Title

Robust Statistical Inference with Privacy and Fairness Constraints

Author

Jin, Yulu

Publication Date

2023

Peer reviewed|Thesis/dissertation

Robust Statistical Inference with Privacy and Fairness Constraints

By

YULU JIN

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Lifeng Lai, Chair

Zhi Ding

Khaled Abdel-Ghaffar

Committee in Charge

2023

Abstract

In recent years, machine learning has witnessed remarkable progress, finding diverse applications and achieving notable success in addressing complex problems. However, these achievements have been accompanied by growing ethical concerns, rooted in the potential of machine learning systems to produce unreliable decisions, inadvertently disclose sensitive information, and exhibit biases. The need for trustworthy machine learning systems, characterized by attributes like privacy, fairness, and robustness, has become increasingly pressing. This dissertation attempts to addressing these critical challenges through an investigation into algorithmic adversarial robustness, the preservation of privacy within cloud-based frameworks, and the development of adversarially robust fairness-aware models.

In the first part, we investigate the adversarial robustness of hypothesis testing rules. In the considered model, after a sample is generated, it will be modified by an adversary before being observed by the decision maker. The decision maker needs to decide the underlying hypothesis that generates the sample from the adversarially-modified data. We formulate this problem as a minimax hypothesis testing problem, in which the goal of the adversary is to design attack strategy to maximize the error probability while the decision maker aims to design decision rules so as to minimize the error probability. We consider both hypothesis-aware case, in which the attacker knows the true underlying hypothesis, and hypothesis-unaware case, in which the attacker does not know the true underlying hypothesis. We solve this minimax problem and characterize the corresponding optimal strategies for both cases.

In the second part, we propose a general framework to provide a desirable trade-off between inference accuracy and privacy protection in the inference as service scenario (IAS). Instead of sending data directly to the server, the user will pre-process the data through a privacy-preserving mapping, which will increase privacy protection but reduce inference accuracy. To properly address the trade-off between privacy protection and inference accuracy, we formulate an optimization problem to find the optimal privacy-preserving mapping. Even though the problem is non-convex

in general, we characterize nice structures of the problem and develop an iterative algorithm to find the desired privacy-preserving mapping, with convergence analysis provided under certain assumptions. From numerical examples, we observe that the proposed method has better performance than gradient ascent method in the convergence speed, solution quality and algorithm stability.

In the third part, we take a first step towards answering the question of how to design fair machine learning algorithms that are robust to adversarial attacks. Using a minimax framework, we aim to design an adversarially robust fair regression model that achieves optimal performance in the presence of an attacker who is able to add a carefully designed adversarial data point to the dataset or perform a rank-one attack on the dataset. By solving the proposed nonsmooth nonconvex-nonconcave minimax problem, the optimal adversary as well as the robust fairness-aware regression model are obtained. For both synthetic data and real-world datasets, numerical results illustrate that the proposed adversarially robust fair models have better performance on poisoned datasets than other fair machine learning models in both prediction accuracy and group-based fairness measure.

Acknowledgement

I stand at this pivotal juncture in my academic journey, deeply indebted to those who have steadfastly accompanied me throughout my PhD. Their unwavering support and encouragement have been the linchpin in achieving this significant milestone.

At the very forefront, I reserve my deepest gratitude for my advisor, Prof. Lifeng Lai. Throughout this pivotal period, his insightful guidance has profoundly shaped my academic journey. From steering me through the labyrinth of research domains to mentoring me in nuanced academic endeavors-literature review, problem analysis, publishing, and presentation-his presence has been emblematic of constant support. Prof. Lai's unwavering support has been more than just instructive; it has acted as a driving force, invigorating my academic pursuits. His encouragement transcends the current scope of my research, inspiring confidence in a lasting academic career. Beyond the scholastic realm, Prof. Lai's wisdom has enriched my perspective on life and charted potential career avenues.

I extend my heartfelt appreciation to my committee members, Prof. Zhi Ding and Prof. Khaled Abdel Ghaffar. Their incisive feedback and encouragement have been invaluable. Moreover, the probing questions they posed have invariably nudged me to broaden my research horizons, exploring varied perspectives and dimensions.

The entire cohort of our research group-Xiaochuan Ma, Guanlin Liu, Puning Zhao, Xinyang Cao, Xinyi Ni, Fuwei Li, Minhui Huang, Chenye Yang, Parisa Oftadeh, and Haodong Liang-deserve special mention. Their camaraderie, unwavering support, and technical guidance have been pivotal in shaping this dissertation.

Lastly, to the anchors of my life-my parents and grandparents, who gave me life and unwavering strength; to the friends who have journeyed alongside for the past 26 years, enriching every moment; and to Kaiming Fu, who brings completion to my being.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning (ML) has found extensive applications across various industrial sectors, spanning from autonomous vehicles [2–4] and medical diagnostics [5,6] to robotics [7,8]. In these domains, ML performs a wide range of tasks, including speech recognition [9,10], object detection [11,12], and decision making [13–17]. However, the proliferation of ML solutions has ushered in a new era, accompanied by significant challenges. These challenges are particularly pronounced in the areas of security [18–25], privacy [26–30], and fairness [31–33], especially within safety-critical applications.

In this chapter, we provide the introduction of this dissertation. In Section 1.1, we introduce the background. In Section 1.2, we introduce basic tools that are used in this dissertation. We then discuss main contributions of this dissertation in Section 1.3.

1.1 Background

In this section, we review recent research on security, privacy and fairness issues associated with machine learning in the following domains:

(1) Security issues;

(2) Privacy concern;

(3) Fairness issues.

1.1.1 Security issues

Machine learning models are not robust to adversarial attacks and are extremely susceptible to a phenomenon called adversarial examples [34]. By adding hardly perceptible perturbations on the input data, the decision of a deep network can be easily manuplicated. For example, an original image of an ice bear is concluded as "ice bear" with 85.8% confidence by the network [1]. Then by adding the carefully constructed adversarial perturbation, an image that looks exactly the same to a human is obtained, which the network thinks with 100% confidence as a "dishwasher". In practical applications, it has been observed that these adversarial examples are consistently misclassified at a notably higher rate than examples perturbed by random noise, even when the magnitude of the noise greatly surpasses that of the adversarial perturbation [34]. Moreover, the issue extends beyond individual models, as adversarial examples often possess a transferability property. An adversarial example engineered to confound one model, say M_1, frequently has the same effect on another model, such as M_2. This property allows for the generation of adversarial examples and the execution of misclassification attacks on machine learning systems without access to the underlying model [35].

