sigma$  by identifying the outliers, and estimates  $\hat{\beta} = \hat{\Sigma}^{-1}\hat{\mu}_d$ . Note that for  $p = 200$ , both the empirical method and robust mean method are inapplicable since the sample covariance matrix is singular. Table 2.2 shows that the calibrated LDA framework is useful for reducing the estimation error for  $\beta^*$  in high dimensions by  $\ell_1$  penalty.

**Misclassification rate.** Table 2.3 presents the misclassification rate of different methods

Table 2.2:  $\ell_2$  error for  $\beta^*$  estimate averaged over 100 replications. The standard deviations are presented in brackets.

AR(1)	$(s, p)$	(5, 100)	(5, 200)
	Empirical	2.793 (0.410)	–
	Calibrated LDA	2.005 (0.010)	2.000 (0.007)
	Robust Mean	14.294 (3.135)	–
	Our method	1.741 (0.105)	2.091 (0.121)
Varying diagonal	$(s, p)$	(5, 100)	(5, 200)
	Empirical	2.305 (0.436)	–
	Calibrated LDA	1.000 (0.002)	1.001 (0.003)
	Robust Mean	9.017 (2.641)	–
	Our method	0.460 (0.108)	0.710 (0.229)
Approximately sparse	$p$	100	200
	Empirical	10.610 (1.740)	–
	Calibrated LDA	1.149 (0.010)	1.145 (0.011)
	Robust Mean	34.464 (7.521)	–
	Our method	1.558 (0.323)	1.929 (0.563)

Table 2.3: Misclassification Rate averaged over 100 replications. The standard deviations are presented in brackets.

	$(s, p)$	(5, 100)	(5, 200)
AR(1)	Empirical	0.491 (0.023)	–
	Calibrated LDA	0.475 (0.022)	0.487 (0.018)
	Robust Mean	0.323 (0.029)	–
	Our method	0.240 (0.022)	0.301 (0.031)
Varying diagonal	$(s, p)$	(5, 100)	(5, 200)
	Empirical	0.488 (0.023)	–
	Calibrated LDA	0.458 (0.022)	0.465 (0.017)
	Robust Mean	0.184 (0.025)	–
Approximately sparse	$p$	100	200
	Empirical	0.447 (0.029)	–
	Calibrated LDA	0.447 (0.016)	0.454 (0.020)
	Robust Mean	0.257 (0.038)	–
Our method		0.110 (0.019)	0.121 (0.011)

over the testing data. We see that by combining the robust mean estimation and a calibration framework, our proposed method yields the lowest misclassification rate. Meanwhile, the methods not considering the outliers are easily corrupted by the contaminated data, with a misclassification rate nearly 50%.

### 2.5.2 MNIST data example

In this subsection, we present the numerical studies using the MNIST dataset [44]. The MNIST dataset contains 60000 training samples and 10000 testing samples, each is an image of hand-written digit (0-9) of  $28 \times 28$  pixels. Using this dataset, we apply different methods of LDA to classify the images of digits ‘1’ and ‘7’ and compare their performance. We first scale the values of each pixel to the range of  $[0, 1]$ . To save the computational time, we resize the images to  $14 \times 14$ , and randomly sample  $n = 100$  images of each class from the training data to train the LDA models. Then we evaluate the performance over the testing data with the two classes. The process is repeated 100 times.

Table 2.4: Misclassification rates over the testing data.

Method	Misclassification rate
Calibrated LDA	0.077 (0.007)
Our method	0.030 (0.023)

We consider the following process for data contamination. We first compute the sample mean of the original data of two classes for each pixel, say  $\mu^{(0)}$  and  $\mu^{(1)}$ . To fool the classifier, we select a random  $\epsilon = 0.2$  fraction of samples from each class, and shift their pixels by  $10(\mu^{(1)} - \mu^{(0)})$ , and finally clip the pixels of resulting images to the range  $[0, 1]$ . We then learn the classifier using the resulting corrupted sample, and evaluate the performance over the testing data with clean images of ‘1’ and ‘7’. Since the empirical method with plug-in Fisher’s discriminant rule is inapplicable for the case where  $p > n$ , here we compare our proposed method with the calibrated LDA method without robust estimation. Table 2.4 presents the misclassification rates over the MNIST testing data, where our method achieves lower misclassification rate by identifying outliers in each class, compared to the calibrated LDA method.

## 2.6 Proofs

This section contains three subsections. In 3.7.1 and 2.6.2, we provide proofs to Theorem 2.4.1 and 2.4.2. In Section 2.6.3, we provide proofs to auxiliary lemmas.

### 2.6.1 Proof of Theorem 4

**Proof** Let  $(\hat{h}^{(i)}, \hat{\mu}^{(i)}, \hat{\beta}, \hat{\tau})$  be a local optimal solution to (2.6). We define

$$\tilde{\tau} = \sqrt{\beta^{*\top} \tilde{\Sigma} \beta^*}. \tag{2.24}$$

Also, we define  $\Delta_{n,\epsilon}$  as the set of probability vector that

$$\Delta_{n,\epsilon} = \left\{ w \in \mathbb{R}^n : 0 \leq w_j \leq \frac{1}{1-\epsilon}, \sum_j w_j = 1 \right\}.$$

Furthermore, we let  $\sigma_{\max}^2 = \max_{k \in [p]} \Sigma_{k,k}$  as the largest diagonal entry of  $\Sigma$ , and similarly we define  $\tilde{\sigma}_{\max}^2 = \max_{k \in [p]} \tilde{\Sigma}_{k,k}$  with slight abuse of notation.

For convenience of presentation, we first define the following events with some universal constant  $C$ .

$$\mathcal{E}_\mu = \left\{ \|\tilde{\mu}^{(i)} - \mu^{(i)}\|_2 \leq \sigma \sqrt{\frac{2 \log p}{(1-\epsilon)n}}, i = 0, 1 \right\} \quad (2.25)$$

$$\mathcal{E}_\tau = \left\{ \left| \beta^{*\top} (\tilde{\Sigma} - \Sigma) \beta^* \right| \leq \frac{1}{2} \beta^{*\top} \Sigma \beta^* \right\}, \quad (2.26)$$

$$\mathcal{E}_{\sigma_{\max}} = \left\{ \left| \tilde{\sigma}_{\max}^2 - \sigma_{\max}^2 \right| \leq \frac{1}{2} \sigma_{\max}^2 \right\}, \quad (2.27)$$

$$\mathcal{E}_1 = \left\{ \|\tilde{\Sigma} - \Sigma\|_\infty \leq 10\sigma \Delta \sqrt{\frac{\log p}{(1-\epsilon)n}} \right\}, \quad (2.28)$$

$$\mathcal{E}_2 = \left\{ \|\tilde{\Sigma} \beta^* - \tilde{\mu}\|_\infty \leq 20\sigma \tilde{\tau} \sqrt{\frac{\log p}{(1-\epsilon)n}} \right\}. \quad (2.29)$$

The following lemma shows that the events defined above occur with high probability.

**Lemma 2.6.1** *For any  $\beta^* \in \mathbb{R}^p$ , we have*

$$\begin{aligned} \mathbb{P}(\mathcal{E}_\mu) &\geq 1 - 4p^{-1}, \\ \mathbb{P}(\mathcal{E}_\tau) &\geq 1 - 2 \exp\left(-\frac{n-1}{16}\right), \\ \mathbb{P}(\mathcal{E}_{\sigma_{\max}}) &\geq 1 - 2p \exp\left(-\frac{n-1}{16}\right), \\ \mathbb{P}(\mathcal{E}_1) &\geq 1 - 2p^{-1}. \end{aligned}$$

Moreover, we have

$$\mathcal{E}_2 \supseteq \left( \mathcal{E}_\tau \cap \mathcal{E}_{\sigma_{\max}} \cap \mathcal{E}_1 \right).$$

We now introduce the following lemma that bounds the difference in weighted mean and covariance with total variation. In fact, it can be deduced from Lemma C.2 in [36].

**Lemma 2.6.2** *Let  $\{x_j : j \in [n]\}$  be a finite set of data points. Let  $w^1$  and  $w^2$  be two probability weight vectors in  $\Delta_{n,0}$ . Let  $\mu_{w^i} = \sum_j w_j^i x_j$  and  $\Sigma_{w^i} = \sum_j w_j^i (x_j - \mu_{w^i})(x_j - \mu_{w^i})^\top$  be the weighted mean and covariance under  $w^i$ . If the total variation between  $w^1$  and  $w^2$  satisfies that  $TV(w^1, w^2) \leq \zeta$ , then we have*

$$\|\mu_{w^1} - \mu_{w^2}\|_2 \leq \left( \sqrt{\|\Sigma_{w^1}\|_2} + \sqrt{\|\Sigma_{w^2}\|_2} \right) \sqrt{\frac{\zeta}{1-\zeta}}.$$

Using Lemma 2, we have

$$\|\tilde{\mu}^{(i)} - \mu^{(i)}\|_2 \leq c_3 \sigma \left( \sqrt{\frac{1}{1-\epsilon}} + \sqrt{\frac{1}{1-c_2}} \right) \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}.$$

Our next lemma upper bound  $\|\widehat{\Sigma} - \tilde{\Sigma}\|_2$  under our contamination model.

**Lemma 2.6.3** *Suppose that the event  $\mathcal{E}_\mu$  holds, and  $\|x_j^{(i)}\|_2 \leq R$  if  $x_j^{(i)} \neq \tilde{x}_j^{(i)}$ . Then with probability at least  $1 - \exp(-n/8)$  we have*

$$\|\widehat{\Sigma} - \tilde{\Sigma}\|_2 \leq \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}} ((2 + 3c_3)\sigma^2 + R).$$

We next present a few lemmas that help us to bound  $\|\delta\|_2$ .

**Lemma 2.6.4** *Let  $(\widehat{h}, \widehat{\mu}, \widehat{\beta}, \widehat{\tau})$  be a local optima for the optimization problem (2.6)-(2.11), and let  $\delta = \widehat{\beta} - \beta^*$ . Let  $S = \text{supp}(\beta^*)$ . Under Assumptions 2.4.1, 2.4.2 and events  $\mathcal{E}_\tau, \mathcal{E}_1, \mathcal{E}_2$ , when  $c_1, c_4$  satisfy (2.18) and (2.19), we have*

$$\|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1.$$

**Lemma 2.6.5** Let  $\tau^* = \sqrt{\beta^{*\top} \widehat{\Sigma} \beta^*}$ . Suppose  $(\widehat{h}, \widehat{\mu}, \beta^*, \tau^*)$  is feasible to (2.6), and  $\mathcal{E}_\tau$  holds. We have

$$\delta^\top \widehat{\Sigma} \delta \leq c_4(3\Delta + \sqrt{\|\delta\|_1/c_1}) \|\delta\|_1.$$

**Lemma 2.6.6** There exist absolute positive constants  $C_1$  and  $C_2$  such that with probability at least  $1 - C_1 \exp(-C_2 n)$ , we have

$$\begin{aligned} \delta^\top \widehat{\Sigma} \delta &\geq \delta^\top \widetilde{\Sigma} \delta - \|\widehat{\Sigma} - \widetilde{\Sigma}\|_2 \|\delta\|_2^2 \\ &\geq \left( \frac{1}{32M} - \sqrt{\frac{\max(\epsilon, c_2)}{1 - \epsilon - c_2}} (3(1 + c_3)\sigma + 2R) \right) \|\delta\|_2^2 \\ &\quad - 81\sigma^2 \frac{\log p}{(1 - \epsilon)n} \|\delta\|_1^2. \end{aligned} \tag{2.30}$$

Combining (2.52) and (2.30), we obtain that

$$\begin{aligned} &c_4(3\Delta + \sqrt{\|\delta\|_1/c_1}) \|\delta\|_1 \\ &\geq \left( \frac{1}{32M} - \sqrt{\frac{\max(\epsilon, c_2)}{1 - \epsilon - c_2}} (3(1 + c_3)\sigma + 2R) \right) \|\delta\|_2^2 \\ &\quad - 81\sigma^2 \frac{\log p}{(1 - \epsilon)n} \|\delta\|_1^2. \end{aligned}$$

Combining the above result with the following inequality that

$$\|\delta\|_1 = \|\delta_S\|_1 + \|\delta_{S^c}\|_1 \leq 4\|\delta_S\|_1 \leq 4\sqrt{s}\|\delta\|_2,$$

we have

$$\|\delta\|_2 \leq C \cdot \frac{c_4 M \Delta \sqrt{s}}{1 - 32M \sqrt{\frac{\max(\epsilon, c_2)}{1 - \epsilon - c_2}} (3(1 + c_3)\sigma + 2R)}$$

for some universal constant  $C$ . This completes our proof for Theorem 2.4.1. ■

### 2.6.2 Proof of Theorem 5

**Proof** Let  $\widehat{\Delta} = \sqrt{\widehat{\beta}^\top \Sigma \widehat{\beta}}$ . The misclassification risk of  $\widehat{\beta}$  is

$$\mathcal{R}(\widehat{\beta}) = \frac{1}{2} \Phi \left( -\frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) + \frac{1}{2} \Phi \left( \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right), \quad (2.31)$$

where  $\Phi(\cdot)$  is the CDF of the standard Gaussian distribution. Recall that the optimal risk achieved by Fisher's rule is  $\mathcal{R}^* = \Phi(-\frac{\Delta}{2})$ . For the first term on the right-hand side of (2.31), its second order Taylor's expansion is

$$\begin{aligned} \Phi \left( -\frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) &= \Phi \left( -\frac{\Delta}{2} \right) + \Phi' \left( -\frac{\Delta}{2} \right) \left( \frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) \\ &\quad + \frac{\Phi''(t_1)}{2} \left( \frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2, \end{aligned} \quad (2.32)$$

where  $t_1 \in \left( \frac{-\Delta}{2}, -\frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)$ . Similarly, for the second term in (2.31), we have

$$\begin{aligned} \Phi \left( \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) &= \Phi \left( -\frac{\Delta}{2} \right) + \Phi' \left( -\frac{\Delta}{2} \right) \left( \frac{\Delta}{2} + \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right) \\ &\quad + \frac{\Phi''(t_2)}{2} \left( \frac{\Delta}{2} + \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2, \end{aligned} \quad (2.33)$$

where  $t_2 \in \left(-\frac{\Delta}{2}, \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}}\right)$ . Combining (2.32) and (2.33), we have

$$\begin{aligned} \mathcal{R}(\widehat{\beta}, \widehat{\mu}_m) - \mathcal{R}^* &= \Phi' \left( -\frac{\Delta}{2} \right) \left( \frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \right) \\ &\quad + \frac{\Phi''(t_1)}{2} \left( \frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2 \\ &\quad + \frac{\Phi''(t_2)}{2} \left( \frac{\Delta}{2} + \frac{(\widehat{\mu}_m - \mu^{(1)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2. \end{aligned} \quad (2.34)$$

We now introduce a lemma that upper bounds the first term on the right-hand side of (2.34).

**Lemma 2.6.7** *Suppose (2.22) holds, and  $n$  satisfies that*

$$n \geq \frac{C}{1-\epsilon} M^2 \sigma^4 s \log p \quad (2.35)$$

for some constant  $C$ , and

$$(\tilde{\kappa} + \widehat{\kappa}) \sqrt{\max(\epsilon, c_2)} \leq \frac{1}{C' M^{3/2} \sigma \sqrt{s}} \quad (2.36)$$

for some constant  $C'$ . Then we have

$$\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \leq \frac{\sigma^2}{2\Delta} \|\delta\|_2^2. \quad (2.37)$$

Note that  $\Phi'(-\Delta/2) = (2\pi)^{-1/2} \exp(-\Delta^2/8)$ . Following Lemma 2.6.7, we have

$$\Phi' \left( -\frac{\Delta}{2} \right) \left( \frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \right) \leq \frac{\sigma^2}{2\sqrt{2\pi}\Delta} \exp \left( -\frac{\Delta^2}{8} \right) \|\delta\|_2^2. \quad (2.38)$$

Now we consider the second-order term in (2.34). First, using Lemma 2.6.7, we have

$$\begin{aligned} \frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} &= \frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} + \frac{\widehat{\beta}^\top (\mu_m - \widehat{\mu}_m)}{\widehat{\Delta}} \\ &\leq \frac{\sigma^2}{2\Delta} \|\delta\|_2^2 + \frac{\widehat{\beta}^\top (\mu^{(0)} - \widehat{\mu}^{(0)}) + \widehat{\beta}^\top (\mu^{(1)} - \widehat{\mu}^{(1)})}{2\widehat{\Delta}}. \end{aligned} \quad (2.39)$$

After taking square, the first term on the right-hand side gives  $\frac{\sigma^2}{4\Delta^2} \|\delta\|_2^4$ , which is negligible compared to the first-order term. Hence it suffices to bound the second term on the right-hand side of (2.39). For this aim we introduce the next lemma.

**Lemma 2.6.8** *Under the same conditions as for Theorem 2.4.1, with probability at least  $1 - 4p^{-1}$  we have*

$$\begin{aligned} &\left( \frac{\widehat{\beta}^\top (\mu^{(0)} - \widehat{\mu}^{(0)}) + \widehat{\beta}^\top (\mu^{(1)} - \widehat{\mu}^{(1)})}{2\widehat{\Delta}} \right)^2 \\ &\leq C \cdot M\sigma^2 \left[ \frac{\log p}{(1-\epsilon)n} + c_3^2 \left( \frac{1}{1-\epsilon} + \frac{1}{1-c_2} \right) \frac{\max(\epsilon, c_2)}{1-\epsilon-c_2} \right] \end{aligned} \quad (2.40)$$

for some constant  $C$ .

Since  $t_1 > -\Delta/2$ , we have  $|\Phi''(t_1)| \leq C \cdot \Delta \exp(-\Delta^2/8)$ . Combining it with (2.40), we bound the second term in (2.34) by

$$\begin{aligned} &\frac{|\Phi''(t_1)|}{2} \left( \frac{\Delta}{2} - \frac{(\widehat{\mu}_m - \mu^{(0)})^\top \widehat{\beta}}{\widehat{\Delta}} \right)^2 \\ &\leq C \cdot \Delta \exp\left(-\frac{\Delta^2}{8}\right) \sigma^2 M \left[ \frac{\log p}{(1-\epsilon)n} + c_3^2 \left( \frac{1}{1-\epsilon} + \frac{1}{1-c_2} \right) \frac{\max(\epsilon, c_2)}{1-\epsilon-c_2} \right] \end{aligned} \quad (2.41)$$

for some constant  $C$ . Likewise, the third term in (2.34) is also subject to this bound.

Finally, plugging (2.38) and (2.41) into (2.34), and using (2.22), and noticing the dom-

inance of the the first-order term, we obtain that

$$\begin{aligned}\mathcal{R}(\widehat{\beta}, \widehat{\mu}_m) - \mathcal{R}(\beta^*) &\leq C \cdot \frac{\sigma^2 M^2 \Delta s}{2\sqrt{2\pi}\Delta} \exp\left(-\frac{\Delta^2}{8}\right) \|\delta\|_2^2 \\ &\leq C \cdot \frac{\sigma^2 M^2 \Delta s}{2\sqrt{2\pi}} \exp\left(-\frac{\Delta^2}{8}\right) c_4^2\end{aligned}$$

for some constant  $C$ , which completes our proof. ■

### 2.6.3 Proof of lemmas

#### *Proof of Lemma 2.6.1*

**Proof** There are four main statements in Lemma 2.6.1, and let us prove them one by one.

- (i) The first inequality of  $\mathbb{P}(\mathcal{E}_\mu)$  is a standard result of concentration from the fact that  $\widetilde{x}_j^{(i)}$ 's are i.i.d.  $\sigma^2$ -subgaussian.
- (ii) It suffices to show that

$$\mathbb{P}(\mathcal{E}_\tau) = \mathbb{P}\left(|\beta^{*\top}(\widehat{\Sigma} - \Sigma)\beta^*| \leq \frac{1}{2}\beta^{*\top}\Sigma\beta^*\right) \geq 1 - 2e^{-(n-1)/16}. \quad (2.42)$$

Let  $\{Y_i\}_{i=1}^{2n-2}$  be i.i.d. random vectors following the multivariate normal distribution  $N(\mathbf{0}, \Sigma)$ . Then

$$\begin{aligned}\widehat{\Sigma} &\stackrel{d}{=} \frac{1}{2n-2} \sum_{i=1}^{2n-2} Y_i Y_i^\top, \\ \beta^{*\top} \widehat{\Sigma} \beta^* &\stackrel{d}{=} \frac{1}{2n-2} \sum_{i=1}^{2n-2} (\beta^{*\top} Y_i)^2,\end{aligned}$$

where  $\stackrel{d}{=}$  denotes equal in distribution. Note that  $\{\beta^{*\top} Y_i\}$  are i.i.d Gaussian r.v.s following distribution  $N(0, \beta^{*\top} \Sigma \beta^*)$ , thus  $\{(\beta^{*\top} Y_i)^2\}$  are i.i.d. sub-exponential r.v.s,

so for any  $t \in (0, \beta^{*\top} \Sigma \beta^*)$ , we have

$$\mathbb{P} \left( \left| \frac{1}{2n-2} \sum_i (\beta^{*\top} Y_i)^2 - \beta^{*\top} \Sigma \beta^* \right| \geq t \right) \leq 2 \exp \left\{ -\frac{(2n-2)t^2}{8(\beta^{*\top} \Sigma \beta^*)^2} \right\}.$$

Relation (2.42) follows directly by taking  $t = \frac{1}{2} \beta^{*\top} \widehat{\Sigma} \beta^*$ , and thus part (i) of Lemma 2.6.1 holds.

(iii) Now we need to show that

$$\begin{aligned} \mathbb{P}(\mathcal{E}_{\sigma_{\max}}) &= \mathbb{P} \left( |\widehat{\sigma}_{\max}^2 - \sigma_{\max}^2| \leq \frac{1}{2} \sigma_{\max}^2 \right) \\ &\geq 1 - 2pe^{-(n-1)/16}. \end{aligned}$$

To prove this, we set  $\beta^* = e_j$  for  $j \in [p]$  and use (2.42) with a union bound argument to obtain that

$$\mathbb{P} \left( |\widetilde{\Sigma}_{j,j} - \Sigma_{j,j}| \leq \frac{1}{2} \Sigma_{j,j}, \forall j \in [p] \right) \geq 1 - 2pe^{-(n-1)/16}, \quad (2.43)$$

where the event on the left-hand side implies that  $|\widetilde{\sigma}_{\max}^2 - \sigma_{\max}^2| \leq \frac{1}{2} \sigma_{\max}^2$ .

(iv) The lower bound of  $\mathbb{P}(\mathcal{E}_1)$  follows an argument in [13]. Since  $\beta^* = \Sigma^{-1} \mu_d$ , we have that  $\widehat{\Sigma} \beta^* - \widehat{\mu}_d = (\widehat{\Sigma} - \Sigma) \beta^* - (\widehat{\mu}_d - \mu_d)$ . By A.5.1 in the supplement of [13], we have that

$$\mathbb{P} \left( |e_j^\top (\widehat{\Sigma} - \Sigma) \beta^*| \leq 10 \sigma_{\max} \Delta \sqrt{\frac{\log p}{n}}, \forall j \in [p] \right) \geq 1 - 2p^{-1}, \quad (2.44)$$

where the event on the left-hand side is equivalent to event  $\mathcal{E}_1$ . Furthermore, recall that  $\Delta^2 = \beta^{*\top} \Sigma \beta^*$ . Therefore, under events  $\mathcal{E}_\tau$  and  $\mathcal{E}_{\sigma_{\max}}$ , we have  $\Delta \leq \sqrt{2} \tau^*$  and  $\sigma_{\max} \leq \sqrt{2} \widehat{\sigma}_{\max}$ . These two conditions and event  $\mathcal{E}_1$  together imply  $\mathcal{E}_2$ . ■

*Proof of Lemma 2.6.3*

**Proof** Define  $\widehat{\omega}_j^{(i)} = \frac{\widehat{w}_j^{(i)}}{\sum_{i,j} \widehat{w}_j^{(i)}}$  and  $\widetilde{\omega}_j^{(i)} = \frac{\widetilde{w}_j^{(i)}}{\sum_{i,j} \widetilde{w}_j^{(i)}}$ . Also, define  $\widehat{\omega}^{(i)} = \sum_j \widehat{\omega}_j^{(i)}$  and  $\widetilde{\omega}^{(i)} = \sum_j \widetilde{\omega}_j^{(i)}$  for  $i = 0, 1$ . Note that by our construction, we have  $\frac{1-c_2}{2-c_2} \leq \widehat{\omega}^{(i)} \leq \frac{1}{2-c_2}$  and  $\frac{1-\epsilon}{2-\epsilon} \leq \widetilde{\omega}^{(i)} \leq \frac{1}{2-\epsilon}$ . For any  $p$ -dimensional vector  $v$  with  $\|v\|_2 = 1$ , we have that

$$\begin{aligned}
& |v^\top (\widehat{\Sigma} - \widetilde{\Sigma}) v| \\
&= \left| \sum_{i,j} \widehat{\omega}_j^{(i)} v^\top (x_j^{(i)} - \widehat{\mu}^{(i)}) (x_j^{(i)} - \widehat{\mu}^{(i)})^\top v \right. \\
&\quad \left. - \sum_{i,j} \widetilde{\omega}_j^{(i)} v^\top (\widetilde{x}_j^{(i)} - \widetilde{\mu}^{(i)}) (\widetilde{x}_j^{(i)} - \widetilde{\mu}^{(i)})^\top v \right| \\
&\leq \left| \sum_{i,j} \widehat{\omega}_j^{(i)} v^\top (x_j^{(i)} - \widetilde{\mu}^{(i)}) (x_j^{(i)} - \widetilde{\mu}^{(i)})^\top v \right. \\
&\quad \left. - \sum_{i,j} \widetilde{\omega}_j^{(i)} v^\top (\widetilde{x}_j^{(i)} - \widetilde{\mu}^{(i)}) (\widetilde{x}_j^{(i)} - \widetilde{\mu}^{(i)})^\top v \right| \\
&\quad + \sum_{i,j} \widehat{\omega}_j^{(i)} v^\top (\widehat{\mu}^{(i)} - \widetilde{\mu}^{(i)}) (\widehat{\mu}^{(i)} - \widetilde{\mu}^{(i)})^\top v \\
&\leq \frac{2 \max(\epsilon, c_2)}{1 - \min(\epsilon, c_2)} \|\widetilde{\Sigma}\|_2^2 + 2\epsilon R^2 \\
&\quad + c_3^2 \sigma^2 \left( \sqrt{\frac{1}{1-\epsilon}} + \sqrt{\frac{1}{1-c_2}} \right)^2 \frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}. \tag{2.45}
\end{aligned}$$

Using Theorem 6.1 in [45], under our sample condition (2.20), we have that

$$\mathbb{P} \left( \|\widetilde{\Sigma}\|_2 \leq 2\sigma^2 \right) \geq 1 - \exp(-n/8)$$

Putting it together, we finally obtain from (2.45) that

$$\|\widehat{\Sigma} - \widetilde{\Sigma}\|_2 \leq \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}} (3(1+c_3)\sigma + 2R).$$

This completes our proof for Lemma 2.6.3. ■

*Proof of Lemma 2.6.4*

**Proof** We define  $\tau^* = \sqrt{\beta^{*\top} \widehat{\Sigma} \beta^*}$ . We first show that  $(\widehat{h}, \widehat{\mu}, \beta^*, \tau^*)$  is a feasible solution. To see this, it suffices to verify the constraint (2.10) is satisfied. Note that under event  $\mathcal{E}_2$  and condition (2.20), we have

$$\begin{aligned}
& \|\widehat{\Sigma} \beta^* - \widehat{\mu}_d\|_\infty \\
&= \|(\widetilde{\Sigma} \beta^* - \widetilde{\mu}_d) + (\widehat{\Sigma} - \widetilde{\Sigma}) \beta^* + (\widehat{\mu}_d - \widetilde{\mu}_d)\|_\infty \\
&\leq \|\widetilde{\Sigma} \beta^* - \widetilde{\mu}_d\|_\infty + \|\widehat{\Sigma} - \widetilde{\Sigma}\|_2 \|\beta^*\|_2 + \|\widehat{\mu}_d - \widetilde{\mu}_d\|_2 \\
&\leq 20\sigma \sqrt{\frac{\log p}{(1-\epsilon)n}} \sqrt{\beta^{*\top} \widetilde{\Sigma} \beta^*} + \|\widehat{\Sigma} - \widetilde{\Sigma}\|_2 \|\beta^*\|_2 + \|\widehat{\mu}_d - \widetilde{\mu}_d\|_2 \\
&\leq 20\sigma \sqrt{\frac{\log p}{(1-\epsilon)n}} \left( \sqrt{\beta^{*\top} \widehat{\Sigma} \beta^*} + \|\beta^*\|_2 \sqrt{\|\widetilde{\Sigma} - \widehat{\Sigma}\|_2} \right) \\
&\quad + \|\widehat{\Sigma} - \widetilde{\Sigma}\|_2 \|\beta^*\|_2 + \|\widehat{\mu}_d - \widetilde{\mu}_d\|_2 \\
&\leq 40\sigma \sqrt{\frac{\log p}{(1-\epsilon)n}} + \sqrt{M} \Delta \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}} (3(1+c_3)\sigma + 2R) \\
&\quad + c_3\sigma \left( \sqrt{\frac{1}{1-\epsilon}} + \sqrt{\frac{1}{1-c_2}} \right) \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}.
\end{aligned}$$

Here we use Lemma 2.6.2 and the fact that  $\|\beta^*\| \leq \sqrt{M} \Delta$ . Also, note that when

$$|\beta^{*\top} (\widehat{\Sigma} - \widetilde{\Sigma}) \beta^*| \leq \frac{1}{2} \beta^{*\top} \widetilde{\Sigma} \beta^* = \widetilde{\tau}^2,$$

we have  $\tau^{*2} \geq \frac{1}{2}\tilde{\tau}^2 \geq \frac{1}{4}\Delta^2$  and  $\tau^{*2} \leq \frac{3}{2}\tilde{\tau}^2 \leq \frac{9}{4}\Delta^2$  under event  $\mathcal{E}_\tau$ . Therefore, under our choice of  $c_4$  such that

$$c_4 = 40\sigma\sqrt{\frac{\log p}{(1-\epsilon)n}} + 2\sqrt{M}\sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}(3(1+c_3)\sigma + 2R) \\ + \frac{2c_3\sigma}{\Delta} \left( \sqrt{\frac{1}{1-\epsilon}} + \sqrt{\frac{1}{1-c_2}} \right) \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}},$$

$(\hat{h}, \hat{\mu}, \beta^*, \tau^*)$  satisfies constraint (2.10), so that it is indeed a feasible solution.

Notice that with fixed  $(\hat{h}, \hat{\mu})$ , the optimization

$$\min_{\beta, \tau} \|\beta\|_1 + c_1\tau^2, \quad (2.46)$$

$$\text{subject to } \|\hat{\Sigma}(\hat{h}, \hat{\mu})\beta - (\hat{\mu}^{(1)} - \hat{\mu}^{(0)})\|_\infty \leq c_4\tau, \quad (2.47)$$

$$\sqrt{\beta^\top \hat{\Sigma}(\hat{h}, \hat{\mu})\beta} \leq \tau \quad (2.48)$$

is convex. Since  $(\hat{h}, \hat{\mu}, \hat{\beta}, \hat{\tau})$  is a local optima, we have

$$\|\hat{\beta}\|_1 + c_1\hat{\tau}^2 \leq \|\beta^*\|_1 + c_1\tau^{*2}.$$

Let  $S = \text{supp}(\beta)$  and  $\delta = \hat{\beta} - \beta^*$ . We have

$$\|\hat{\beta}_S\|_1 - \|\beta^*_S\|_1 + \|\hat{\beta}_{S^c}\|_1 \\ \leq 2c_1\beta^{*\top}\hat{\Sigma}(\beta^* - \hat{\beta}) \leq 2c_1\|\hat{\Sigma}\beta^*\|_\infty\|\delta\|_1 \\ \leq 2c_1(\|\mu_d\|_\infty + 10\sigma\Delta\sqrt{\frac{\log p}{(1-\epsilon)n}} + \|\hat{\Sigma} - \tilde{\Sigma}\|_2\|\beta^*\|_\infty)\|\delta\|_1.$$

Here the last inequality uses event  $\mathcal{E}_1$ . When  $c_1$  satisfies (2.18) we have  $\|\delta_{S^c}\|_1 \leq 3\|\delta_S\|_1$ .