There are many attack algorithms designed to find adversarial examples efficiently [36–38]. At the same time, there are significant amount of research works that focus on developing defense strategies with the goal of constructing robust classifiers that can work well in the presence of adversarial perturbations [35, 39, 40]. While a proposed defense is often empirically shown to be successful against the set of attacks known at the time, new stronger attacks are subsequently discovered that render the defense useless. For example, defensive distillation [35] and adversarial training against the Fast Gradient Sign Method [36] were two defenses that were later shown to be ineffective against stronger attacks [41, 42]. In order to break this arms race between attackers and defenders, there are many studies establishing the fundamental limits on the robustness of classifiers [43, 44]. Most of these works rely on tools from concentration of measure [45] and

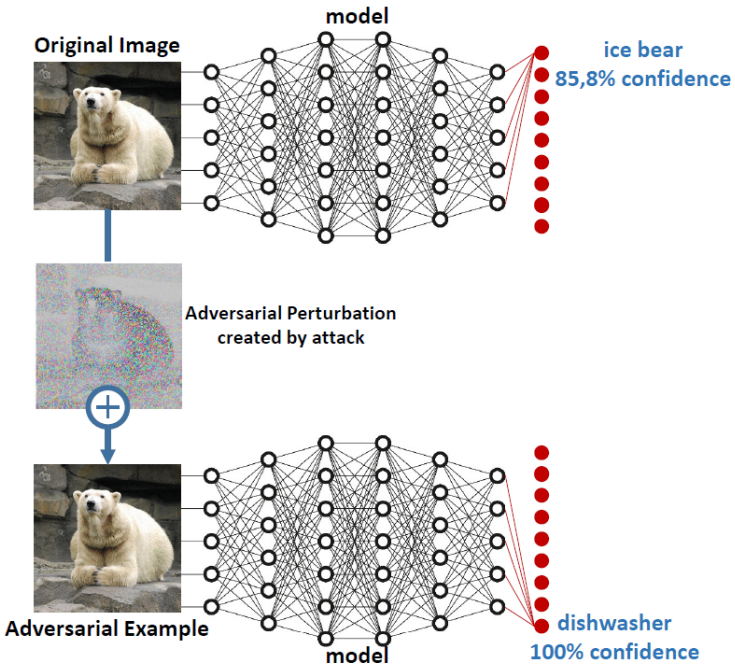


Figure 1.1: Adversarial example [1].

provide interesting results when the dimension of data is high and the distribution of data satisfies certain conditions.

At the meantime, motivated by growing applications of various signal processing and statistical inference algorithms in safety and security-related applications [20, 21], there is also an increasing interest in the study of adversary robustness of statistical inference algorithms [34, 36–38, 46–49]. The purpose of these studies is to understand the robustness of these algorithms in the adversarial setup, so as to properly design systems that are safe and secure even under adversarial attacks. The investigation of adversary robustness of statistical algorithms is related to, but different from, the large volume of work on classic robust statistics [50–55]. The classic robust statistical inference mainly focuses on distributional robustness, in which the true distributions of data lie in the neighborhood of nominal distributions [53, 56, 57]. On the other hand, the attack in the adversary robustness model is more powerful. In particular, in the adversarial robustness models, an adversary is typically assumed to have access to the data sample and can make data-dependent changes. The decision maker then has to make statistical inference based on the adversarially-modified data [58].

Hence, it is of our interest to study the performance bound of the powerful adversarial models. Under this setting, to reveal the structures of the optimal attack and defense strategies, as the adversary and the defender perform opposite rules, the game between them can be formulated as a minimax problem and will be studied in Chapter 2.

1.1.2 Privacy concern

The impressive accuracy achieved by modern ML models in business, medicine and communication motivates many data holders to apply ML to their own datasets. Existing ML frameworks, however, are not easy to deploy by non-expert users due to a large number of configuration parameters and general lack of understanding of why and how modern ML works [59]. Furthermore, ML expertise is scarce and often unrelated to data holders' primary competency. At the same time, the Internet of Things (IoT) is an emerging communication paradigm that aims at connecting different kinds of devices to the Internet [60–62]. Within the past decade, the number of IoT devices being introduced in the market has increased dramatically due to its low cost and convenience [63]. Sensors of IoT devices could generate contexts at a high velocity and the inference with the contexts becomes an essential component for IoT applications [64]. However, considering the complexity of state-of-the-art machine learning models, it is difficult to run them on IoT devices. Thus, one of the emerging solutions to solve the two problems mentioned above is so called inference-as-a-service (IAS) [65, 66]. IAS is known as a cloud service that manages various types of inferences effectively.

With cloud services, machine learning algorithms can be run on the cloud providers' infrastructure where training and deploying machine learning models are performed on cloud servers. Once the models are deployed, users can use these models to make predictions without having to worry about maintaining the models and the service [67]. Several such services are currently offered including Microsoft Azure Machine Learning, Google Prediction API, GraphLab, and Ersatz Labs. However, such service brings privacy issues, as the devices will send their data to the cloud without knowing where these data is stored or what future purposes these data might serve.

There are some interesting works that attempt to address this issue using Homomorphic Encryption (HE) technique [68–70]. Unfortunately, the complexity of HE based solution is very high, and its privacy relies on the (unproved) assumption that certain mathematical problems are difficult to solve. The most notable shortcoming of practical Homomorphic Encryption schemes is that operations in practical schemes are limited to addition and multiplication. Consequently, we need to adopt algorithms within these limitations. However, the computation performed over sensitive data by machine learning models, especially neural networks, is usually very complex and cannot be simply translated to encrypted versions without modification.

There exist many other privacy-preserving techniques that are based on perturbations of data, which provide privacy guarantees at the expense of a loss of accuracy [28–30]. For example, k-anonymity is proposed by Samarati and Sweeney [26], which requires that each record is indistinguishable from at least k–1 other records within the dataset. Differential privacy works by adding a pre-determined amount of randomness into a computation performed on a data set [27]. These concepts and techniques are very useful for the privacy protection of data analysis through a dataset or database. Moreover, various minimax formulations and algorithms have also been proposed to defend against inference attack in different scenarios [71–73]. Bertran et al. [71] proposed an optimization problem where the terms in the objective function were defined in terms of mutual information, showed the performance bound for the optimization problem and learned the sanitization transform in a data-driven fashion using an adversarial approach with Deep Neural Networks (DNNs). Under their formulation, they analyzed a trade-off between utility loss and attribute obfuscation under the constraint of the attribute obfuscation $I(A; Z) \leq k$. Feutry et al. [72] measured the utility and privacy by expected risks, formulated the utility-privacy trade-off as a min-diff-max optimization problem and proposed a learning-based and task-dependent approach to solve this problem, while only deterministic mechanisms are considered. To address this issue, a privacy-preserving adversarial network was proposed in [73] by employing adversarially-trained neural networks to implement randomized mechanisms and to perform a variational approximation of mutual information privacy.