This completes our proof for Lemma 2.6.4. ■

*Proof of Lemma 2.6.5*

**Proof** When  $(\widehat{h}, \widehat{\mu}, \beta^*, \tau^*)$  is feasible to (2.6), from constraint (2.10) we have

$$\|\widehat{\Sigma}\delta\|_\infty \leq \|\widehat{\Sigma}\widehat{\beta} - \widehat{\mu}_d\|_\infty + \|\widehat{\Sigma}\beta^* - \widehat{\mu}_d\|_\infty \leq c_4(\widehat{\tau} + \tau^*). \quad (2.49)$$

In addition, from

$$\|\widehat{\beta}\|_1 + c_1\widehat{\tau}^2 \leq \|\beta\|_1 + c_1\tau^{*2}$$

we have

$$\widehat{\tau} \leq \tau^* + \sqrt{\|\delta\|/c_1}. \quad (2.50)$$

We combine (2.49) and (2.50), and use  $\tau^* \leq \frac{3}{2}\Delta$  to obtain

$$\|\widehat{\Sigma}\delta\|_\infty \leq c_4(3\Delta + \sqrt{\|\delta\|_1/c_1}) \quad (2.51)$$

Finally, using Hölder's inequality, we have

$$\delta^\top \widehat{\Sigma}\delta \leq c_4(3\Delta + \sqrt{\|\delta\|_1/c_1})\|\delta\|_1, \quad (2.52)$$

which completes our proof. ■

*Proof of Lemma 2.6.6*

**Proof** In this proof, we will use the following result in [19].

**Lemma 2.6.9 (Theorem 1 of [19])** *For any Gaussian random design  $Z \in \mathbb{R}^{n \times p}$  with i.i.d.*

$N(\mathbf{0}, \Sigma)$  rows, there exist absolute positive constants  $C_1, C_2$  such that

$$\frac{\|Z\delta\|_2}{\sqrt{n}} \geq \frac{1}{4} \|\Sigma^{1/2}\delta\|_2 - 9\sigma \sqrt{\frac{\log p}{n}} \|\delta\|_1, \quad \forall \delta \in \mathbb{R}^p,$$

with probability at least  $1 - C_1 \exp(-C_2 n)$ .

Now we ready to prove Lemma 2.6.6. Suppose  $n \geq 2$ . Then we have

$$\tilde{\Sigma} = \frac{1}{\sum_{i,j} \tilde{w}_j^{(i)}} \left[ \sum_{i,j} \tilde{w}_j^{(i)} \left( x_i^{(0)} - \tilde{\mu}^{(0)} \right) \left( x_i^{(0)} - \tilde{\mu}^{(0)} \right)^\top \right],$$

where  $\tilde{w}_j^{(i)} \in \{0, 1\}$ .  $\hat{\Sigma}$  has the same distribution as

$$\tilde{\Sigma} = \frac{1}{N} \sum_{j=1}^{N-2} z_j z_j^\top,$$

where  $Z_j$ 's are i.i.d. samples from  $N(0, \Sigma)$ , and  $N = \sum_{i,j} \tilde{w}_j^{(i)} \geq 2(1 - \epsilon)n$ . Hence  $\hat{\Sigma}$  can be viewed as a sample covariance matrix of a Gaussian random design with 0 mean, up to a factor of  $(N - 2)/N$ .

By Lemma 2.6.9, there exist absolute positive constants  $C_1$  and  $C_2$  such that with probability at least  $1 - C_1 \exp(-C_2 n)$ ,

$$\|\tilde{\Sigma}^{1/2}\delta\|_2 \geq \frac{1}{4} \|\Sigma^{1/2}\delta\|_2 - 9\sigma \sqrt{\frac{\log p}{2(1 - \epsilon)n}} \|\delta\|_1.$$

When  $n \geq 2$  and  $\lambda_{\min}(\Sigma) \geq M^{-1}$ , we have

$$\|\tilde{\Sigma}^{1/2}\delta\|_2 \geq \frac{1}{4\sqrt{M}} \|\delta\|_2 - 9\sigma \sqrt{\frac{\log p}{(1 - \epsilon)n}} \|\delta\|_1,$$

and thus

$$\begin{aligned}\delta^\top \tilde{\Sigma} \delta &\geq \left( \frac{1}{4\sqrt{M}} \|\delta\|_2 - 9\sigma \sqrt{\frac{\log p}{(1-\epsilon)n}} \|\delta\|_1 \right)^2 \\ &\geq \frac{1}{32M} \|\delta\|_2^2 - 81\sigma^2 \frac{\log p}{(1-\epsilon)n} \|\delta\|_1^2.\end{aligned}$$

Here the last inequality follows from the fact that

$$(a-b)^2 = \left( \frac{1}{2}a^2 - 2ab + 2b^2 \right) + \frac{1}{2}a^2 - b^2 \geq \frac{1}{2}a^2 - b^2$$

for any number  $a, b \geq 0$ . Finally, we have

$$\begin{aligned}\delta^\top \hat{\Sigma} \delta &\geq \delta^\top \tilde{\Sigma} \delta - \|\hat{\Sigma} - \tilde{\Sigma}\|_2 \|\delta\|_2^2 \\ &\geq \left( \frac{1}{32M} - \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}} (3(1+c_3)\sigma + 2R) \right) \|\delta\|_2^2 \\ &\quad - 81\sigma^2 \frac{\log p}{(1-\epsilon)n} \|\delta\|_1^2.\end{aligned}$$

This completes our proof for Lemma 2.6.6. ■

*Proof of Lemma 2.6.7*

**Proof** Note that

$$\begin{aligned}\hat{\Delta} &= \sqrt{\hat{\beta}^\top \Sigma \hat{\beta}} = \sqrt{\beta^\top \Sigma \beta + 2\beta^\top \Sigma \delta + \delta^\top \Sigma \delta} \\ &\leq \sqrt{\beta^\top \Sigma \beta} \left( 1 + \frac{2\beta^\top \Sigma \delta + \delta^\top \Sigma \delta}{2\beta^\top \Sigma \beta} \right) \\ &= \Delta + \frac{2\mu_d^\top \delta + \delta^\top \Sigma \delta}{2\Delta}.\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} &= \frac{1}{2\widehat{\Delta}} (\Delta\widehat{\Delta} - \mu_d^\top \widehat{\beta}) \\
&\leq \frac{1}{2\widehat{\Delta}} \left( \Delta^2 + \mu_d^\top (\delta - \widehat{\beta}) + \frac{1}{2} \delta^\top \Sigma \delta \right) \\
&= \frac{1}{4\widehat{\Delta}} \delta^\top \Sigma \delta \leq \frac{\delta^\top \Sigma \delta}{4(\Delta + \frac{\mu_d^\top \delta}{\Delta})}.
\end{aligned} \tag{2.53}$$

Note that  $|\mu_d^\top \delta| \leq \|\mu_d\|_2 \|\delta\|_2 \leq \sigma \Delta \|\delta\|_2$ . Using the convergence rate of  $\|\delta\|_2$  in (2.22) from Theorem 2.4.1, when (2.35) and (2.36) are satisfied, we have that  $|\mu_d^\top \delta| \leq \Delta^2/2$ , and thus it follows from (2.53) that

$$\frac{\Delta}{2} - \frac{\mu_d^\top \widehat{\beta}}{2\widehat{\Delta}} \leq \frac{\delta^\top \Sigma \delta}{2\Delta} \leq \frac{\sigma^2}{2\Delta} \|\delta\|_2^2.$$

■

*Proof of Lemma 2.6.8*

**Proof** First, note that  $\widetilde{x}_j^{(i)}$ 's are i.i.d.  $\sigma^2$ -subgaussian, so that we have

$$P \left( \|\widetilde{\mu}^{(i)} - \mu^{(i)}\|_2 \leq \sigma \sqrt{\frac{2 \log p}{(1-\epsilon)n}}, i = 0, 1 \right) \geq 1 - 4p^{-1}. \tag{2.54}$$

Also, from Lemma 2.6.2, we have

$$\|\widehat{\mu}^{(i)} - \widetilde{\mu}^{(i)}\|_2 \leq c_3 \sigma \left( \sqrt{\frac{1}{1-\epsilon}} + \sqrt{\frac{1}{1-c_2}} \right) \sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}. \tag{2.55}$$

Combining (2.54) and (2.55) gives us an upper bound for  $\|\widehat{\mu}^{(i)} - \mu^{(i)}\|_2$  for  $i = 0, 1$ .

We now bound the term  $\widehat{\beta}^\top(\mu^{(i)} - \widehat{\mu}^{(i)})$ . Note that

$$\begin{aligned}\widehat{\beta}^\top(\mu^{(i)} - \widehat{\mu}^{(i)}) &= (\beta^* + \delta)^\top(\mu^{(i)} - \widehat{\mu}^{(i)}) \\ &\leq (\|\beta^*\|_2 + \|\delta\|_2)\|\mu^{(i)} - \widehat{\mu}^{(i)}\|_2 \\ &\leq (M^{1/2}\Delta + \|\delta\|_2)\|\mu^{(i)} - \widehat{\mu}^{(i)}\|_2.\end{aligned}$$

When  $n$  satisfies that

$$n \geq \frac{C}{1-\epsilon} M\sigma^2 s \log p$$

for some constant  $C$ , and

$$\sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}(3(1+c_3)\sigma + 2R) \leq \frac{1}{C'M\sqrt{s}}$$

for some constant  $C'$ , we have that

$$\begin{aligned}&\widehat{\beta}^\top(\mu^{(i)} - \widehat{\mu}^{(i)}) \\ &\leq 2M^{1/2}\Delta\sigma\sqrt{\frac{2\log p}{(1-\epsilon)n}} + 2c_3M^{1/2}\Delta\sigma\left(\sqrt{\frac{1}{1-\epsilon}} + \sqrt{\frac{1}{1-c_2}}\right)\sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}}.\end{aligned}\quad (2.56)$$

We then consider the term  $\widehat{\Delta} = \sqrt{\widehat{\beta}^\top \Sigma \widehat{\beta}}$ . Note that

$$\widehat{\Delta}^2 = \widehat{\beta}^\top \Sigma \widehat{\beta} = \Delta^2 + 2\mu_d^\top \delta + \delta^\top \Sigma \delta,$$

Hence we have

$$|\widehat{\Delta}^2 - \Delta^2| \leq 2\|\mu_d\|_2\|\delta\|_2 + \sigma^2\|\delta\|_2^2.$$

When

$$n \geq \frac{C}{1-\epsilon} M^2 \sigma^4 s \log p$$

and

$$\sqrt{\frac{\max(\epsilon, c_2)}{1-\epsilon-c_2}} (3(1+c_3)\sigma + 2R) \leq \frac{1}{C' M^{3/2} \sigma \sqrt{s}}$$

for some constants  $C$  and  $C'$ , we have  $|\widehat{\Delta}^2 - \Delta^2| \leq \frac{1}{2}\Delta^2$ . Combining this with (2.56), we have that

$$\begin{aligned} & \left( \frac{\widehat{\beta}^\top(\mu^{(0)} - \widehat{\mu}^{(0)}) + \widehat{\beta}^\top(\mu^{(1)} - \widehat{\mu}^{(1)})}{2\widehat{\Delta}} \right)^2 \\ & \leq C \cdot M\sigma^2 \left[ \frac{\log p}{(1-\epsilon)n} + c_3^2 \left( \frac{1}{1-\epsilon} + \frac{1}{1-c_2} \right) \frac{\max(\epsilon, c_2)}{1-\epsilon-c_2} \right] \end{aligned}$$

for some constant  $C$ . This completes our proof for Lemma 2.6.8. ■

## CHAPTER 3

### AN EFFICIENT ALGORITHM FOR SEQUENTIAL MATCHING PROBLEMS WITH UNKNOWN UTILITY

Motivated by the scenario of mentor-mentee matching in US colleges, in this chapter we introduce a useful algorithm for sequential assignment problems with unknown utility based on upper confidence bound (UCB) method and Hungarian algorithm. We derive the theoretical bounds of our algorithm for both the estimation error and the total regret, and numerical studies are also conducted to illustrate the usefulness of our algorithm.

#### 3.1 Introduction

The assignment problem is a classical problem in combinatoric optimization, with many classical real-world applications such as allocation of workers or resources for optimal utility gain. Under a general setup, it is assumed that we are given equal number of agents and tasks along with the utility associated with every possible agent-task pair, then we want to find the optimal assignment, i.e. a one-to-one mapping between the agents and tasks that yields maximal total utility. When the underlying utilities are known, the problem is well-studied in the combinatoric optimization literature. For instance, [46] first proposed the well-known Hungarian algorithm, which gives the optimal solution in polynomial time.

However, in many real-world applications, the underlying utility is often unknown and needs to be learned from data on the fly. The motivating example of our research is the mentor-mentee program of the Office of Alumni Relations (OAR) in many U.S. colleges and universities. Such mentoring programs are usually held every year or semester, and at each matching cycle, the Alumni office needs to decide how to suitably pair between mentors and mentees, so as to maximize the overall satisfaction constantly. At the end of the each program, a survey will be distributed to both mentors and mentees to collect

their feedback of satisfaction. The feedback can typically be regarded as binary outcome (i.e., satisfied/unsatisfied). In some colleges, the assignment is conducted manually, and the process can cost considerable human-labor and time every year. This inspires us to develop an algorithm for general sequential assignment problems with unknown utility and bandit feedback. Our goal is to learn the utility from data, and find near-optimal assignment plan for every round, so that the overall utility gain is maximized.

In this paper, we propose an efficient sequential assignment algorithm, with the objective of nearly maximizing the overall utility simultaneously for each round. Our algorithm consists of two key components at each time: 1) Constructing the upper confidence bound of the utility for every agent-task pair based on previous outcomes through logistic regression; 2) Conducting the assignment that maximizes the overall upper confidence bound of utility. By considering the upper confidence bound, our algorithm is able to balance the trade-off between exploration and exploitation, as compared to the greedy matching approach. Numerical experiments are conducted to show the usefulness of our algorithm. A further study of mentor-mentee matching scenario is discussed to illustrate our method.

Below it is useful to provide a brief literature review. Our work is closely related to the area of combinatorial semi-bandit, where each time the player need to pull a collection of arms (called super-arm) subjected to certain constraints, see for example [47, 48, 49]. In these studies, the agents and tasks are fixed at each round, and the player wants to learn the optimal assignment through bandit feedback of utility from matched pairs, which is different from our setting, where the agents and tasks can change constantly. Based on previous works, [50] investigate the contextual version of the combinatorial semi-bandit with a linear model for the payoff, and consider the general oracle algorithm for the combinatorial optimization problem as a sub-routine. Different from their study, our work utilizes the concrete Hungarian algorithm for the specific assignment problem at each time, combining with the logistic regression model to learn the utility function from binary outcomes. Another track of similar studies focus on a different goal of learning stable assignments

through the bandit feedback, see [51, 52, 53] for example.

Here we further note some fields related to our study. First, our problem involves parameter estimation in a sequential manner, which is related to the field of sequential estimation [54, 55], as well as online learning and optimization [56, 57, 58], while in our problem we further consider the optimal assignment based on the estimation, and want to nearly maximize the total utility through all times. Our research is also remotely related to the so-called reciprocal recommendation systems in applications such as online dating recommendation [59, 60], where the system recommends users potential partners based on their profiles, and learn the strategy for finding good pairs. Such a system gives a number of top recommendations for each user without conducting the assignments among users, as different from our setting.

The rest of the paper is organized as below. Section 3.2 introduces the problem formulation and relative background. Section 3.3 develops our proposed UCB-based sequential assignment algorithm, and Section 3.4 presents the theoretical results of our algorithm. Section 3.5 presents the results of numerical studies, and Section 3.6 includes a further study under the mentor-mentee matching scenario.

**Notations.** For  $n \in \mathbb{N}$ , we denote  $[n]$  the set  $\{1, 2, \dots, n\}$ . For a  $d$ -dimensional vector  $\mathbf{v} = (v_1, \dots, v_d)$ , we define the vector  $\ell_2$  norm  $\|\mathbf{v}\|_2^2 = \sum_{k=1}^d v_k^2$  and the matrix norm  $\|\mathbf{v}\|_M = \sqrt{\mathbf{v}^\top M \mathbf{v}}$ , where  $M$  is a  $d \times d$  positive definite matrix. We use  $P(\cdot)$  to denote the probability of events, and  $\mathbb{E}[\cdot]$  to denote the expectation of random variables. We use  $I(\cdot)$  to denote the indicator function.