Hence, it is of our interest to address the fundamental trade-off between inference accuracy and privacy protection from information theory perspective. Instead of sending data directly to the server, the user will preprocess the data through a privacy-preserving mapping. This privacy-preserving mapping has two opposing effects. On one hand, it will prevent the server from observing the data directly and hence enhance the privacy protection. On the other hand, this might reduce the inference accuracy. To properly address the trade-off between these two competing goals, an optimization problem can be formulated and will be studied in Chapter 3.

1.1.3 Fairness issues

ML models have been used in various domains, including several security and safety critical applications, such as banking, education, healthcare, law enforcement etc. However, it has become increasingly evident that ML algorithms can inadvertently perpetuate or even exacerbate biases [31, 32], particularly those related to race or gender, thereby raising concerns about fairness and equity in their outcomes. For instance, notable instances of bias have been observed in systems like the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a tool used by judges to assess an offender's risk of recommitting a crime [74]. Investigations into COMPAS revealed that it displayed bias against African-American individuals, assigning them higher risk scores than their Caucasian counterparts with similar profiles [74]. Similar findings have been made in other areas, such as an AI system that judges beauty pageant winners but was biased against darker-skinned contestants [31], or facial recognition software in digital cameras that over-predicts Asians as blinking [75].

These fairness issues in machine learning outcomes can be attributed to at least two main sources: biases present in the data and biases introduced by the algorithms themselves. Data, especially in big data scenarios, often reflect inherent heterogeneities arising from subgroups with distinct characteristics and behaviors. These variations can introduce bias into the data, which, when used to train models, may result in unfair and inaccurate predictions. Bias in data can stem from multiple sources, such as historical bias, representation bias, measurement bias, evaluation

bias, aggregation bias,population bias, etc. [76]. In order to mitigate the effects of bias in data, some general methods have been proposed that advocate having good practices while using data. For example, [77] proposes having labels, just like nutrition labels on food, to better categorize each data for each task.

For the algorithmic fairness, one should first define the notion of fairness to fight against discrimination and achieve fairness. However, the absence of a universally accepted definition of fairness highlights the complexity of this challenge. Different cultures and perspectives may favor distinct interpretations of fairness, making it difficult to arrive at a single, universally applicable definition. Consequently, a range of fairness definitions and corresponding methods have emerged, tailored to specific applications or preferences. These methods can be categorized into pre-processing, which modifies the data that the algorithm learns from [78]; in-processing, which modifies the algorithm's objective function to incorporate a fairness constraint or penalty [79–81]; post-processing, which modifies the predictions produced by the algorithm [82].

In the meantime, a large body of work has shown that ML models are vulnerable to various types of attacks [22–24]. Thus, a major and natural concern for fair machine learning algorithms is their robustness in adversarial environments. Recent works show that well-designed adversarial samples can significantly reduce the test accuracy as well as exacerbating the fairness gap of ML models [83–86].

In light of the vulnerabilities of existing fair machine learning algorithms, there is a pressing need to design fairness-aware learning algorithms that are robust to adversarial attacks. As the first step towards this goal, we focus on regression problems and design a fair regression model that is robust to adversarial attacks, as discussed in Chapter 4.

1.2 Preliminaries

In this section, we introduce basic tools that will be used in this dissertation.

1.2.1 Minimax problem and saddle points property

Given $\phi : X \times Z \to \mathbb{R}$, where $X \subset \mathbb{R}^n, Z \subset \mathbb{R}^m$, consider

$$\inf_{x\in X} \sup_{z\in Z} \phi(x, z) \text{ and } \sup_{z\in Z} \inf_{x\in X} \phi(x, z).$$

The minimax inequality gives that

$$\sup_{z\in Z} \inf_{x\in X} \phi(x, z) \leq \inf_{x\in X} \sup_{z\in Z} \phi(x, z). \tag{1.1}$$

Definition 1. *(x^*, z^*) is called a saddle point of ϕ if*

$$\phi(x^*, z) \leq \phi(x^*, z^*) \leq \phi(x, z^*), \forall x \in X, \forall z \in Z. \tag{1.2}$$

Proposition 1. *(x^*, z^*) is a saddle point of ϕ if and only if the minimax equality holds and*

$$x^* \in \arg\inf_{x\in X} \sup_{z\in Z} \phi(x, z), z^* \in \arg\sup_{z\in Z} \inf_{x\in X} \phi(x, z). \tag{1.3}$$

Then we have the minimax theorem,

Theorem 1. *If*

- *X and Z are convex and compact sets;*
- *$\phi(\cdot, z)$ is a continuous function;*
- *For each $z \in Z$, the function $\phi(\cdot, z)$ is convex;*
- *For each $x \in X$, the function $\phi(x, \cdot)$ is closed and concave;*

then the minimax equality holds [87].

The saddle point analysis will be used in our proposed minimax problem in Chapter 2.

1.2.2 Robust Hypothesis Testing

The detection of the presence or absence of an event with a specified accuracy is fundamental to statistical inference and binary hypothesis testing is the usual starting point. Formally, any real world example of binary decision making problem can be modeled by a binary hypothesis test, where under each hypothesis H_j, a received data $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ follows a particular probability distribution $f_j, j \in \{0, 1\}$, i.e.

- $\mathcal{H}_0$: $X \sim f_0$.
- $\mathcal{H}_1$: $X \sim f_1$.

Consider the general decision function $\delta(x_1, x_2, \cdots, x_n)$, where $\delta(x_1, x_2, \cdots, x_n) = 0$ means that $\mathcal{H}_0$ is accepted and $\delta(x_1, x_2, \cdots, x_n) = 1$ means that $\mathcal{H}_1$ is accepted. Since the function takes on only two values, the test can also be specified by the set A over which $\delta(x_1, x_2, \cdots, x_n) = 0$ while the complement of this set is the set where $\delta(x_1, x_2, \cdots, x_n) = 1$. Define two probabilities of error

$$\alpha = Pr[\delta(x_1, x_2, \cdots, x_n) = 1 | \mathcal{H}_0 \text{ is true}], \tag{1.4}$$

$$\beta = Pr[\delta(x_1, x_2, \cdots, x_n) = 0 | \mathcal{H}_1 \text{ is true}]. \tag{1.5}$$

In general, we wish to minimize both probabilities, but there is a trade-off between them. We can either minimize one of the probabilities of error subject to a constraint on the other probability of error or construct an error probability function consisting of both probabilities of error. Assuming that the prior probability of two hypotheses are Pr($\mathcal{H}_0$) and Pr($\mathcal{H}_1$), then the error probability P_E can be written as

$$P_E(\delta(\cdot)) = \alpha \text{Pr}(\mathcal{H}_0) + \beta \text{Pr}(\mathcal{H}_1). \tag{1.6}$$