## 3.2 Problem Formulation and Background

In this section, we first introduce the formulation of our sequential assignments with unknown utility, and then briefly introduce the optimal assignment with known utility and the classical Hungarian algorithm, which will be adapted into our method. Suppose that

initially (denoted as time 0) we are given past data of size  $n_0$ :

$$\mathcal{D}_0 := \{(\mathbf{x}_i^0, \mathbf{z}_i^0, U_i^0) : i \in [n_0]\},$$

where  $\mathbf{x}_i^0$ 's and  $\mathbf{z}_i^0$ 's are task and agent covariates, and  $U_i^0$ 's are corresponding outcome for every pair of  $(\mathbf{x}_i^0, \mathbf{z}_i^0)$ . For each time  $t = 1, 2, \dots, T$ , we are given  $n_t$  agents and tasks, where each agent or task is associated a vector of covariates, which is also referred to as side information or context. Let  $\{\mathbf{x}_i^t : i \in [n_t]\} \subset \mathcal{X}$  and  $\{\mathbf{z}_i^t : i \in [n_t]\} \subset \mathcal{Z}$  be the collection of the covariates of agents and tasks at time  $t$ , where  $\mathcal{X}$  and  $\mathcal{Z}$  are the spaces of the corresponding covariates. We then need to conduct an assignment, denoted by  $\delta_t$  between these agents and tasks. After the assignment being conducted, we observe the utility associated to every matched pair  $(\mathbf{x}_i^t, \mathbf{z}_{\delta_t(i)}^t)$ , denoted by  $U(\mathbf{x}_i^t, \mathbf{z}_{\delta_t(i)}^t)$ . Our goal is to maximize the total utility gained up to time  $T$ . We summarize the procedure in an online learning framework as below. For round  $t = 1, 2, \dots, T$ :

- (i) The player is given the agent and task covariates,  $\{\mathbf{x}_i^t : i \in [n_t]\}$  and  $\{\mathbf{z}_i^t : i \in [n_t]\}$ .
- (ii) The player decides an assignment, denoted by a one-to-one mapping  $\delta_t : [n_t] \rightarrow [n_t]$ .
- (iii) The player observes the utility feedback for every assigned pair  $U(\mathbf{x}_i^t, \mathbf{z}_{i, \delta_t(i)}^t)$ .

Loosely speaking, our goal is to decide the assignment  $\delta_t$  at every round  $t$  such that the overall expected utility  $\sum_{t=1}^T \sum_{i=1}^{n_t} \mathbb{E} \left[ U(\mathbf{x}_i^t, \mathbf{z}_{\delta_t(i)}^t) \right]$  is maximized.

While the underlying utility of any agent-task pair is unknown by the time of assignment, we assume that the utility is related to the covariates of the agent and task through some noisy function. More specifically, for any pair of covariates  $(\mathbf{x}_i^t, \mathbf{z}_j^t)$ , we suppose that the associated utility  $U(\mathbf{x}_i^t, \mathbf{z}_j^t)$  is binary that satisfies

$$\log \frac{P(U(\mathbf{x}_i^t, \mathbf{z}_j^t) = 1)}{P(U(\mathbf{x}_i^t, \mathbf{z}_j^t) = 0)} = \sum_{k=1}^d x_{i,k}^t z_{j,k}^t \theta_k^*, \quad (3.1)$$

where  $\theta^*$  is a  $d$ -dimensional unknown parameter,  $x_{i,k}^t$  is the  $k$ -th entry of the vector  $\mathbf{x}_i^t$ , and likewise for  $z_{j,k}^t$  and  $\theta_k^*$ . By defining  $\circ$  the entry-wise product, we can rewrite the right-hand side in (3.1) as  $(\mathbf{x}_i^t \circ \mathbf{z}_j^t)^\top \theta^*$ .

We further define

$$u(\mathbf{x}_i^t, \mathbf{z}_j^t) = \mathbb{E}[U(\mathbf{x}_i^t, \mathbf{z}_j^t)] = \frac{\exp((\mathbf{x}_i^t \circ \mathbf{z}_j^t)^\top \theta^*)}{1 + \exp((\mathbf{x}_i^t \circ \mathbf{z}_j^t)^\top \theta^*)}.$$

as the expectation of the utility associated with the pair. For notational simplicity, in the rest of the paper, we may use the shorthands  $U_{i,j}^t = U(\mathbf{x}_i^t, \mathbf{z}_j^t)$  and  $u_{i,j}^t = u(\mathbf{x}_i^t, \mathbf{z}_j^t)$  when there is no ambiguity. Also, we define the notation  $\phi_{i,j}^t = (\mathbf{x}_i^t \circ \mathbf{z}_j^t)$ , which will be used in the rest of the paper.

Note that in our work, we assume that  $\mathbf{x}$  and  $\mathbf{z}$  has same dimension, and the utility is related to these covariates through  $\mathbf{x} \circ \mathbf{z}$ . This setting is natural in scenarios (such as mentor-mentee matching) with specific data form, where the preference or attributes of the two parts are aligned well. For other applications, our method need to be extended to accommodate more general form of the utility functions. Nevertheless, finding the suitable transformation function class can be challenging in general, as it usually depends on the problem and data implicitly. In this case one might consider non-parametric models to approximate the underlying utility function, or conduct variable selection through the learning process, which is of independent interest to our work.

To measure the loss of assignments conducted by certain algorithm  $\mathcal{A}$ , we first define the oracle assignment  $\delta_t^*$  at each round  $t$  as the assignment that maximizes the total expected utility, i.e.,

$$\delta_t^* \in \arg \max_{\delta} \sum_{i=1}^{n_t} u(\mathbf{x}_i^t, \mathbf{z}_{\delta(i)}^t).$$

Note that since  $u$  depends on the unknown parameter  $\theta^*$ ,  $\delta_t^*$  is also unknown in practice when one conducts the assignment. With the definition of  $\delta_t^*$ , we further define the cumu-

lative regret of an algorithm  $\mathcal{A}$  to be

$$R_T(\mathcal{A}) = \sum_{t=1}^T \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} \left[ u(\mathbf{x}_i^t, \mathbf{z}_{\delta_t^*(i)}^t) - u(\mathbf{x}_i^t, \mathbf{z}_{\delta_t(i)}^t) \right] \right\}, \quad (3.2)$$

where  $\delta_t$  is the assignment conducted by  $\mathcal{A}$ , and  $\delta_t^*$  is the oracle assignment at time  $t$ . By the definition of  $\delta^*$ ,  $R_T(\mathcal{A})$  captures the performance gap between  $\mathcal{A}$  and the oracle performance in expected utility. Ideally, one wish to design an algorithm with the total regret  $R_T$  as small as possible, at least sublinear in  $T$ .

Note that in (3.2), at each time we scale the utility gap by the number of pairs  $n_t$ , so that we eliminate the difference of the size across time and focus on the average utility gap. Alternatively, one might also consider the regret defined as the summation of the total utility gap directly, without taking the average at each time. The choice of the criteria for regret can depend on the specific problem of interest. In this paper we will discuss the regret bound with regret defined as in (3.2), while the result can be naturally adapted to the alternative definition.

### 3.2.1 Optimal assignment with known utilities

In this section, we briefly review the optimal assignment problem with known utilities and the Hungarian algorithm, which will be combined in our method for deciding the assignment at each time.

A typical optimal assignment problem can be expressed as follows. Suppose there are  $n$  agents and  $n$  tasks. Let  $U$  being a  $n \times n$  matrix of utility, where  $U_{i,j}$  is the utility gain when assigning agent  $i$  to task  $j$ . The objective is to find the optimal assignment plan (i.e., a one-to-one mapping between agents and tasks), such that the total utility gain is maximized.

Mathematically, one solves the following optimization problem for optimal assignment:

$$\begin{aligned}
& \max_{\delta_{i,j}} \sum_{i \in [n]} \sum_{j \in [n]} U_{i,j} \delta_{i,j} & (3.3) \\
& \text{subject to} \quad \sum_{i=1}^n \delta_{i,j} = 1, \quad \forall j, \\
& \quad \quad \quad \sum_{j=1}^n \delta_{i,j} = 1, \quad \forall i, \\
& \quad \quad \quad \delta_{i,j} \in \{0, 1\}, \quad \forall i, j.
\end{aligned}$$

Here  $\delta_{i,j}$ 's are binary decision variables, and  $\delta_{i,j} = 1$  indicates that  $i$  is assigned to task  $j$ , and  $\delta_{i,j} = 0$  otherwise.

We now introduce the Hungarian algorithm, which is a classical one for solving the optimal assignment above, developed by [46]. In the high level, it considers the dual problem of (3.3), which can be expressed as below:

$$\begin{aligned}
& \min_{u,v} \sum_{i \in [n]} u_i + \sum_{j \in [n]} v_j, \\
& \text{subject to} \quad u_i + v_j \geq U_{i,j}, \quad \forall i, j.
\end{aligned}$$

The algorithm utilizes the primal-dual method to update the solution until the optimal objective is reached for both primal and dual problems. Algorithm 3 below describes the classical Hungarian algorithm in [46] in the matrix form.

Note that the computational complexity of Algorithm 3 is  $O(n^4)$ . The classical Hungarian algorithm is later improved by [61] and [62] to achieve an  $O(n^3)$  complexity. In addition, there are many approximation algorithms for optimal assignment with less computational cost and near optimality, see [63, 64, 65] among others. One might consider adopt such approximation algorithms for assignment problems with large scales.

---

**Algorithm 3** The Hungarian algorithm for optimal assignment with known utility

---

Specify  $n$  the number of agents and tasks; Specify the utility matrix  $U \in \mathbb{R}^{n \times n}$ . Set the matrix  $C \in \mathbb{R}^{n \times n}$  such that  $C_{i,j} = \max_{k,\ell} U_{k,\ell} - U_{i,j}$ .

- Step i. Subtract  $\min_{i,j} C_{i,j}$  from each element of  $C$  and obtain a matrix  $C_1$ .
  - Step ii. Find a minimum set  $S_1$  of lines (rows or columns) that includes all null elements in  $C_1$ . Let  $n_1 = |S_1|$ . If  $n_1 = n$ , then report the  $n$  positions of null elements as the required solution.
  - Step iii. If  $n_1 < n$ , let  $h_1$  be the smallest element in  $C_1$  that is not in any line in  $S_1$ . Add  $h_1$  to any elements in a line of  $S_1$  and subtract  $h_1$  from any elements in  $C_1$ . Let the resulting matrix be  $C_2$ .
  - Step iv. Repeat the Steps 2 and 3 starting with  $C_2$ , until  $n_k = n$  at some stage. Report the positions of these null elements.
- 

### 3.3 Our Proposed Method

In this section, we introduce our proposed UCB-based algorithm for the sequential assignment problem with unknown utility. At a high level, our method consists of two components for every round: 1) Constructing the upper confidence bound for every agent-task pair based on past data, and 2) Conducting the assignment by maximizing the total upper confidence bound of the utility. In order to obtain a reliable initial estimate for  $\theta^*$ , we conduct random assignment up to certain time  $t_1$  for pure exploration. Below we first describe our proposed UCB-based method in details, and later summarize our algorithm.

#### 3.3.1 Constructing upper confidence bound

In this subsection, we specify the construction of the confidence bound based on past data in our method. Loosely speaking, at the beginning of each round after certain time  $t_1$ , we use the logistic regression to fit the past observations and obtain the estimate for  $\theta^*$ , and construct the confidence bound accordingly. For logistic regression, we penalize the  $\ell_2$  norm of the estimator to avoid overfitting in early stages. More specifically, for every  $t > t_1$ , we first estimate the parameter  $\theta^*$  by maximizing the  $\ell_2$ -penalized log-likelihood

based on past observations  $\{(\mathbf{x}_i^\tau, \mathbf{z}_{\delta_\tau(i)}^\tau, U_{i,\delta_\tau(i)}^\tau) : \tau \in [t-1], i \in [n_\tau]\}$ , or,

$$\hat{\theta}^t \in \operatorname{argmin}_{\theta} \sum_{\tau=0}^{t-1} \sum_{i \in [n_i]} \left\{ \log(1 + \exp(\phi_{i,\delta_\tau(i)}^{\tau\top} \theta)) - U_{i,\delta_\tau(i)}^\tau \phi_{i,\delta_\tau(i)}^{\tau\top} \theta \right\} + \frac{r}{2} \|\theta\|_2^2,$$

where  $r$  is a penalty coefficient, and  $\delta_0(i) = i$  for consistency. Also recall that we define  $\phi_{i,j}^t = (\mathbf{x}_i^t \circ \mathbf{z}_j^t)$ . It is equivalent to solving the following equation to obtain  $\hat{\theta}^t$ :

$$\sum_{\tau=0}^{t-1} \sum_{i=1}^{n_\tau} \phi_{i,\delta_\tau(i)}^\tau \left( U_i^\tau - \frac{1}{1 + \exp(-\phi_{i,\delta_\tau(i)}^{\tau\top} \theta)} \right) = r\theta. \quad (3.4)$$

Let  $\hat{\theta}^t$  be the estimate of  $\theta^*$  at the beginning of time  $t$ . Then for any new pair  $(\mathbf{x}_i^t, \mathbf{z}_j^t)$ , we construct the following upper confidence bound  $b_{i,j}^t$  on the expected utility for the pair:

$$b_{i,j}^t = \frac{1}{1 + \exp \left\{ -\phi_{i,j}^{\tau\top} \hat{\theta}^t - \lambda \sqrt{\phi_{i,j}^{\tau\top} \bar{M}_{t-1}^{-1} \phi_{i,j}^\tau} \right\}}, \quad (3.5)$$

where  $\lambda$  is a parameter, and

$$\bar{M}_{t-1} = rI + \sum_{\tau=1}^{t-1} \sum_{i=1}^{n_\tau} \phi_{i,\delta_\tau(i)}^\tau \phi_{i,\delta_\tau(i)}^{\tau\top}.$$

### 3.3.2 Assignment by maximizing upper confidence bound

After constructing the upper confidence bound  $b_{i,j}^t$  for every possible agent-task pair  $(\mathbf{x}_i^t, \mathbf{z}_j^t)$  at time  $t$ , we conduct the assignment for time  $t$  by solving the following optimization prob-

lem:

$$\begin{aligned}
& \max_{\delta_{i,j}} \sum_{i \in [n_t]} \sum_{j \in [n_t]} b_{i,j}^t \delta_{i,j} & (3.6) \\
& \text{subject to} \quad \sum_{i=1}^{n_t} \delta_{i,j} = 1, \quad \forall j, \\
& \quad \quad \quad \sum_{j=1}^{n_t} \delta_{i,j} = 1, \quad \forall i, \\
& \quad \quad \quad \delta_{i,j} \in \{0, 1\}, \quad \forall i, j.
\end{aligned}$$

Here  $\delta_{i,j}$ 's are binary decision variables, where  $\delta_{i,j} = 1$  indicates  $\mathbf{x}_i^t$  and  $\mathbf{z}_j^t$  are matched with each other, and  $\delta_{i,j} = 0$  otherwise. Note that here we maximize the objective of the total upper confidence bound of the utility. To solve such optimization problems, we may apply the classical Hungarian algorithm, or adopt some approximation or heuristic algorithms with less computational cost, depending on the problem scale and demand. Our approach is inspired by the UCB-algorithm for classical multi-armed bandit and contextual bandit, as discussed in [66], [67] and [68] among others, where the player pull the arm with highest upper confidence bound of reward at each round. By considering the upper confidence bound instead of the utility estimate, one can balance the exploitation with exploration. In later sections, we will show that with proper choice of  $\lambda$ , our algorithm can balance the exploration and exploitation, and thus enjoys desired theoretical property in cumulative regret. To solve problem (3.6), one can utilize many off-the-shelf packages providing algorithms for solving the assignment problem, such as the classical Hungarian algorithm, or approximation algorithms.

In summary, our proposed algorithm conducts assignment based on the two subsections above. For better understanding, we present our method as Algorithm 4.

---

**Algorithm 4** UCB-based algorithm for sequential assignment with logistic model

---

Set  $\mathcal{D}_0 = \{(\mathbf{x}_i^0, \mathbf{z}_i^0, U_i^0) : i \in [n_0]\}$ . Specify the parameters  $t_1 > 0$  and  $\lambda > 0$ .

**for**  $t = 1$  to  $T$  **do**

Observe covariates  $\mathbf{x}_1^t, \dots, \mathbf{x}_{n_t}^t$  and  $\mathbf{z}_1^t, \dots, \mathbf{z}_{n_t}^t$ .

**if**  $t \leq t_1$  **then**

Conduct random assignment through a random permutation  $\delta_t$ .

**else**

Obtain  $\hat{\theta}^t$  by solving (3.4) based on  $\mathcal{D}_{t-1}$ .

Let  $\phi_{i,j}^t = (\mathbf{x}_i^t \circ \mathbf{z}_j^t)$  for every  $i$  and  $j$  in  $[n_t]$ . Construct the upper confidence bound  $b_{i,j}^t$  for the associated utility as

$$b_{i,j}^t = \frac{1}{1 + \exp\left(-\phi_{i,j}^{t\top} \hat{\theta}^t - \lambda \|\phi_{i,j}^t\|_{\bar{M}_{t-1}^{-1}}\right)}$$

where  $\bar{M}_{t-1} = rI + \sum_{\tau=1}^{t-1} \sum_{i=1}^{n_\tau} \phi_{i,j}^\tau \phi_{i,j}^{\tau\top}$ .

Finding the assignment  $\delta_t$  by solving the optimization problem (3.6), using a sub-routine such as the Hungarian algorithm.

**end if**

Observe feedback  $\{U_{i,\delta_t(i)}^t : i \in [n_t]\}$ .

Update

$$\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{(\mathbf{x}_i^t, \mathbf{z}_{\delta_t(i)}^t, U_{i,\delta_t(i)}^t) : i \in [n_t]\}.$$

**end for**

---

### 3.4 Theoretical Results

In this section, we discuss the theoretical property of our proposed algorithm on the estimation error and regret bound. Before we move on, we first introduce some mild assumptions for our results to hold, which are standard in the statistics and bandit literature, see for example [67, 68].

**Assumption 3.4.1** For every  $\mathbf{x} \in \mathcal{X}$  and  $\mathbf{z} \in \mathcal{Z}$ ,  $\|\mathbf{x} \circ \mathbf{z}\|_2 \leq 1$ . Meanwhile, we assume  $\|\theta^*\|_2 \leq 1$ .

**Assumption 3.4.2** We assume that for every  $t \geq 1$ ,  $\mathbf{x}_i^t$ 's and  $\mathbf{z}_j^t$ 's are i.i.d. sample drawn

from distributions  $\mathcal{P}_X$  and  $\mathcal{P}_Z$ , respectively. Furthermore,  $\mathcal{P}_X$  and  $\mathcal{P}_Z$  satisfies that

$$\lambda_{\min} \left( \mathbb{E}_{\mathbf{x} \sim \mathcal{P}_X, \mathbf{z} \sim \mathcal{P}_Z} [(\mathbf{x} \circ \mathbf{z})(\mathbf{x} \circ \mathbf{z})^\top] \right) \geq \sigma^2 > 0.$$

Below we first state the result on the estimation error of our proposed UCB-based algorithm in Theorem 3.4.1. It shows that with sufficient past observations, our estimates  $\hat{\theta}^t$  are pretty close to the underlying true parameter  $\theta^*$  with high probability.

**Theorem 3.4.1** *Suppose Assumptions 3.4.1 and 3.4.2 hold. Define*

$$\kappa = \inf_{\|\phi\|_2 \leq 1, \|\theta - \theta^*\|_2 \leq 1} \frac{\exp(\phi^\top \theta)}{[1 + \exp(\phi^\top \theta)]^2}.$$

*If there exists some  $t_1$  such that  $M_{t_1}$  satisfies that*

$$\lambda_{\min}(M_{t_1}) \geq \frac{32(d + \log(T/\delta))}{\kappa^2} + \frac{2r}{\kappa}, \quad (3.7)$$

*then with probability at least  $1 - 2\delta$ , we have*

$$\|\hat{\theta}^t - \theta^*\|_{M_{t-1}}^2 \leq \frac{1}{\kappa^2} \left[ d \log \left( 1 + \frac{\sum_{t=1}^T n_t}{rd} \right) + 2 \log \frac{1}{\delta} + r \right]$$

*for all  $t \geq t_1 + 1$ .*

Based on Theorem 3.4.1, our next theorem presents the result on the cumulative regret bound of our method.

**Theorem 3.4.2** *Suppose that Assumptions 3.4.1 and 3.4.2 hold. Then for any  $\delta \in (0, 1)$ , there exists universal constants  $C_1$ ,  $C_2$  and  $C_3$ , such that as long as*

$$\sum_{\tau=1}^{t_1} n_\tau \geq C_1 \frac{d + \log(1/\delta)}{\sigma^4} + C_2 \frac{d + \log(T/\delta)}{\kappa^2 \sigma^2} + C_3 \frac{r}{\kappa \sigma^2}, \quad (3.8)$$

with the choice that

$$\lambda = \frac{1}{\kappa} \sqrt{d \log \left( 1 + \frac{\sum_{t=1}^T n_t}{rd} \right) + 2 \log(1/\delta) + r}, \quad (3.9)$$

with probability at least  $1 - 3\delta$ , the total regret of Algorithm 4 satisfies that

$$R_T \leq t_1 + \frac{2d}{\kappa} \sqrt{T \log \left( 1 + \frac{\sum_{t=1}^T n_t}{d} \right)} \sqrt{\log \left( 1 + \frac{\sum_{t=1}^T n_t}{rd} \right) + 2 \log(1/\delta) + r}. \quad (3.10)$$

**Remark 3.4.1** Here the condition (3.8) is to guarantee that the minimum eigenvalue of  $M_{t_1}$  is sufficiently large under the worst case that there is no observation at time 0 (i.e.  $\mathcal{D}_0 = \emptyset$ ), such that with high probability  $\hat{\theta}^t$  is close to  $\theta^*$  for every  $t > t_1$ . Meanwhile, if we have sufficient observations at time 0 such that  $M_0$  is well conditioned, this requirement on  $t_1$  can be relaxed. Moreover, the parameter  $\lambda$  is important in our algorithm that trades off between exploration and exploitation. Specially, when  $\lambda = 0$ , the algorithm reduces to the greedy method, while when  $\lambda$  is very large, the algorithm tends to find the best design for  $\theta^*$  estimate. For most desired performance, the choice of  $\lambda$  can depend on data in practice.

As can be seen, our high-probability regret bound in (3.10) is of the rate  $\tilde{O}(d\sqrt{T})$  when neglecting the logarithm factors. This rate is consistent with the result of regret for the logistic bandit, as presented in [68]. We defer the proof for Theorems 3.4.1 and 3.4.2 to Sections 3.7.1 and 3.7.2 in Appendix.

### 3.5 Numerical Studies

In this section, we conduct simulations to demonstrate the usefulness of our proposed UCB-based algorithm. More specifically, we investigate the total regret and parameter estimation under several examples, with various data dimension  $d$  and choice of tuning parameter  $\lambda$ . For convenience, in most settings below we take  $n_t = n$  as a constant.

### 3.5.1 Settings

In this subsection, we specify the construction of our simulation examples. Specifically, we consider the following two settings of  $\theta^*$ .

(i)  $\theta^* = \frac{1}{\sqrt{d}}(1, 1, \dots, 1, -1, -1 \dots, -1)$ , where the first  $d/2$  entries positive and the second  $d/2$  entries negative.

(ii)  $\theta^* = \frac{1}{\sqrt{31}}(-1, 1, 2, 3, 4, 0, \dots, 0)$ , with  $\|\theta^*\|_2 = 1$ .

When generating the utility outcome, we add a random noise  $\epsilon$  that follows a normal distribution with mean 0 and variance 1. At each round, we randomly sample  $\mathbf{x}_i^t$ 's and  $\mathbf{z}_i^t$ 's from the multivariate normal distribution  $N(0, I_d)$ . While we fix  $T = 20$ ,  $n = 50$ , we vary the dimension  $d$  as 10 and 100. We set the penalty parameter  $r = 0.2$ . For every example, we have 20 replications of randomly sampled data, and finally we present the average performance. Note that we do not have history data at time 0, and we set  $t_1 = 1$ .

### 3.5.2 Performance

In this subsection, we compare the performance between our method and the  $\epsilon$ -greedy method with various  $\epsilon$  values. We first present the figures that characterize the growth rate of the cumulative regret in  $t$ , followed by a table with detailed performance for different  $d$  and  $\lambda$ . Figure 3.1 presents the cumulative regret of our algorithm with respect to  $t$  under settings (i) and (ii), with different choice of  $\lambda$ . As can be seen, the growth of the total regret is indeed sublinear in  $t$ . Also, note that with the choice  $\lambda = 0$ , then the algorithm is greedy that does pure exploitation. Tables 3.1 and 3.2 below provide more detailed numerical results, including the standard deviations for the total regret up to  $T = 20$  rounds under settings (a) and (b) with different  $d$  and  $\lambda$  averaged over 50 replications. Note that in these tables we also present the regret for the random match (i.e. assign by a random permutation with equal probability) as a baseline.

From the figures and tables, we can see that with a proper choice of  $\lambda$ , one can achieve a lower regret than the pure greedy method with  $\lambda = 0$ , by balancing the exploration and exploitation. Moreover, the performance is sensitive to the tuning parameter  $\lambda$ . When  $\lambda$  is too large, the performance can be worse-off. Meanwhile, we also observe that our proposed method takes more advantage when the unknown parameter dimension  $d$  is relatively small.

Table 3.1: Total regret up to  $T = 20$  under setting (a)

Regret (std)	$d = 10$	$d = 100$
Random Match	6.87 (0.12)	7.45 (0.13)
$\lambda = 0$ (Greedy)	0.72 (0.14)	3.07 (0.22)
$\lambda = 0.2$	0.63 (0.11)	3.08 (0.23)
$\lambda = 0.5$	0.64 (0.11)	3.06 (0.24)
$\lambda = 1.0$	0.64 (0.11)	3.04 (0.20)
$\lambda = 5$	0.77 (0.12)	3.22 (0.26)
$\lambda = 10$	0.94 (0.11)	3.40 (0.20)
$\lambda = 20$	1.44 (0.09)	3.98 (0.20)
$\lambda = 50$	2.65 (0.15)	5.17 (0.21)

Table 3.2: Total regret up to  $T = 20$  under setting (b)

Regret (std)	$d = 10$	$d = 100$
Random Match	6.29 (0.13)	7.39 (0.10)
$\lambda = 0$ (Greedy)	0.81 (0.19)	3.01 (0.27)
$\lambda = 0.2$	0.66 (0.12)	3.00 (0.26)
$\lambda = 0.5$	0.67 (0.16)	3.01 (0.21)
$\lambda = 1.0$	0.70 (0.15)	3.08 (0.23)
$\lambda = 5$	0.88 (0.14)	3.16 (0.25)
$\lambda = 10$	1.08 (0.13)	3.41 (0.18)
$\lambda = 20$	1.50 (0.10)	3.88 (0.18)
$\lambda = 50$	2.52 (0.10)	5.11 (0.18)

Furthermore, it is also worth noticing that while the choice of  $\lambda$  in (3.9) guarantees the theoretical property, in practice one might prefer to tune  $\lambda$  for better empirical performance. As for in this example, while the choice of  $\lambda$  suggested by (3.9) can be much larger than 1, the empirical performance seems better with a  $\lambda$  less than 1.

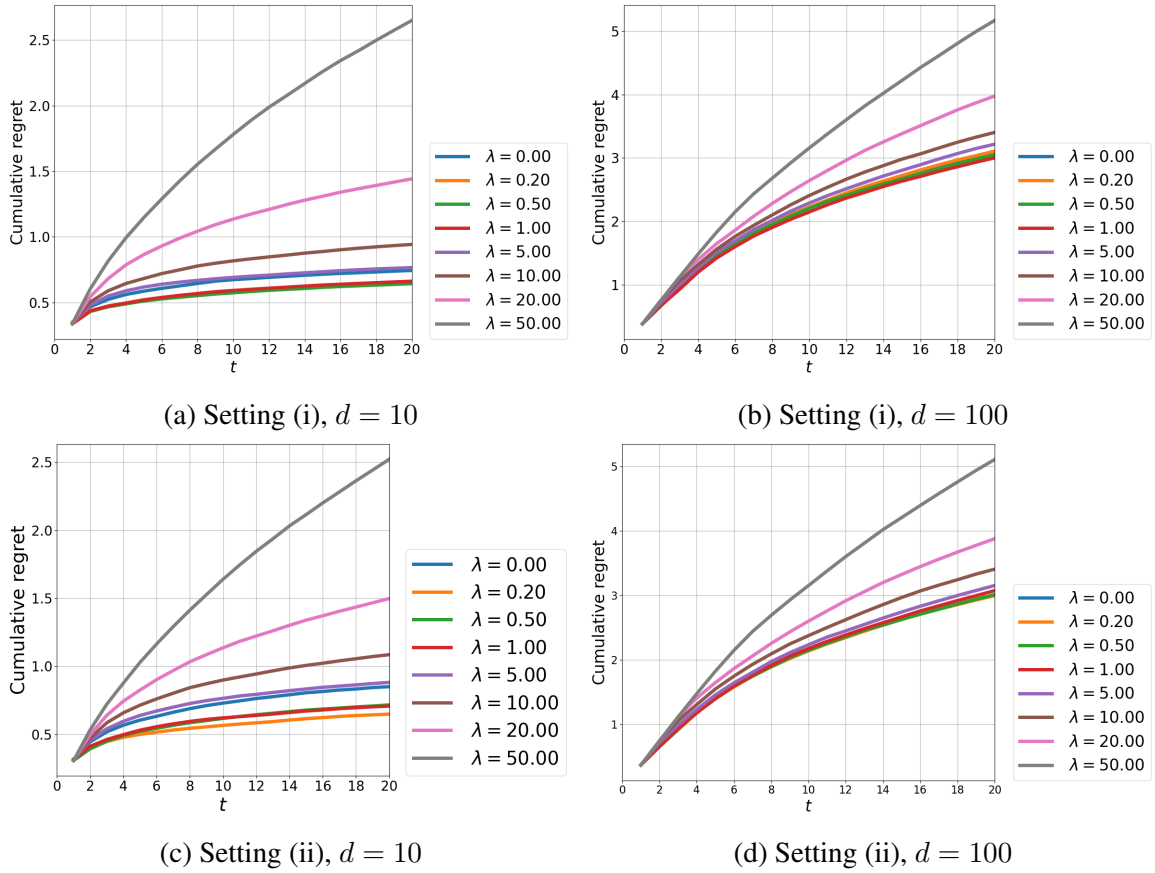


Figure 3.1: Cumulative regret with different  $d$  and  $\lambda$  under settings (i) and (ii). The curves of cumulative regret are sub-linear in  $t$ . With proper choice of  $\lambda$ , the performance is improved over the pure-greedy algorithm when  $\lambda = 0$ .

### 3.5.3 Comparison with $\epsilon$ -greedy method

In this subsection, we compare our method with the  $\epsilon$ -greedy method. By  $\epsilon$ -greedy method, at each time every agent is chosen with probability  $\epsilon$  to be paired with a random task, and then the remaining agents and tasks are assigned by the greedy method. Figure 3.2 shows the cumulative regret of the  $\epsilon$ -greedy method with  $\epsilon = 0, 0.1, 0.2, 1$  (note that  $\epsilon = 0$  reduces to the greedy algorithm, and  $\epsilon = 1$  reduces to random assignment at all times), together with our proposed method with  $\lambda = 0.5$ . The setting is same as (i), with  $d = 10$  and 100. As can be seen, the  $\epsilon$ -greedy method is worse than the pure-greedy method under this setting. Overall, these results validates the usefulness of our UCB-based assignment approach.

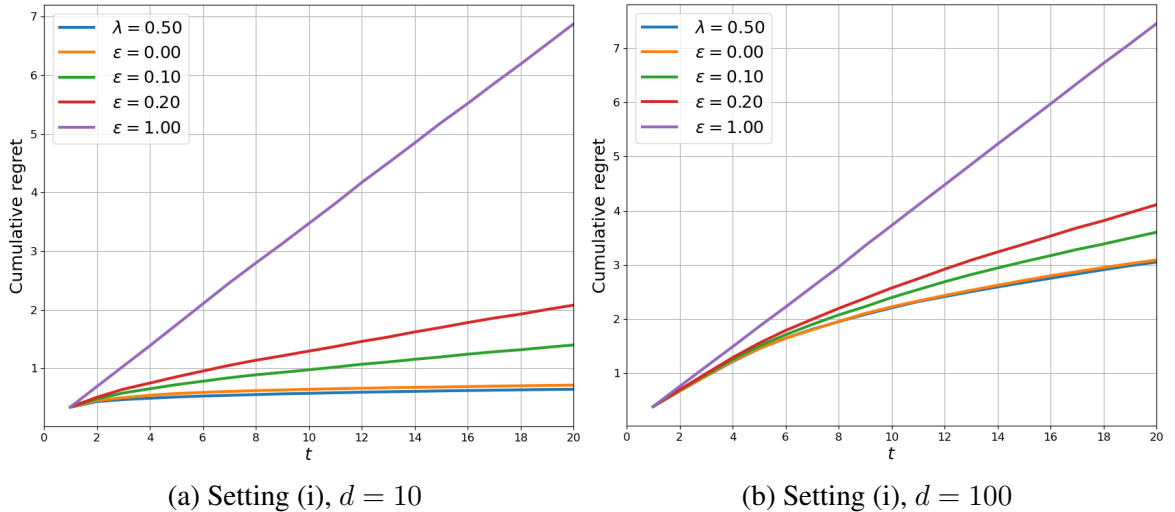


Figure 3.2: Cumulative regret for  $\epsilon$ -greedy method and our UCB-based method with  $\lambda = 0.5$  under settings (i) with  $d = 10, 100$ . The performance of  $\epsilon$ -greedy method is worse than the pure-greedy method under this setting.