Suppose P_0 and P_1 under $\mathcal{H}_0$ and $\mathcal{H}_1$ are known and using Bayes' rule, we can obtain the

posterior probabilities of $\mathcal{H}_0$ and $\mathcal{H}_1$:

$$P(\mathcal{H}_0|x_1, x_2, \cdots, x_n) = \frac{f_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n|\mathcal{H}_0)\text{Pr}(\mathcal{H}_0)}{f_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n)}, \tag{1.7}$$

$$P(\mathcal{H}_1|x_1, x_2, \cdots, x_n) = \frac{f_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n|\mathcal{H}_1)\text{Pr}(\mathcal{H}_1)}{f_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n)}. \tag{1.8}$$

If one further assumes that the prior probability of each hypothesis is the same, then the optimal way to decide between $\mathcal{H}_0$ and $\mathcal{H}_1$ is to compare $P(\mathcal{H}_0|x_1, x_2, \cdots, x_n)$ and $P(\mathcal{H}_1|x_1, x_2, \cdots, x_n)$, and accept the hypothesis with the higher posterior probability. This is the idea behind the maximum a posterior (MAP) test. Recall the definition of error probability in (1.6), and we can see that the error probability is minimized by the MAP test since we are choosing the hypothesis with the highest posterior probability.

However, in practice, for random observation $Y \in \mathbb{R}$, when the true distribution $g_j(y)$ deviates from the assumed nominal distribution $f_j(y)$, the performance of the likelihood ratio detector is no longer optimal and it may perform poorly. Various robust hypothesis testing frameworks have been developed to address the issue with distribution misspecification and outliers [88, 89].

The robust detectors are constructed by introducing various uncertainty sets for the distributions under the null and the alternative hypotheses. In non-parametric setting, Huber's work [52] considers the so-called ϵ-contamination sets, which contain distributions that are close to the nominal distributions in terms of total variation metric. [57] considers uncertainty set induced by Kullback-Leibler divergence around a nominal distribution. Under this setup, the actual density $g_j(y)$ of Y under $\mathcal{H}_j$ is not known exactly and belongs to the neighborhood

$$\mathcal{F}_j = \{g_j : D_{\text{KL}}(g_j|f_j) \leq \epsilon_j\},$$

where

$$D_{\text{KL}}(g|f) = \int_{-\infty}^{\infty} \log\left[\frac{g(y)}{f(y)}\right] dy.$$

Use $\mathcal{D}$ to denote the class of pointwise randomized decision rules $\delta(y)$ such that if $Y = y$, $\mathcal{H}_1$ is selected with probability $\delta(y)$ and $\mathcal{H}_0$ is selected with probability $1 - \delta(y)$. Then denote the probability of false alarm and the probability of miss detection as

$$P_F(\delta, g_0) = \int_{-\infty}^{\infty} \delta(y) g_0(y) dy,$$
$$P_M(\delta, g_1) = \int_{-\infty}^{\infty} (1 - \delta(y)) g_1(y) dy.$$

Then $P_F(\delta, g_0)$ is separately linear in δ and g_0. $P_M(\delta, g_1)$ is separately linear in δ and g_1. Assume that the two hypotheses are equally likely, then the probability of error is given by

$$P_E(\delta, g_0, g_1) = \frac{1}{2} \left[P_F(\delta, g_0) + P_M(\delta, g_1) \right].$$

Then the robust hypothesis testing problem is solved via the minimax problem

$$\min_{\delta \in \mathcal{D}} \max_{(g_0, g_1) \in \mathcal{F}_0 \times \mathcal{F}_1} P_E(\delta, g_0, g_1). \tag{1.9}$$

Note that $P_E(\delta, g_0, g_1)$ is convex in δ and concave in g_0 and g_1. The set $\mathcal{F}_0 \times \mathcal{F}_1$ is convex and compact. $\mathcal{D}$ is convex and compact with respect to the infinity norm. By Theorem 1, there exists a saddle point $(\delta_R, (g_0^L, g_1^L))$ for the minimax problem. Here, δ_R is the robust/minimax test, whereas g_0^L and g_1^L are the least favorable densities in $\mathcal{F}_0 \times \mathcal{F}_1$. Then the saddle point has the property

$$P_E(\delta, g_0^L, g_1^L) \geq P_E(\delta_R, g_0^L, g_1^L) \geq P_E(\delta_R, g_0, g_1). \tag{1.10}$$

Then the first inequality indicates that the robust test δ_R is the optimal Bayesian test for the least-favorable pair (g_0^L, g_1^L). In particular, for the likelihood ratio function defined as $L_L(y) = \frac{g_1^L(y)}{g_0^L(y)}$, the

decision test is

$$\delta_R(y) = \begin{cases} 1, & L_L(y) > 1 \\ arbitrary, & L_L(y) = 1 \\ 0, & L_L(y) < 1. \end{cases}$$

Thus, robust detectors usually depend on the least-favorable distributions, which are designed by solving the second inequality of (1.10). Given δ_R, to solve the maximization problem

$$\max_{(g_0,g_1)\in\mathcal{F}_0\times\mathcal{F}_1} P_E(\delta_R, g_0, g_1),$$

[57] made two strong assumptions: 1) the nominal likelihood ratio $L_L(y) = \frac{f_1(y)}{f_0(y)}$ is a monotone increasing function of y; 2) f_0 and f_1 admit the symmetry $f_1(y) = f_0(-y)$.

Moreover, [53, 56] also explore the robust hypothesis testing problem with different assumptions and different measures in defining the neighborhoods $\mathcal{F}_j$. Although there has been much success in theoretical results, computation remains a major challenge in finding robust detectors and finding he least-favorable distributions in general.

1.2.3 ADMM

The alternating direction method of multipliers (ADMM) is an algorithm that solves complex optimization problems by breaking them into smaller problems, each of which is then easier to handle. It takes the form of a decomposition-coordination procedure, in which the solutions to small local sub-problems are coordinated to find a solution to a large global problem [90]. ADMM can be viewed as an attempt to blend the benefits of dual decomposition and augmented Lagrangian methods for constrained optimization [90].

With $f(\cdot), g(\cdot)$ being convex functions, ADMM solves problems in the form

$$\begin{aligned} \min \quad & f(x) + g(y), \\ \text{subject to} \quad & Ax + By = c, \end{aligned} \tag{1.11}$$

with variables $x \in \mathcal{R}^n, y \in \mathcal{R}^m$ and $A \in \mathcal{R}^{p\times n}, B \in \mathcal{R}^{p\times m}, c \in A \in \mathcal{R}^p$.

As in the method of multipliers, the augmented Lagrangian function is

$$\mathcal{L}_\rho(x, y; \lambda) = f(x) + g(y) - < \lambda, Ax + By - c > + \frac{\rho}{2}\|Ax + By - c\|^2. \tag{1.12}$$

With $\rho > 0$, ADMM consists of the iterations

- $x^{k+1} = \arg\min_x \mathcal{L}_\rho(x, y^k; \lambda^k)$,
- $y^{k+1} = \arg\min_y \mathcal{L}_\rho(x^{k+1}, y; \lambda^k)$,
- $\lambda^{k+1} = \lambda^k - \rho(Ax^{k+1} + By^{k+1} - c)$.