## 3.6 Further Study

In this section, we go back to the motivating example of mentor-mentee matching for university mentoring programs, and illustrate the usefulness of our proposed algorithm.

At each semester/round, the university alumni office receives a number of mentees and mentors with context information on their background and preference. The office

then needs to decide how to pair between the mentees and mentors. At the end of each semester, a survey will be distributed to each participant to ask about their satisfaction on the experience. Due to the privacy constraints, we are unable to share the concrete dataset, but we will use similar data format to mimic the real dataset.

### 3.6.1 Data format and settings

In the mentor-mentee matching, the data includes the background and preference of each matched mentor-mentee pair, including their major, degree, location, industry, etc, together with their feedback. The feedback of each pair is then summarized as two outcomes: satisfied/unsatisfied. With preprocessing of the raw data and variable selection in linear regression, we select three variables (Major, Location and Industry) that are important to the experience of participants. Table 3.3 specifies these their categories.

Table 3.3: Feature variables for mentees

Var Name	Categories
Major	1: Science; 2: Engineering; 3: Business; 4: Other.
Location	1: Local; 2: East Coast of US; 3: West Coast of US; 4: Other in US; 5: International.
Industry	1: Academia; 2: Information Technology; 3: Transportation; 4: Finance; 5: Healthcare; 6: Other.

Since the variables are all categorical, for any mentor-mentee pair, we introduce the indicator variables  $I_{\text{major}}$ ,  $I_{\text{location}}$  and  $I_{\text{industry}}$  to indicate whether the pair is matched for each variable. For example,  $I_{\text{major}} = 1$  if the mentee and mentor have the same major, and  $I_{\text{major}} = 0$  otherwise. Besides, let  $Y \in \{0, 1\}$  be the binary variable of outcome, such that  $Y = 1$  indicates ‘satisfied’, and  $Y = 0$  indicates ‘unsatisfied’. With the above variables, we fit the following model for the rating using the data:

$$\log \frac{P(Y = 1)}{P(Y = 0)} = -0.325 + 0.189I_{\text{major}} + 0.219I_{\text{location}} + 0.173I_{\text{industry}}. \quad (3.11)$$

Alternatively, we can also represent the categorical variables with one-hot vectors for

mentee and mentor, concatenated with an additional scalar 1 for the intercept. Let  $\mathbf{x}$  and  $\mathbf{z}$  denote such vector of covariates for mentee and mentor, respectively. Then (3.11) is equivalent to

$$\log \frac{P(Y = 1)}{P(Y = 0)} = (\mathbf{x} \circ \mathbf{z})^\top \beta, \quad \text{where } \beta = (-0.325, 0.189\mathbb{1}_4, 0.219\mathbb{1}_5, 0.173\mathbb{1}_6). \quad (3.12)$$

Here  $\mathbb{1}_k$  denotes the vector of  $k$  dimensions with all the entries being 1. In the following study, after the assignment is decided for each round, we randomly generate the outcomes based on (3.11). We then use logistic regression to fit the binary outcomes with the covariates previously described.

Now we specify the generation of the covariates for mentors and mentees at each round. We randomly generate the data for  $T = 20$  rounds. At each round, we generate the covariates of  $n = 200$  mentors and mentees. For convenience, we draw from each category with equal probability for every variable, independent from other variables.

### 3.6.2 Performance

We run our proposed algorithm with various choices of  $\lambda$  over 50 repetitions and record the cumulative regret. Figure 3.3 shows the cumulative regret with different choice of  $\lambda$  averaged over 50 repetitions, with the standard deviation indicated by the shadow. The regret of the random match approach is also presented in the figure for comparison. Also, Table 3.4 presents the cumulative regrets and their standard deviations up to  $T = 20$  under different choice of  $\lambda$ . As can be seen, in this study we observe a larger gap between our method with a properly chosen  $\lambda$  and the pure-greedy algorithm when  $\lambda = 0$ . Intuitively, this is related to the condition number of the design matrix. Because the variables are categorical, there are many zero entries in  $\phi(\mathbf{x}, \mathbf{z})$ , which increases the difficulty to capture the underlying  $\theta^*$ . In fact, it can be verified that with random pairing, the minimum eigenvalue of the covariance matrix  $\mathbb{E}_{\mathbf{x} \sim P_{\mathcal{X}}, \mathbf{z} \sim P_{\mathcal{Z}}}[(\mathbf{x} \circ \mathbf{z})(\mathbf{x} \circ \mathbf{z})^\top]$  is much smaller in this example, compared to

the previous simulations. In this case, a larger  $\lambda$  value (such as 20) can better encourage the exploration in the early stages for a more accurate estimation and thus balance the trade-off between exploration and exploitation, resulting in the significant improvement compared to the pure-greedy algorithm. This observation also suggests that the optimal choice of  $\lambda$  and  $t_1$  can depend on the problem and data in practice, and one might want fine-tuning for more desired performance.

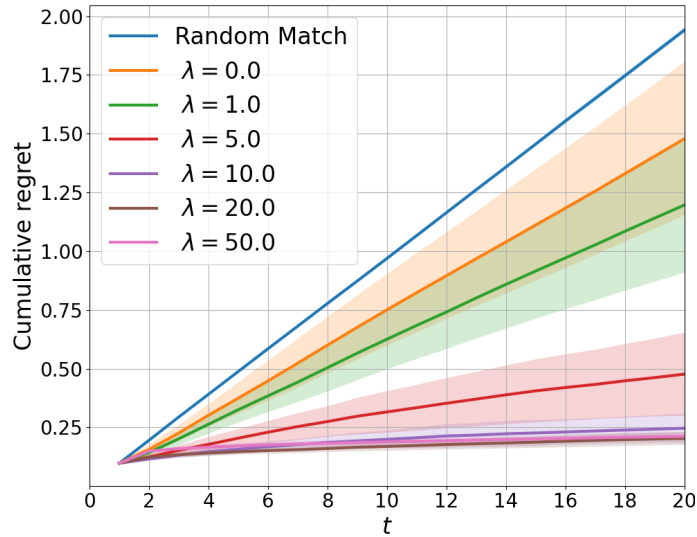


Figure 3.3: Cumulative regret with different  $\lambda$  for the mentor-mentee matching scenario. With a proper  $\lambda$  value (e.g. 20), our method has significant improvement in the total utility compared to the greedy method.

Table 3.4: Total regret up to  $T = 20$  for the simulated data for mentor-mentee matching scenario averaged over 50 repetitions. The standard deviations are presented in brackets.

Random Match	1.94 (0.01)
$\lambda = 0$ (Greedy)	1.48 (0.33)
$\lambda = 1$	1.20 (0.29)
$\lambda = 5$	0.48 (0.18)
$\lambda = 10$	0.25 (0.06)
$\lambda = 20$	0.20 (0.03)
$\lambda = 50$	0.21 (0.02)

### 3.7 Proofs

This section contains three subsections. Subsections 3.7.1 and 3.7.2 present the proof for Theorems 3.4.1 and 3.4.2, respectively. Subsection 3.7.3 presents the proof for Lemma 3.7.4, which is used in the proof for Theorem 3.4.2.

#### 3.7.1 Proof of Theorem 7

**Proof** We first define that

$$M_{t-1} = \sum_{\tau=1}^{t-1} \sum_{i \in \mathcal{I}_\tau} \phi_{i, \delta_\tau(i)}^\tau \phi_{i, \delta_\tau(i)}^{\tau \top},$$

$$\bar{M}_{t-1} = M_{t-1} + rI.$$

Our first step is to show that with high probability,  $\|\hat{\theta}^t - \theta^*\|_2 \leq 1$  for all  $t \geq t_1 + 1$  under the given condition.

Consider some  $\eta$ -neighborhood of  $\theta^*$ ,  $\mathcal{B}_\eta(\theta^*) := \{\theta : \|\theta - \theta^*\|_2 \leq \eta\}$ . Define

$$\kappa_\eta := \inf_{\|\phi\|_2 \leq 1, \theta \in \mathcal{B}_\eta(\theta^*)} \frac{\exp(\phi^\top \theta)}{1 + \exp(\phi^\top \theta)}.$$

Note that at time  $t$ , the estimator  $\hat{\theta}^t$  from penalized maximum-likelihood is the solution to the following equation:

$$\sum_{\tau=1}^{t-1} \sum_{i \in \mathcal{I}_\tau} \left( U_{i, \delta_\tau(i)}^t - \frac{\exp(\phi_{i, \delta_\tau(i)}^{\tau \top} \theta)}{1 + \exp(\phi_{i, \delta_\tau(i)}^{\tau \top} \theta)} \right) \phi_{i, \delta_\tau(i)}^\tau - r\theta = 0.$$

Define

$$G(\theta) := \sum_{\tau=1}^{t-1} \sum_{i \in \mathcal{I}_\tau} \left( \frac{\exp(\phi_{i, \delta_\tau(i)}^{\tau \top} \theta)}{1 + \exp(\phi_{i, \delta_\tau(i)}^{\tau \top} \theta)} - \frac{\exp(\phi_{i, \delta_\tau(i)}^{\tau \top} \theta^*)}{1 + \exp(\phi_{i, \delta_\tau(i)}^{\tau \top} \theta^*)} \right) \phi_{i, \delta_\tau(i)}^\tau + r\theta.$$

Then we have  $G(\theta^*) = r\theta^*$  and  $G(\hat{\theta}^t) = \sum_{\tau=1}^{t-1} \sum_{i \in [n_\tau]} \epsilon_{i, \delta_\tau(i)}^\tau \phi_{i, \delta_\tau(i)}^\tau$ . We next show that

$G(\cdot)$  is an injection from  $\mathbb{R}^d$  to  $\mathbb{R}^d$ . Note that for any  $\theta_1 \neq \theta_2$ , by the mean value theorem, we have

$$G(\theta_1) - G(\theta_2) = \left[ \sum_{\tau=1}^{t-1} \sum_{i \in [n_\tau]} \frac{\exp(\phi_{i, \delta_\tau(i)}^{\tau \top} \bar{\theta})}{\left[1 + \exp(\phi_{i, \delta_\tau(i)}^{\tau \top} \bar{\theta})\right]^2} \phi_{i, \delta_\tau(i)}^{\tau} \phi_{i, \delta_\tau(i)}^{\tau \top} + rI \right] (\theta_1 - \theta_2),$$

where  $\bar{\theta} = v\theta_1 + (1-v)\theta_2$  for some  $v \in [0, 1]$ . It follows that

$$(\theta_1 - \theta_2)^\top (G(\theta_1) - G(\theta_2)) \geq r \|\theta_1 - \theta_2\|_2^2 > 0.$$

Hence,  $G(\cdot)$  is an injection from  $\mathbb{R}^d$  to  $\mathbb{R}^d$ , and  $G^{-1}$  is well-defined.

Before we continue, we introduce two useful lemmas that help bound  $\|\hat{\theta}^t - \theta^*\|_2$ .

**Lemma 3.7.1 (Lemma A in [69])** *Let  $G(\cdot)$  be a smooth injection from  $\mathbb{R}^d$  to  $\mathbb{R}^d$ , with  $G(x_0) = y_0$ . Define  $\mathcal{B}_\eta(x_0) = \{x : \|x - x_0\| \leq \eta\}$ . Then  $\inf_{x \in \partial \mathcal{B}_\eta(x_0)} \|G(x) - y_0\| \geq s$  implies that*

$$\mathcal{B}_r(y_0) = \{y : \|y - y_0\| \leq s\} \subseteq G(\mathcal{B}_\eta(x_0)).$$

**Lemma 3.7.2 (Lemma 7 in [68])** *For any  $\delta > 0$ , the event*

$$\left\{ \|G(\hat{\theta}^t)\|_{M_{t-1}^{-1}} \leq 4\sqrt{d + \log(1/\delta)} \right\}$$

*holds with probability at least  $1 - \delta$ .*

For any  $\theta \in \mathcal{B}_\eta(\theta^*)$ , since  $0 < \kappa_\eta < 1$ , we have

$$\begin{aligned}
\|G(\theta) - G(\theta^*)\|_{M_{t-1}^{-1}}^2 &\geq (\theta - \theta^*)^\top (\kappa_\eta M_{t-1} + rI)(M_{t-1})^{-1} (\kappa_\eta M_{t-1} + rI)(\theta - \theta^*) \\
&\geq (\theta - \theta^*)^\top (\kappa_\eta M_{t-1} + rI)(M_{t-1} + \frac{r}{\kappa_\eta} I)^{-1} (\kappa_\eta M_{t-1} + rI)(\theta - \theta^*) \\
&\geq \kappa_\eta^2 (\theta - \theta^*)^\top (M_{t-1} + \frac{r}{\kappa_\eta} I)(\theta - \theta^*) \\
&\geq \kappa_\eta^2 \lambda_{\min}(M_{t-1}) \|\theta - \theta^*\|_2^2.
\end{aligned}$$

By Lemma 3.7.1, we have that

$$\left\{ \theta : \|G(\theta) - G(\theta^*)\|_{M_{t-1}^{-1}} \leq \kappa_\eta \eta \sqrt{\lambda_{\min}(M_{t-1})} \right\} \subseteq \mathcal{B}_\eta(\theta^*). \quad (3.13)$$

By Lemma 3.7.2, and using a union bound argument, we have that with probability at least  $1 - \delta$ ,

$$\|G(\hat{\theta}^t)\|_{M_{t-1}^{-1}} \leq 4\sqrt{d + \log(T/\delta)}, \text{ for } t \in [T]. \quad (3.14)$$

Also, note that

$$\|G(\theta^*)\|_{M_{t-1}^{-1}}^2 = r^2 \theta^{*\top} M_{t-1}^{-1} \theta^* \leq \frac{r^2}{\lambda_{\min}(M_{t-1})} \|\theta^*\|_2^2 \leq \frac{r^2}{\lambda_{\min}(M_{t-1})}. \quad (3.15)$$

Combining (3.13), (3.14) and (3.15), we have that when

$$\lambda_{\min}(M_{t_1}) \geq \frac{16(d + \log(T/\delta))}{\kappa_\eta^2 \eta^2} + \frac{8\sqrt{rd + r \log(T/d)}}{(\kappa_\eta \eta)^{3/2}} + \frac{r}{\kappa_\eta \eta},$$

with probability at least  $1 - \delta$ ,  $\hat{\theta}^t \in \mathcal{B}_\eta(\theta^*)$  for all  $t \geq t_1 + 1$ . Taking  $\eta = 1$ , the above condition is met in (3.7), so that with probability at least  $1 - \delta$ , we have  $\|\hat{\theta}^t - \theta^*\|_2 \leq 1$  for all  $t \geq t_1 + 1$ , which we assume to hold in the followings.

We now introduce the following lemma to help uniformly bound the terms  $\|\hat{\theta}^t - \theta^*\|_{M_{t-1}}^2$

for all  $t \geq t_1 + 1$ .