The procedure consists of an x-minimization step, a y-minimization step and a dual variable update. The convergence properties of ADMM have been studied extensively in the literature [91–93]. Because of its wide applicability in multiple fields, ADMM is a popular means of solving optimization problems. However, the original method only considers the two-block separable structure.

For the case of $n \geq 3$, numerous research efforts have been devoted to analyzing the convergence of multi-block ADMM and its variants for the linearly constrained separable convex optimization model. Recent work [94] has shown that the n-block ADMM does not necessarily converge, even for a nonsingular square system of linear equations. Various methods have been proposed to overcome the divergence issue of multi-block ADMM. One typical solution is to combine correction steps with the output of n-block ADMM [92, 95]. If at least $n - 2$ functions in the objective are strongly convex, it has been shown that the ADMM process is globally convergent, provided that the penalty parameter λ is restricted to a specific range [96, 97]. Without strong convexity, it has been shown in [98] that the n-block ADMM with a small dual step size is linearly convergent provided that the objective function satisfies certain error bound conditions. Some recent studies [99] have demonstrated the convergence of multi-block ADMM under some other conditions, and some convergent proximal variants of the multi-block ADMM have been proposed for solving convex linear or quadratic conic programming problems [97]. Moreover, [100] pro-

posed a randomly modified variant of the multi-block ADMM, called randomly permuted ADMM (RPADMM). At each step, RPADMM forms a random permutation of $\{1, 2, \cdots, n\}$ (known as block sampling without replacement), and updates the primal variables $x_i, i = \{1, 2, \cdots, n\}$ in the order of the chosen permutation followed by the regular multiplier update. Surprisingly, RPADMM is convergent in expectation for any non-singular square system of linear equations.

In contrast to the separable case, studies on the convergence properties of n-block ADMM with non-separable objective, even for $n = 2$, are limited. In [101], the authors demonstrated that when the problem is convex but not necessarily separable, and certain error bound conditions are satisfied, the ADMM iteration converges to some primal-dual optimal solution, provided that the step size in the update of the multiplier is sufficiently small. However, the step size usually depends on some unknown parameters associated with the error bound, and may thus be difficult to compute, which often makes the algorithm less efficient. [102] investigated the convergence of a majorized ADMM for the convex optimization problem with a coupled smooth objective function, which includes the 2-block ADMM as a special case. Convergence was established for the case when some restrictions are satisfied and the sub-problems of the ADMM admit unique solutions. [103] studied the convergence and ergodic complexity of a 2-block proximal ADMM and its variants for the non-separable convex optimization by assuming some additional conditions on the problem data.

1.2.4 Privacy metrics

A technical privacy metric takes properties of a system as an input (e.g., the amount of sensitive information leaked or the number of users who are indistinguishable with respect to some characteristic) and yields a numerical value, which allows us to quantify the privacy level in a system. Privacy metrics can be used in different contexts, and they can differ with regard to the kind of adversary they consider, the data sources they assume to be available to the adversary, and the aspects of privacy they measure. However, the diversity and complexity of privacy metrics in the literature makes an informed choice of metrics challenging. As a result, instead of using existing

metrics, new metrics are proposed frequently, and privacy studies are often incomparable. In the following, we will explain and discuss a selection of privacy metrics.

- **Uncertainty**

Uncertainty metrics assume that an adversary who is uncertain of his estimate cannot breach privacy as effectively as one who is certain. Many uncertainty metrics build on entropy, an information theoretic notion to measure uncertainty [104].

Shannon entropy is the basis for many other metrics. In general, entropy measures the uncertainty associated with predicting the value of a random variable. As a privacy metric, it can be interpreted as the number of bits of additional information the adversary needs to identify a user [105].

Rényi entropy is a generalization of Shannon entropy that also quantifies the uncertainty in a random variable. It uses an additional parameter α, and Shannon entropy is the special case with $\alpha \to 1$. In particular, we have the privacy with respect to Rényi entropy as

$$\text{privacy}_{\text{RE}} = H_\alpha(X) = \frac{1}{1-\alpha} \log_2 \sum_{x \in X} p(x)^\alpha.$$

Hartley entropy H_0 or max-entropy is the special case with $\alpha = 0$. It depends only on the number of users and is a best-case scenario because it represents the ideal privacy situation for a user. Min-entropy H_∞ is the special case with $\alpha = \infty$ which is a worst-case scenario because it only depends on the user for whom the adversary has the highest probability [106].

- **Data Similarity**

Data similarity metrics measure properties of observable or published data. They are usually independent of the adversary and derive the privacy level solely from the features of observable data.

k-Anonymity was originally proposed to prepare statistical databases for publication. For example, a medical database would contain both identifying information (e.g., the names of individuals) and sensitive information (e.g., their medical conditions). k-Anonymity assumes that identifying columns are removed from a database before publication, and then demands that the database

table can be grouped into equivalence classes with at least k rows that are indistinguishable with respect to their quasi-identifiers q [107]. Each equivalence class E contains all rows that have the same values for each quasi-identifier q. To increase the size of equivalence classes to a minimum of k rows, several algorithms exist to transform a given database to make it k-anonymous, such as suppression, generalization and random sampling [108]. However, studies have shown k-anonymity to be insufficient, especially for high-dimensional data and against correlation with other data sets [109], because it fails to protect against attribute disclosure [110].

The l-diversity principle modifies k-anonymity to bound the diversity of published sensitive information. It states that every equivalence class E must contain at least l well-represented sensitive values. This general principle can be instantiated in different ways. In the simplest form, the l-diversity principle requires l distinct values in each equivalence class. However, this simple instantiation does not prevent probabilistic inference attacks [111]. Although l-diversity is an improvement to k-anonymity, it has been shown to offer insufficient protection against some attacks. In particular, it does not protect privacy when the distribution of sensitive values is skewed, or when sensitive attributes are semantically similar [111].

- **Indistinguishability**

Indistinguishability metrics indicate whether the adversary can distinguish between two items of interest. Many of these metrics are associated with privacy mechanisms that provide formal privacy guarantees.

First, we discuss differential privacy. In statistical databases, differential privacy guarantees that any disclosure is equally likely (within a small multiplicative factor ϵ) regardless of whether or not an item is in the database [112]. This guarantee is usually achieved by adding a small amount of random noise to the results of database queries. Formally, differential privacy is defined using two data sets D_1 and D_2 that differ in at most a single row, i.e., the Hamming distance between the two data sets is at most 1. A privacy mechanism, realized as a randomized function $\mathcal{K}$, operating on these data sets is ϵ-differentially private if for all sets of query responses, the output random variables for the two data sets differ by at most $\exp(\epsilon)$. However, the choice of the parameter ϵ

is difficult. It has also been shows that differential privacy's guarantees degrade in the case of correlated data, for example when nodes are added to a social network graph [113].