**Lemma 3.7.3 (Theorem 1 in [70])** *Let  $\{F_t\}_{t=0}^\infty$  be a filtration. Let  $\{\epsilon_t\}_{t=1}^\infty$  be a real-valued stochastic process such that  $\epsilon_t$  is  $F_{t-1}$ -measurable, and  $\epsilon_t$  is conditionally  $R$ -sub-Gaussian, i.e.,*

$$\mathbb{E}[\exp(\lambda\epsilon_t)|F_{t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \quad \forall \lambda \in \mathbb{R}.$$

*Let  $\{X_t\}_{t=1}^\infty$  be an  $\mathbb{R}^d$ -valued stochastic process, such that  $X_t$  is  $F_{t-1}$ -measurable. Assume  $M$  is a  $d \times d$  positive definite matrix. Define*

$$\bar{M}_t = M + \sum_{s=1}^t X_s X_s^\top, \quad S_t = \sum_{s=1}^t \epsilon_s X_s.$$

*Then for any  $\delta > 0$ , with probability at least  $1 - \delta$ ,*

$$\|S_t\|_{\bar{M}_t^{-1}}^2 \leq 2R^2 \log\left(\frac{\det(\bar{M}_t)^{1/2} \det(M)^{-1/2}}{\delta}\right), \quad \text{for all } t \geq 0.$$

Let  $N_t = \sum_{\tau=1}^t n_\tau$ . By Lemma 3.7.3, with probability at least  $1 - \delta$ , we have that

$$\begin{aligned} \|G(\hat{\theta}^t)\|_{\bar{M}_{t-1}^{-1}}^2 &\leq 2 \log\left(\frac{\det(\bar{M}_{t-1})^{1/2} \det(rI)^{-1/2}}{\delta}\right) \\ &\leq 2 \log\left(\frac{r^{-d/2} (r + \frac{N_{t-1}}{d})^{d/2}}{\delta}\right) \\ &\leq d \log\left(1 + \frac{N_{t-1}}{rd}\right) + 2 \log(1/\delta). \end{aligned}$$

Meanwhile, we have  $\|G(\theta^*)\|_{\bar{M}_{t-1}^{-1}}^2 \leq r$  for all  $t$ . Note that when  $\|\theta - \theta^*\|_2 \leq 1$ ,

$$\begin{aligned}
\|G(\theta) - G(\theta^*)\|_{\bar{M}_{t-1}^{-1}}^2 &\geq (\theta - \theta^*)^\top (\kappa M_{t-1} + rI)(M_{t-1})^{-1} (\kappa M_{t-1} + rI)(\theta - \theta^*) \\
&\geq (\theta - \theta^*)^\top (\kappa M_{t-1} + rI) \left(M_{t-1} + \frac{r}{\kappa} I\right)^{-1} (\kappa_\eta M_{t-1} + rI)(\theta - \theta^*) \\
&\geq \kappa^2 (\theta - \theta^*)^\top \left(M_{t-1} + \frac{r}{\kappa} I\right)(\theta - \theta^*) \\
&\geq \kappa^2 \|\theta - \theta^*\|_{\bar{M}_{t-1}}^2.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\|\hat{\theta}^t - \theta^*\|_{\bar{M}_{t-1}}^2 &\leq \frac{1}{\kappa^2} \|G(\theta) - G(\theta^*)\|_{\bar{M}_{t-1}^{-1}}^2 \\
&\leq \frac{1}{\kappa^2} \left[ d \log \left( 1 + \frac{N_{t-1}}{rd} \right) + 2 \log(1/\delta) + r \right] \\
&\leq \frac{1}{\kappa^2} \left[ d \log \left( 1 + \frac{N_T}{rd} \right) + 2 \log(1/\delta) + r \right]
\end{aligned}$$

for all  $t \geq t_1$ . This completes our proof for Theorem 3.4.1.

### 3.7.2 Proof of Theorem 8

In this section, we provide the proof for our Theorem 3.4.2. Recall that we define  $\phi_{i,j}^t := \mathbf{x}_i^t \circ \mathbf{z}_j^t$ . We first introduce a useful lemma.

**Lemma 3.7.4** *Let  $\{\phi_i^t : t \in [T], i \in [n_t]\}$  be an arbitrary collection of  $d$ -dimensional vectors satisfying  $\|\phi_i^t\|_2 \leq 1$  for every  $t$  and  $i$ . Suppose  $M_0 \succeq I_d$ ,  $M_t = M_{t-1} + \sum_{i=1}^{n_t} \phi_i^t \phi_i^{t\top}$ . Denote  $\|\phi\|_M = \sqrt{\phi^\top M \phi}$  the norm induced by a positive definite matrix  $M$ . Then we have*

$$\sum_{t=t_1+1}^T \sum_{i=1}^{n_t} \frac{\|\phi_i^t\|_{M_{t-1}}^2}{n_t} \leq 2d \log \left( 1 + \frac{\sum_{t=1}^T n_t}{d} \right).$$

Lemma 3.7.4 will be used to upper bound the total uncertainty for our algorithm, which is closely related to the regret bound. The proof of The proof for Lemma 3.7.4 adapts the

argument for UCB method in linear contextual bandits, see [71] and [70] for example. We defer the detailed proof for Lemma 3.7.4 in Section 3.7.3 of Appendix.

Below we provide the proof to Theorem 3.4.2. In the high level, we first make use of Theorem 3.4.1 to bound the total regret in terms of the summation of  $\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}$ 's. Then we use Lemma 3.7.4 to derive an upper bound for  $\sum_t \sum_i \|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}$ , which leads to an upper bound for the total regret.

**Proof** For notational simplicity, we first define

$$V_t = \frac{1}{n_t} \sum_{i=1}^{n_t} u_{i,\delta_t(i)}^t, \quad V_t^* = \frac{1}{n_t} \sum_{i=1}^{n_t} u_{i,\delta_t^*(i)}^t.$$

where  $\delta_t$  and  $\delta_t^*$  denotes the assignment decided by the algorithm and the oracle assignment at  $t$ , respectively. Then we have

$$R_T \leq t_1 + \sum_{t=t_1+1}^T (V_t^* - V_t).$$

We first upper bound the performance gap  $(V_t^* - V_t)$  for each single round  $t \geq t_1 + 1$ , and then upper bound their summation as the total regret.

First, under Assumption 3.4.2, by Proposition 1 in [68], when  $t_1$  satisfies (3.8), with probability at least  $1 - \delta$ , we have

$$\lambda_{\min}(M_{t_1}) \geq \frac{32(d + \log(T/\delta))}{\kappa^2} + \frac{2r}{\kappa},$$

so that with probability at least  $1 - 2\delta$ , we have

$$\|\hat{\theta}^t - \theta^*\|_{\bar{M}_{t-1}}^2 \leq \frac{1}{\kappa^2} \left[ d \log \left( 1 + \frac{\sum_{t=1}^T n_t}{rd} \right) + 2 \log(1/\delta) + r \right] \quad (3.16)$$

for all  $t \geq t_1 + 1$ , according to Theorem 3.4.1.

Now we upper bound the difference between  $\hat{u}_{i,j}^t - u_{i,j}^t$  for  $t \geq t_1 + 1$  and  $i, j \in [n_t]$ .

Note that

$$\widehat{u}_{i,j}^t = \frac{1}{1 + \exp(-\phi_{i,j}^{t\top} \widehat{\theta}^t)}, \quad u_{i,j}^t = \frac{1}{1 + \exp(-\phi_{i,j}^{t\top} \theta^*)}.$$

Since the function  $\frac{1}{1 + \exp(-\phi_{i,j}^{t\top} \theta)}$  is convex in  $\theta$ , we have

$$\begin{aligned} |\widehat{u}_{i,j}^t - u_{i,j}^t| &\leq \frac{\exp(-\phi_{i,j}^{t\top} \widehat{\theta}^t)}{[1 + \exp(-\phi_{i,j}^{t\top} \widehat{\theta}^t)]^2} \left| \phi_{i,j}^{t\top} (\widehat{\theta}^t - \theta^*) \right| \\ &\leq \frac{1}{2} \|\phi_{i,j}^t\|_{\bar{M}_{t-1}^{-1}} \|\widehat{\theta}^t - \theta^*\|_{\bar{M}_{t-1}} \\ &\leq \frac{\lambda}{2} \|\phi_{i,j}^t\|_{\bar{M}_{t-1}^{-1}}. \end{aligned} \tag{3.17}$$

Here in the last inequality, we use (3.9) and (3.16). Then for every  $t \geq t_1 + 1$ , we have that

$$\begin{aligned} V_t^* - V_t &= \frac{1}{n_t} \sum_{i \in [n_t]} u_{i, \delta_t^*(i)}^t - u_{i, \delta_t(i)}^t \\ &\leq \frac{1}{n_t} \sum_{i \in [n_t]} \left[ \left( \widehat{u}_{i, \delta_t^*(i)}^t + \frac{\lambda}{2} \|\phi_{i, \delta_t^*(i)}^t\|_{\bar{M}_{t-1}^{-1}} \right) - u_{i, \delta_t(i)}^t \right] \\ &\leq \frac{1}{n_t} \sum_{i \in [n_t]} \left[ \left( \widehat{u}_{i, \delta_t(i)}^t + \frac{\lambda}{2} \|\phi_{i, \delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}} \right) - u_{i, \delta_t(i)}^t \right] \\ &\leq \frac{\lambda}{n_t} \sum_{i \in [n_t]} \|\phi_{i, \delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}. \end{aligned} \tag{3.18}$$

Here the second inequality is by the construction of our assignment  $\delta_t$  that maximizes the total upper confidence bound, and the first and third inequalities use (3.17). From (3.18) we have that

$$R_T \leq t_1 + \lambda \sum_{t=t_1+1}^T \sum_{i \in [n_t]} \frac{\|\phi_{i, \delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}}{n_t}.$$

To further bound the right-hand side, we use Lemma 3.7.4 to obtain that

$$\begin{aligned}
\sum_{t=t_1}^T \sum_{i=1}^{n_t} \frac{\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}}{n_t} &\leq \sqrt{T \sum_{t=1}^T \left( \sum_{i=1}^{n_t} \frac{\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}}{n_t} \right)^2} \\
&\leq \sqrt{T \sum_{t=1}^T \sum_{i=1}^{n_t} \frac{\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}^{-1}}^2}{n_t}} \\
&\leq \sqrt{2dT \log \left( 1 + \frac{\sum_{t=1}^T n_t}{d} \right)}. \tag{3.19}
\end{aligned}$$

Combining (3.19) with our choice of  $\lambda$  in (3.9), we obtain that

$$\begin{aligned}
R_T &\leq t_1 + \lambda \sum_{t=t_1+1}^T \sum_{i=1}^{n_t} \frac{s_{i,\delta_t(i)}^t}{n_t} \\
&\leq t_1 + \frac{2d}{\kappa} \sqrt{T \log \left( 1 + \frac{\sum_{t=1}^T n_t}{d} \right)} \sqrt{\log \left( 1 + \frac{\sum_{t=1}^T n_t}{rd} \right) + 2 \log(1/\delta) + r},
\end{aligned}$$

which completes our proof for Theorem 3.4.2. ■

Note that for any  $t, i, j$ , we have

$$\left| \phi_{i,j}^{t\top} (\hat{\theta}^t - \theta^*) \right| \leq \|\phi_{i,j}^t\|_{V_{t-1}^{-1}} \|\hat{\theta}^t - \theta^*\|_{V_{t-1}}.$$

We now introduce the following result as in [68] to bound the term  $\|\hat{\theta}^t - \theta^*\|_{V_{t-1}^{-1}}$ .

**Lemma 3.7.5 (Lemma 3 in [68])** *Suppose  $\lambda_{\min}(V_0) \geq 1$ . Then for any  $\delta \in [1/T, 1)$ , define the event*

$$\mathcal{E} := \left\{ \|\hat{\theta}^t - \theta^*\|_{V_{t-1}} \leq \frac{1}{\kappa} \sqrt{\frac{d}{2} \log(1 + 2t/d) + \log(1/\delta)}, \text{ for } t \geq 1 \right\}.$$

Then  $\mathcal{E}$  holds with probability at least  $1 - \delta$ .

Using the result above and our choice of  $\lambda$  in (3.9), we conclude that  $\mathcal{E}_{i,j}^t$  holds simultaneously for every  $t, i, j$  with probability at least  $1 - \delta$ . ■

### 3.7.3 Proof of Lemma 3.7.4

**Proof** We first introduce the following lemma that will be used to bound  $\|\phi_i^t\|_{M_{t-1}^{-1}}^2$ .

**Lemma 3.7.6** *Let  $M$  be a  $d \times d$  positive definite matrix with the minimum eigenvalue  $\lambda_{\min}(M) \geq 1$ . Let  $\phi$  be a  $d$ -dimensional vector with  $\|\phi\|_2 \leq 1$ . Then we have*

$$\|\phi\|_{M^{-1}}^2 \leq 2 \log \frac{\det(M + \phi\phi^\top)}{\det(M)}.$$

To prove Lemma 3.7.6, we note that

$$\begin{aligned} \det(M + \phi\phi^\top) &= \det(M) \det(I_d + M^{-1/2}\phi(M^{-1/2}\phi)^\top) \\ &= \det(M) (1 + \|\phi\|_{M^{-1}}^2) \end{aligned}$$

Since  $\|\phi\|_2 \leq 1$  and  $\lambda_{\min}(M) \geq 1$ , we have  $\|\phi\|_{M^{-1}}^2 \leq 1$ . Using the fact that  $x \leq 2 \log(1 + x)$  for  $x \in [0, 1]$ , we get

$$\|\phi\|_{M^{-1}}^2 \leq 2 \log(1 + \|\phi\|_{M^{-1}}^2) = 2 \log \frac{\det(M + \phi\phi^\top)}{\det(M)}.$$

The proof for Lemma 3.7.6 is thus completed.

We now continue to prove Lemma 3.7.4. Note that we assume  $\|\phi_{i,j}^t\|_2 \leq 1$  for all  $t, i, j$ . Also, by our construction of  $\bar{M}_t$ 's, we have  $\lambda_{\min}(\bar{M}_t) \geq r$  for all  $t \geq t_1 + 1$ . By defining

$\bar{M}_{t-1,i} = \bar{M}_{t-1} + \phi_{i,\delta_t(i)}^t \phi_{i,\delta_t(i)}^{t\top}$  and using Lemma 3.7.6, we have

$$\begin{aligned}
\sum_{t=t_1+1}^T \sum_{i=1}^{n_t} \frac{\|\phi_{i,\delta_t(i)}^t\|_{\bar{M}_{t-1}}^2}{n_t} &\leq \sum_{t=t_1+1}^T \sum_{i=1}^{n_t} 2 \log \left[ \frac{\det(\bar{M}_{t-1,i})}{\det(\bar{M}_{t-1})} \right]^{\frac{1}{n_t}} \\
&= 2 \log \prod_{t=t_1+1}^T \frac{[\prod_{i=1}^{n_t} \det(\bar{M}_{t-1,i})]^{\frac{1}{n_t}}}{\det(\bar{M}_{t-1})} \\
&\leq 2 \log \frac{[\prod_{i=1}^{n_T} \det(\bar{M}_{T-1,i})]^{\frac{1}{n_T}}}{\det(\bar{M}_{t_1})}. \tag{3.20}
\end{aligned}$$

Here the last inequality uses the fact that  $\det(M_{t-1,i}) \leq \det(M_t)$  for every  $t$  and  $i$ , by our construction. Since  $M_{t_1} \succeq I_d$  and  $\|\phi_{i,j}^t\|_2 \leq 1$ , the maximum value  $\det(M_{T-1,i})$  can take is  $\left(1 + \frac{(\sum_{t=1}^{T-1} n_t) + 1}{d}\right)^d$ , which can be further upper bounded by  $\left(1 + \frac{\sum_{t=1}^T n_t}{d}\right)^d$ . Therefore, following (3.20), we have

$$\sum_{t=t_1}^T \sum_{i=1}^{n_t} \frac{\|\phi_{i,\delta_t(i)}^t\|_{M_{t-1}}^2}{n_t} \leq 2d \log \left(1 + \frac{\sum_{t=1}^T n_t}{d}\right).$$

This completes our proof for Lemma 3.7.4. ■

## CHAPTER 4

### CONCLUSION AND FUTURE RESEARCH

In this chapter, we summarize our contributions and also highlight several potential research topics for future investigation.

#### 4.1 Summary of Contribution

This dissertation investigates the following three specific areas in statistical machine learning: pivotal method for high-dimensional LDA, robust method for high-dimensional LDA under data contamination, and efficient algorithm for sequential assignment problems with unknown utility. Here we remark that the three areas we investigate have numerous real-world applications. Specifically, the LDA or linear classification is widely applied in various situations such as image classification [72, 41], fraud detection [73, 74], spam filtering [75, 76, 77], medical diagnosis [78, 79], biomarker discovery [80, 81, 82], among many others. The approach of robust LDA helps improve the robustness and efficiency in the presence of data contamination or outliers. In addition, algorithms for sequential assignment problems with unknown utility can also be applied in scenarios such as online resource allocation [83], online dating recommendation [59, 60], etc., aside from the mentor-mentee matching. Below is the summary of our contributions.