Distributional privacy extends differential privacy to a setting in which the data sets themselves do not need to be protected, but instead the parameters governing the generation of data need to be protected. Distributional privacy assumes a distributed setting in which smart meters apply noise to their local data, limiting the energy provider to querying this distributed database. Formally, distributional privacy uses two parameter sets θ_1 and θ_2 which govern the creation of two data sets and differ in at most one element. Furthermore, The privacy mechanism $\mathcal{K}$ is distributionally ϵ-differentially private if the probability of generating query response $\mathcal{K}_j$ is roughly the same, regardless of whether the underlying parameter set is θ_1 or θ_2 [114].

- **Error**

Error-based metrics quantify the error an adversary makes in creating his estimate. The adversary's expected estimation error measures the adversary's correctness by computing the expected distance between the true location x^* and the estimated location x using a distance metric $d()$, for example the Euclidean distance or an indicator function (in this case, the metric reduces to the adversary's probability of error). The expectation is computed over the posterior probability of the adversary's estimated locations x based on the observations y [115].

$$\text{privacy}_{\text{AEE}} = \sum_{x \in X} p(x|y) d(x, x^*).$$

In statistical parameter estimations, a common goal is to minimize the mean squared error. As a privacy metric, the mean squared error describes the error between observations y by the adversary and the true outcome x^*.

$$\text{privacy}_{\text{MSE}} = \frac{1}{|X^*|} \sum_{x^* \in X^*} \|x^* - y\|^2.$$

1.2.5 Fairness metrics

In recent years, the research community has put forth many formal and mathematical definitions of fairness to assist practitioners in developing equitable risk assessment tools. Broadly speaking, there are two main categories of definitions for algorithmic fairness: group fairness and individual fairness.

- **Group fairness**

Group fairness partition individuals into "protected groups" (often based on race, gender, or some other binary protected attribute) and ask that some statistic of a machine learning model (error rate, false positive rate, positive classification rate, etc.) be approximately equalized across those groups. To this end, numerous group fairness measures have been proposed, such as demographic parity [116], equality of opportunity [82], equalized odds [82], envy-free group fairness [117], etc. Suppose that A is the protected attribute, Y is the outcome and $\hat{Y}$ is the predictor. In the following, we present several commonly used definitions of group fairness.

Definition 2. *(Demographic Parity) $\hat{Y}$ satisfies demographic parity if $\hat{Y}$ is independent of A, i.e.*

$$\mathbb{P}(\hat{Y} = 1|A = 0) = \mathbb{P}(\hat{Y} = 1|A = 1).$$

This definition indicates that positive outcome is given to the two groups at the same rate. However, demographic parity may cripple the utility, especially in the common scenario when $\mathbb{P}(A = 0, Y = 1) \neq \mathbb{P}(A = 1, Y = 1)$ [82]. In light of this, an alternative definition is equal odds or equal opportunity.

Definition 3. *(Equal odds) $\hat{Y}$ satisfies equal odds if $\hat{Y}$ and A are independent conditional on Y, i.e.*

$$\mathbb{P}(\hat{Y} = 1|A = 0, Y = y) = \mathbb{P}(\hat{Y} = 1|A = 1, Y = y), y \in \{0, 1\}.$$

This metric essentially requires equal true positive and false positive rates between different

groups. A relaxed version of equal odds is equal opportunity, which demands only the equality of true positive rates.

Definition 4. *(Equal opportunity)* $\hat{Y}$ *satisfies equal opportunity if*

$$\mathbb{P}(\hat{Y} = 1|A = 0, Y = 1) = \mathbb{P}(\hat{Y} = 1|A = 1, Y = 1).$$

However, in certain decision making scenarios, the existing parity-based fairness notions may be too stringent and precluding more accurate decisions. To relax these parity-based notions, a preference-based notions of fairness is proposed–given the choice between various sets of decision treatments or outcomes, any group of users would collectively prefer its treatment or outcomes, regardless of the (dis)parity as compared to the other groups [117]. Other definitions of group fairness include calibration [118, 119], disparate mistreatment [120], counterfactual fairness [121], etc.

One problem for group fairness measures is that they are only suited to a limited number of coarse-grained, prescribed protected groups [122]. For groups at the intersection of multiple discriminations, or groups that have not yet been defined but may need protection [123], group fairness measures may ignore the underlying bias.

- **Individual fairness**

Individual definitions of fairness have no notion of protected groups, and instead ask for constraints on pairs of individuals. These constraints can have the semantics that "similar individuals should be treated similarly" [124], or that "less qualified individuals should not be preferentially favored over more qualified individuals" [125]. In particular, [124] first proposed a technical definition of individual fairness by presupposing a task-specific quality metric on individuals and proposing that fair algorithms should satisfy a Lipschitz condition on this metric. [125] has similar definition of fairness and requires equal false positive rate across all pairs of individuals who have negative labels.

1.3 Main contributions

In this dissertation, we contribute to the advancement of machine learning models by addressing three primary concerns: security, privacy protection, and fairness.

1.3.1 Adversarial Robustness of Hypothesis Testing

For the analysis of adversarial robustness, our goal is to understand adversarial robustness of hypothesis testing rules. In the considered model, after data samples are generated by the underlying hypothesis, an adversary can observe the samples and then modify them to other values. The decision maker only observes the modified data but still needs to determine which underlying hypothesis is true. We formulate this as a minimax hypothesis testing problem, in which the adversary aims at designing attack strategies to modify the data so as to maximize the error probability while the goal of the decision maker is to design decision rules to minimize the error probability. Our work is related to several recent interesting papers [126–128], which characterize the asymptotic equilibrium of the games between the attacker and detector, as the number of samples increases. Different from these papers, we focus on the non-asymptotic case and use the exact error probability as the performance metric to characterize the corresponding optimal attack and defense strategies.

We first focus on the hypothesis-aware scenario, in which the adversary knows which hypothesis is used to generate the data sample. The study of this powerful adversarial model can provide performance bounds for other attack models. Under this setting, we show that the formulated minimax problem has a saddle-point solution, which reveals the structures of the optimal attack and defense strategies. In this dissertation, we solve this problem for a special case where the optimal Bayesian decision regions corresponding to the PMFs before attack consist of two consecutive regions. Under this assumption, we first derive an upper-bound on the prediction error, which only depends on the PMFs before attack. Afterwards, we design a specific attack scheme and show that the designed attack scheme achieves the upper-bound. This implies that the specific attack scheme is optimal. We also note that the attack strategy that achieves the maximum error probability is not

unique.

We then study a more practical and challenging hypothesis-unaware scenario, where the attacker does not know the prior information about the underlying hypothesis. Despite the additional challenge, we show that the method developed for the hypothesis-aware case can be properly modified and extended to this scenario. In particular, following a similar saddle-point analysis, we reveal the structure of the optimal attack and defense strategy and convert the problem into a complicated non-convex optimization problem over the attack strategy. We then derive an upper-bound on the error probability and design a specific attack strategy to achieve the upper-bound.

The derived algorithms could potentially be useful for the quickest detection setup [129–143]. In particular, consider a system where an attacker appears at an unknown time, and we are interested in detecting the presence of attacks with minimum delay (under certain delay metric). The presence of the attacker is reflected on the change of the distribution of the data, and hence the quickest detection framework can be employed. Most of the existing works on quickest detection assume that post-change distribution is known. In the setup with an attacker, this assumption may not be practical. The algorithms developed in our work could be used to identify which distribution is most beneficial to the attacker and hence could be the most likely post-change distribution used by the attacker. This work has been published in [23, 144].

1.3.2 Privacy-Accuracy Trade-off

To analyze the privacy-protection, we address the fundamental trade-off between inference accuracy and privacy protection from information theory perspective.

There exist many privacy-preserving techniques that are based on perturbations of data, which provide privacy guarantees at the expense of a loss of accuracy [28–30]. For example, k-anonymity is proposed by Samarati and Sweeney [26], which requires that each record is indistinguishable from at least k–1 other records within the dataset. Differential privacy works by adding a predetermined amount of randomness into a computation performed on a data set [27]. These concepts and techniques are very useful for the privacy protection of data analysis through a dataset

or database, which is different from the setup considered in this dissertation. Moreover, various minimax formulations and algorithms have also been proposed to defend against inference attack in different scenarios [71–73]. Bertran et al. [71] proposed an optimization problem where the terms in the objective function were defined in terms of mutual information, showed the performance bound for the optimization problem and learned the sanitization transform in a data-driven fashion using an adversarial approach with Deep Neural Networks (DNNs). Under their formulation, they analyzed a trade-off between utility loss and attribute obfuscation under the constraint of the attribute obfuscation $I(A;Z) \leq k$. Feutry et al. [72] measured the utility and privacy by expected risks, formulated the utility-privacy trade-off as a min-diff-max optimization problem and proposed a learning-based and task-dependent approach to solve this problem, while only deterministic mechanisms are considered. To address this issue, a privacy-preserving adversarial network was proposed in [73] by employing adversarially-trained neural networks to implement randomized mechanisms and to perform a variational approximation of mutual information privacy.

Different from them, we propose a more general privacy metric and avoid the reliance on DNNs to derive the privacy-mapping. In our problem formulation, instead of using a specific privacy leakage measure, we propose a general framework to measure privacy leakage. The proposed privacy leakage metric is defined by a continuous function f with certain properties. Different choices of f lead to different privacy measures. For example, if f is chosen to be $\log$ function, the proposed privacy leakage metric is the same as mutual information, a widely used information leakage measure. Moreover, we introduce a parameter β to represent the relative weight between these two measures. Thus, the trade-off problem between privacy and accuracy can be solved through a maximization problem where the objective function is composed of a weighted sum of accuracy and privacy terms. To solve the maximization problem, if we optimize over the space of the privacy-preserving mapping directly, the formulated problem is a complicated non-concave problem with multiple constraints. Through various transformations and variable augmentations, we transform the optimization problem into a form that has three dominating arguments with cer-

tain nice concavity properties. In particular, if any two arguments are fixed, the problem is concave in the remaining argument. We then exploit this structure and design an algorithm with two nested loops to solve the optimization problem for general f by iterating between those three dominating arguments until reaching convergence. For the outer loop, we solve the optimization on the first dominating argument, for which we have a closed-form update formula. For the inner loop, using certain concavity properties of the objective function on the other two dominating arguments, we apply the Alternating Direction Method of Multipliers (ADMM) to solve the non-convex problem efficiently. Compared with solving the optimization problem using gradient ascent in the space of the privacy-preserving mapping directly, the proposed method does not need parameter tuning, converges much faster and finds solutions that have much better qualities. Moreover, we provide the convergence analysis of the proposed method. Since there are two nested loops in the proposed method, we first prove the convergence of the inner loop, which is the convergence proof of the ADMM process. Although there exists convergence proof for typical ADMM, it handles two-block separable problems only. In our case, the considered optimization problem is non-convex and multi-block with a non-separable structure. Hence, we come up with two proofs with different assumptions on f. Based on the convergence proof of the ADMM procedure, we further prove that the function value is non-decreasing between two iterations in the outer-loop. Then with a guarantee that the objective function is upper-bounded, the proposed algorithm is shown to converge. To further illustrate the proposed framework and algorithm, we also provide several examples by specializing f to particular function choices and provide numerical results.

This work has been published in [145, 146].

1.3.3 Robust and Fairness-aware regression

Fairness and robustness are critical elements of trustworthy artificial intelligence that need to be addressed together [147]. Firstly, in the field of adversarial training, several research works are proposed to interpret the accuracy/robustness disparity phenomenon and to mitigate the fairness issue [147–149]. For example, [148] presents an adversarially-trained neural network that is closer

to achieve some fairness measures than the standard model on the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset. Secondly, a class-wise loss reweighting method is shown to obtain more fair standard and robust classifiers [150]. Moreover, [151] and [152] argue that traditional notions of fairness are not sufficient when the model is vulnerable to adversarial attacks, investigate the class-wise robustness and propose methods to improve the robustness of the most vulnerable class, so as to obtain a fairer robust model.

In this dissertation, we focus on regression problems and design a fair regression model that is robust to adversarial attacks. In particular, we consider two increasingly complex attack models. We first consider a scenario where the adversary is able to add one carefully designed adversarial data point to the dataset. We then consider a more powerful adversary who can directly modify the existing data points in the feature matrix. Particularly, we consider a rank-one modification attack, where the attacker carefully designs a rank-one matrix and adds it to the existing data matrix.

To design the robust fairness-aware model, we formulate a game between a defender aiming to minimize the accuracy loss and bias, and an attacker aiming to maximize these objectives. To characterize both the prediction and fairness performance of a model, the objective function is selected to be a combination of prediction accuracy loss and group fairness gap. Since the goals of the adversary and the fairness-aware defender are opposite, a minimax framework is introduced to characterize the considered problem. By solving the minimax problem, the optimal adversary as well as the robust fair regression model can be derived.

To solve the problem, one major challenge is that the proposed minimax problem is nonsmooth nonconvex-nonconcave, which may not have a local saddle point in general [153]. Although there exist many iterative methods for finding stationary points or local optima of nonconvex-concave or nonconvex-nonconcave minimax problems [154–157], there are usually specific assumptions that are not satisfied in our proposed realistic problems. To solve the complicated minimax problems in hand, we carefully examine the underlying structure of the inner maximization problem and the outer minimization problem, and then exploit the identified structure to design efficient algorithms.