- In Chapter 1, we introduce a novel tuning-insensitive method named PANDA for high-dimensional LDA. Different from existing methods, our method conducts parameter estimation under a pivotal estimation framework and only needs to solve a single convex optimization problem by simultaneously estimate  $\beta^*$  and  $\Delta$ . Compared with existing methods, we observe that our method yields equal or better performance, and requires less effort in parameter tuning.

- In Chapter 2, we robustify the PANDA method under the new scenario of data contamination, where a fraction of the observed data might be corrupted by an adversary. By combining with a robust mean estimation approach and filtering the outliers, we enhance the robustness of the PANDA method.
- In Chapter 3, motivated by the mentor-mentee matching practice in college, we present a simple but useful algorithm combining the UCB and Hungarian method for a new scenario of bandit problem, which involves a sequence of assignment problems with unknown utility and stochastic feedback.

## 4.2 Future Research

Below we highlight some interesting topics that worth further investigation.

- While we have derived the mis-classification error properties of the proposed PANDA method, it remains an open problem to investigate its theoretical properties on the variable selection. Moreover, it is also interesting to extend the similar tuning-insensitive framework to quadratic discriminant analysis (QDA).
- In Chapter 2, our theoretical analysis for our method for robust high-dimensional LDA involves a weaker contamination model on the outliers, i.e., the adversary can replace the inliers with arbitrary data points lying in a bounded ball of radius  $R < \infty$  around the inliers. It will be interesting to develop computationally efficient robust LDA algorithms for the stronger contamination model, where  $R$  is allowed to go to  $\infty$ . Also, while we can derive a loose lower bound for the breakdown point of our method, further analysis of the exact breakdown point is of interest.
- In real-world applications of the assignment problems, the unknown underlying utility function might have complicated form, and thus we might need to adopt a more sophisticated model such as deep neural networks or non-parametric models. In addition, it is also interesting to investigate when the utility function is non-stationary,

e.g., changing over time, by adapting the time-varying bandit algorithms in [84] and [85] to our context. Moreover, it is important to develop distributed algorithm for learning the assignment strategy, especially when we face the problem of large-scale assignments.

## REFERENCES

- [1] T. W. Anderson, *An Introduction To Multivariate Statistical Analysis*, 3rd ed. New York: Wiley Interscience, 2003.
- [2] O. Yamashita, M.-a. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse Estimation Automatically Selects Voxels Relevant for the Decoding of fMRI Activity Patterns," *NeuroImage*, vol. 42, no. 4, pp. 1414–1429, 2008.
- [3] Y. Shi, D. Dai, C. Liu, and H. Yan, "Sparse Discriminant Analysis for Breast Cancer Biomarker Identification and Classification," *Progress in Natural Science*, vol. 19, no. 11, pp. 1635–1641, 2009.
- [4] P. J. Bickel, E. Levina, *et al.*, "Some Theory for Fisher's Linear Discriminant Function, 'Naive Bayes', and some Alternatives when there are many more Variables than Observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.
- [5] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of Multiple Cancer Types by Shrunk Centroids of Gene Expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6567–6572, 2002.
- [6] J. Fan and Y. Fan, "High-Dimensional Classification Using Features Annealed Independence Rules," *The Annals of Statistics*, vol. 36, no. 6, p. 2605, 2008.
- [7] J. Shao, Y. Wang, X. Deng, and S. Wang, "Sparse Linear Discriminant Analysis by Thresholding for High Dimensional Data," *The Annals of Statistics*, vol. 39, no. 2, pp. 1241–1265, 2011.
- [8] D. M. Witten and R. Tibshirani, "Penalized Classification using Fisher's Linear Discriminant," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 5, pp. 753–772, 2011.
- [9] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, "Sparse Discriminant Analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.
- [10] T. Cai and W. Liu, "A Direct Estimation Approach to Sparse Linear Discriminant Analysis," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1566–1577, 2011.
- [11] Q. Mai, H. Zou, and M. Yuan, "A Direct Approach to Sparse Discriminant Analysis in Ultra-high Dimensions," *Biometrika*, vol. 99, no. 1, pp. 29–42, 2012.

- [12] J. Fan, Y. Feng, and X. Tong, “A Road to Classification in High Dimensional Space: the Regularized Optimal Affine Discriminant,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 4, pp. 745–771, 2012.
- [13] T. Cai and L. Zhang, “High Dimensional Linear Discriminant Analysis: Optimality, Adaptive Algorithm and Missing Data,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 81, no. 4, pp. 675–705, 2019.
- [14] E. Gautier, A. Tsybakov, and C. Rose, “High-Dimensional Instrumental Variables Regression and Confidence Sets,” *arXiv preprint arXiv:1105.2454*, 2011.
- [15] E. Candes and T. Tao, “The Dantzig Selector: Statistical Estimation When  $p$  is Much Larger than  $n$ ,” *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.
- [16] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994, vol. 13.
- [17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
- [18] T. R. Golub *et al.*, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” *Science*, vol. 286, no. 5439, pp. 531–537, 1999.
- [19] G. Raskutti, M. J. Wainwright, and B. Yu, “Restricted Eigenvalue Properties for Correlated Gaussian Designs,” *Journal of Machine Learning Research*, vol. 11, no. Aug, pp. 2241–2259, 2010.
- [20] A. Belloni, V. Chernozhukov, and L. Wang, “Square-root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming,” *Biometrika*, vol. 98, no. 4, pp. 791–806, 2011.
- [21] T. Sun and C.-H. Zhang, “Scaled Sparse Linear Regression,” *Biometrika*, vol. 99, no. 4, pp. 879–898, 2012.
- [22] A. Belloni, V. Chernozhukov, and L. Wang, “Pivotal Estimation via Square-Root Lasso in Nonparametric Regression,” *The Annals of Statistics*, vol. 42, no. 2, pp. 757–788, 2014.
- [23] H. Liu, L. Wang, and T. Zhao, “Calibrated Multivariate Regression with Application to Neural Semantic Basis Discovery,” *Journal of Machine Learning Research: JMLR*, vol. 16, p. 1579, 2015.

- [24] F. Bunea, J. Lederer, and Y. She, “The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms,” *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1313–1325, 2013.
- [25] T. Sun and C.-H. Zhang, “Sparse Matrix Inversion with Scaled Lasso,” *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3385–3418, 2013.
- [26] H. Liu and L. Wang, “TIGER: A Tuning-Insensitive Approach for Optimally Estimating Gaussian Graphical Models,” *Electronic Journal of Statistics*, vol. 11, no. 1, pp. 241–294, 2017.
- [27] T. Zhao and H. Liu, “Sparse Precision Matrix Estimation with Calibration,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, 2013, pp. 2274–2282.
- [28] A. Belloni and V. Chernozhukov, “ $\ell_1$ -Penalized Quantile Regression in High-Dimensional Sparse Models,” *The Annals of Statistics*, vol. 39, no. 1, pp. 82–130, 2011.
- [29] L. Wang, “The  $L_1$  Penalized LAD Estimator for High Dimensional Linear Regression,” *Journal of Multivariate Analysis*, vol. 120, pp. 135–151, 2013.
- [30] J. W. Tukey, “Mathematics and the picturing of data,” in *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, vol. 2, 1975, pp. 523–531.
- [31] I. Diakonikolas, G. Kamath, D. M. Kane, J. Z. Li, A. Moitra, and A. Stewart, “Robust estimators in high dimensions without the computational intractability,” *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 655–664, 2016.
- [32] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, “Being robust (in high dimensions) can be practical,” in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML’17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 999–1008.
- [33] J. Steinhardt, “Robust learning: Information theory and algorithms,” AAI28115249, Ph.D. dissertation, Stanford, CA, USA, 2018, ISBN: 9798662559240.
- [34] Y. Cheng, I. Diakonikolas, and R. Ge, “High-dimensional robust mean estimation in nearly-linear time,” in *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA ’19, San Diego, California: Society for Industrial and Applied Mathematics, 2019, pp. 2755–2771.
- [35] Y. Cheng, I. Diakonikolas, R. Ge, and M. Soltanolkotabi, “High-dimensional robust mean estimation via gradient descent,” in *Proceedings of the 37th International*

*Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 13–18 Jul 2020, pp. 1768–1778.

- [36] B. Zhu, J. Jiao, and J. Steinhardt, “Robust estimation via generalized quasi-gradients,” *Information and Inference: A Journal of the IMA*, vol. 11, no. 2, pp. 581–636, Aug. 2021. eprint: <https://academic.oup.com/imaiai/article-pdf/11/2/581/44020638/iaab018.pdf>.
- [37] A. Deshmukh, J. Liu, and V. V. Veeravalli, “Robust mean estimation in high dimensions: An outlier fraction agnostic and efficient algorithm,” in *2022 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2022, pp. 1115–1120.
- [38] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart, “Robust estimators in high-dimensions without the computational intractability,” *SIAM Journal on Computing*, vol. 48, no. 2, pp. 742–864, 2019. eprint: <https://doi.org/10.1137/17M1126680>.
- [39] M. Avella-Medina, H. S. Battey, J. Fan, and Q. Li, “Robust estimation of high-dimensional covariance and precision matrices,” *Biometrika*, vol. 105, no. 2, pp. 271–284, 2018.
- [40] L. Liu, Y. Shen, T. Li, and C. Caramanis, “High dimensional robust sparse regression,” in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 411–421.
- [41] J. Wen *et al.*, “Robust sparse linear discriminant analysis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 390–403, 2018.
- [42] C.-N. Li, Y.-H. Shao, W. Yin, and M.-Z. Liu, “Robust and sparse linear discriminant analysis via an alternating direction method of multipliers,” *IEEE transactions on neural networks and learning systems*, vol. 31, no. 3, pp. 915–926, 2019.
- [43] Z. Allen-Zhu, Y. T. Lee, and L. Orecchia, “Using optimization to obtain a width-independent, parallel, simpler, and faster positive sdp solver,” in *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA ’16, Arlington, Virginia: Society for Industrial and Applied Mathematics, 2016, pp. 1824–1831, ISBN: 9781611974331.
- [44] Y. LeCun, C. Cortes, and C. Burges, “Mnist handwritten digit database,” *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, vol. 2, 2010.
- [45] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, 2019.

- [46] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [47] N. Cesa-Bianchi and G. Lugosi, “Combinatorial bandits,” *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1404–1422, 2012.
- [48] Y. Gai, B. Krishnamachari, and R. Jain, “Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations,” *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, 2012.
- [49] W. Chen, Y. Wang, and Y. Yuan, “Combinatorial multi-armed bandit: General framework and applications,” in *International Conference on Machine Learning*, PMLR, 2013, pp. 151–159.
- [50] Z. Wen, B. Kveton, and A. Ashkan, “Efficient learning in large-scale combinatorial semi-bandits,” in *International Conference on Machine Learning*, PMLR, 2015, pp. 1113–1122.
- [51] S. Das and E. Kamenica, “Two-sided bandits and the dating market.,” in *IJCAI*, Citeseer, vol. 5, 2005, p. 19.
- [52] L. T. Liu, H. Mania, and M. Jordan, “Competing bandits in matching markets,” in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 1618–1628.
- [53] M. Jagadeesan, A. Wei, Y. Wang, M. Jordan, and J. Steinhardt, “Learning equilibria in matching markets from bandit feedback,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [54] F. J. Anscombe, “Sequential estimation,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 15, no. 1, pp. 1–21, 1953.
- [55] M. Ghosh, N. Mukhopadhyay, and P. K. Sen, *Sequential estimation*. John Wiley & Sons, 2011.
- [56] T. Anderson, *The theory and practice of online learning*. Athabasca University Press, 2008.
- [57] S. Shalev-Shwartz *et al.*, “Online learning and online convex optimization,” *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.
- [58] E. Hazan *et al.*, “Introduction to online convex optimization,” *Foundations and Trends® in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

- [59] L. Pizzato, T. Rej, T. Chung, I. Koprinska, and J. Kay, “Recon: A reciprocal recommender for online dating,” in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 207–214.
- [60] P. Xia, B. Liu, Y. Sun, and C. Chen, “Reciprocal recommendation system for on-line dating,” in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2015, pp. 234–241.
- [61] N. Tomizawa, “On some techniques useful for solution of transportation network problems,” *Networks*, vol. 1, no. 2, pp. 173–194, 1971.
- [62] J. Edmonds and R. M. Karp, “Theoretical improvements in algorithmic efficiency for network flow problems,” *Journal of the ACM (JACM)*, vol. 19, no. 2, pp. 248–264, 1972.
- [63] J. M. Kurtzberg, “On approximation methods for the assignment problem,” *Journal of the ACM (JACM)*, vol. 9, no. 4, pp. 419–439, 1962.
- [64] D. Avis, “A survey of heuristics for the weighted matching problem,” *Networks*, vol. 13, no. 4, pp. 475–493, 1983.
- [65] R. Duan and S. Pettie, “Linear-time approximation for maximum weight matching,” *Journal of the ACM (JACM)*, vol. 61, no. 1, pp. 1–23, 2014.
- [66] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.
- [67] W. Chu, L. Li, L. Reyzin, and R. Schapire, “Contextual bandits with linear payoff functions,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings*, 2011, pp. 208–214.
- [68] L. Li, Y. Lu, and D. Zhou, “Provably optimal algorithms for generalized linear contextual bandits,” in *International Conference on Machine Learning*, PMLR, 2017, pp. 2071–2080.
- [69] K. Chen, I. Hu, and Z. Ying, “Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs,” *The Annals of Statistics*, vol. 27, no. 4, pp. 1155–1163, 1999.
- [70] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, “Improved algorithms for linear stochastic bandits,” *Advances in neural information processing systems*, vol. 24, pp. 2312–2320, 2011.

- [71] V. Dani, T. P. Hayes, and S. M. Kakade, “Stochastic linear optimization under bandit feedback,” in *Conference on Learning Theory*, 2008.
- [72] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee, “Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields,” *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1386–1398, 2015.
- [73] E. W. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature,” *Decision support systems*, vol. 50, no. 3, pp. 559–569, 2011.
- [74] S. Maniraj, A. Saini, S. Ahmed, and S. Sarkar, “Credit card fraud detection using machine learning and data science,” *International Journal of Engineering Research*, vol. 8, no. 9, pp. 110–115, 2019.
- [75] A. Bratko, B. Filipič, G. V. Cormack, T. R. Lynam, and B. Zupan, “Spam filtering using statistical data compression models,” *The Journal of Machine Learning Research*, vol. 7, pp. 2673–2698, 2006.
- [76] T. S. Guzella and W. M. Caminhas, “A review of machine learning approaches to spam filtering,” *Expert Systems with Applications*, vol. 36, no. 7, pp. 10 206–10 222, 2009.
- [77] E. G. Dada, J. S. Bassi, H. Chiroma, A. O. Adetunmbi, O. E. Ajibuwa, *et al.*, “Machine learning for email spam filtering: Review, approaches and open research problems,” *Heliyon*, vol. 5, no. 6, e01802, 2019.
- [78] G.-C. Lin, W.-J. Wang, C.-M. Wang, and S.-Y. Sun, “Automated classification of multi-spectral mr images using linear discriminant analysis,” *Computerized Medical Imaging and Graphics*, vol. 34, no. 4, pp. 251–268, 2010.
- [79] N. Jothi, W. Husain, *et al.*, “Data mining in healthcare—a review,” *Procedia computer science*, vol. 72, pp. 306–313, 2015.
- [80] V. Zuber and K. Strimmer, “Gene ranking and biomarker discovery under correlation,” *Bioinformatics*, vol. 25, no. 20, pp. 2700–2707, 2009.
- [81] K. Kim *et al.*, “Urine metabolomics analysis for kidney cancer detection and biomarker discovery,” *Molecular & cellular proteomics*, vol. 8, no. 3, pp. 558–570, 2009.
- [82] F. Chang, S. He, and C. Dang, “Assisted selection of biomarkers by linear discriminant analysis effect size (lefse) in microbiome data,” *JoVE (Journal of Visualized Experiments)*, no. 183, e61715, 2022.

- [83] T. Lattimore, K. Crammer, and C. Szepesvári, “Linear multi-resource allocation with semi-bandit feedback,” in *NIPS*, 2015, pp. 964–972.
- [84] S. Vakili, Q. Zhao, and Y. Zhou, “Time-varying stochastic multi-armed bandit problems,” in *2014 48th Asilomar Conference on Signals, Systems and Computers*, IEEE, 2014, pp. 2103–2107.
- [85] X. Xu, F. Dong, Y. Li, S. He, and X. Li, “Contextual-bandit based personalized recommendation with time-varying user interests,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 6518–6525.