For the scenario where the adversary adds a poisoned data point into the dataset, when solving

the inner maximization problem, we deal with the non-smooth nature of the objective function and obtain a structure that characterizes the best adversary, which is a function of the regression coefficient $\boldsymbol{\beta}$ of the defense model. We then analyze the minimization problem by transforming it to four sub-problems where each sub-problem is a non-convex quadratic minimization problem with multiple quadratic constraints, which is usually NP hard [158, 159], and finding a global minimizer is very challenging. By exploring the underlying properties of a specific sub-problem, we investigate 8 different cases, and obtain a global minimizer to such sub-problem. Then the minimum point of the proposed four sub-problems, $\boldsymbol{\beta}^*_{rob}$, corresponds to the optimal robust fairness-aware model, and the best adversarial data sample is obtained by fitting $\boldsymbol{\beta}^*_{rob}$ to the derived optimal attack strategy. On both synthetic data and real-world datasets, numerical results illustrate that the proposed robust fairness-aware regression model has better performance than the unrobust fair model as well as the ordinary linear regression model in both prediction accuracy and group-based fairness.

For the rank-one attack scheme, we transform the maximization problem into a form with five arguments, four of which can be solved exactly. With this transformation, the original nonconvex-nonconcave minimax problem for two vectors can be converted into several weakly-convex-weakly-concave minimax problems for one vector and one scalar, which can be approximately solved using existing algorithms such as [160]. With the proposed algorithm, the optimal attack scheme of the adversary and the adversarially robust fairness-aware model can be obtained simultaneously. On two real-world datasets, numerical results illustrate that the performance of the adversarially robust model relies on the trade-off parameter between prediction accuracy and fairness guarantee. By properly choosing such parameter, the robust model can achieve desirable performance in both prediction accuracy and group-based fairness. On the other hand, for other fair regression models, at least one performance metric will be severely affected by the rank-one attack.

This work has been published in [161, 162].

Chapter 2

Adversarial Robustness of Hypothesis Testing

In this chapter, we investigate the adversarial robustness of hypothesis testing rules. In the considered model, after a sample is generated, it will be modified by an adversary before being observed by the decision maker. The decision maker needs to decide the underlying hypothesis that generates the sample from the adversarially-modified data. We formulate this problem as a minimax hypothesis testing problem, in which the goal of the adversary is to design attack strategy to maximize the error probability while the decision maker aims to design decision rules so as to minimize the error probability. We consider both hypothesis-aware case, in which the attacker knows the true underlying hypothesis, and hypothesis-unaware case, in which the attacker does not know the true underlying hypothesis.

Particularly, in Section 2.1, we present our problem formulation. In Section 2.2, we depict the optimal solution for the hypothesis-aware setting. In Section 2.3, we focus on the hypothesis-unaware case. In Section 2.4, we provide numerical examples to illustrate the analytical results. In Section 2.5, we offer concluding remarks.

2.1 Problem Formulation

Suppose there is a discrete random variable X defined on a finite set $\mathcal{X} = \{x_1, x_1, ..., x_n\}$. Consider the binary hypothesis testing problem:

$$\mathcal{H}_0 : X \sim \boldsymbol{p}_0, \mathcal{H}_1 : X \sim \boldsymbol{p}_1,$$

in which $\boldsymbol{p}_0$ is a $1 \times n$ PMF vector with $p_{0,j} = \Pr(X = x_j|\mathcal{H}_0)$. $\boldsymbol{p}_1$ is defined in a similar manner. Here, $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are assumed to be known to both the adversary and the decision maker.

In this chapter, we focus on adversary hypothesis testing problem. In the considered model, after a sample is generated, an adversary can modify it to another value. The decision maker then observes the corrupted data. We consider two different adversary models with different capabilities.

2.1.1 Hypothesis-aware adversary

We first consider a powerful hypothesis-aware adversary, who knows the true underlying hypothesis with which the sample is generated. The study of this worst-case scenario will provide performance limits of other adversary models. In the considered model, the attacker can conduct randomized attacks. In particular, after observing sample $X = x_i$, the adversary can change it to an attacked sample $X' = x_j$ with a certain probability, where X' is also a random variable defined on $\mathcal{X}$. Since the adversary knows the true underlying hypothesis, different attack rules can be applied depending on whether the true hypothesis is $\mathcal{H}_0$ or $\mathcal{H}_1$. We denote the attack strategy of the attacker as $(\boldsymbol{A}, \boldsymbol{B})$, in which the components of $\boldsymbol{A}$ are $A_{i,j} = \Pr(X' = x_j|X = x_i, \mathcal{H}_0)$ and the components of $\boldsymbol{B}$ are $B_{i,j} = \Pr(X' = x_j|X = x_i, \mathcal{H}_1)$.

Motivated by adversarial example phenomena studied in deep neural networks, we assume that the change introduced by the adversary has limited amplitude. In particular, an adversarial example is data that has been modified by the attacker to fool the classifier. However, to avoid human eye detection, the amplitude of these modifications should be limited so that they are not perceptible to human eyes [34–40,43,44,47,163,164]. Formally, we assume $A_{i,j} = B_{i,j} = 0$ when

$|i-j| > \delta$, in which δ denotes the largest change allowed. We will use $\mathcal{A}, \mathcal{B}$ to denote the whole sets of all amplitude-constrained attackers under $\mathcal{H}_0, \mathcal{H}_1$ correspondingly. For any given attack rule $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}$, the PMF of X' can be written as $\boldsymbol{q}_0 = \boldsymbol{p}_0\boldsymbol{A}$ under $\mathcal{H}_0$ and $\boldsymbol{q}_1 = \boldsymbol{p}_1\boldsymbol{B}$ under $\mathcal{H}_1$, with $q_{k,j} = \Pr(X' = x_j|\mathcal{H}_k)$, with $k = 0, 1$.

Let $\mathcal{T} = [0,1]^n$ be the set of all decision rules. Denote $\boldsymbol{t} = [t_1, \cdots, t_n] \in \mathcal{T}$ as a randomized decision rule such that if $X = x_i$, the detector selects $\mathcal{H}_1$ with probability t_i, where $0 \leq t_i \leq 1$. For decision rule $\boldsymbol{t}$, the probability of false alarm and miss detection are

$$P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}) = \boldsymbol{p}_0\boldsymbol{A}\boldsymbol{t}^T, P_M(\boldsymbol{p}_1, \boldsymbol{B}, \boldsymbol{t}) = \boldsymbol{p}_1\boldsymbol{B}(\boldsymbol{1} - \boldsymbol{t})^T. \tag{2.1}$$

Assuming that the prior probability of two hypotheses are equal, i.e., $\Pr(\mathcal{H}_0) = \Pr(\mathcal{H}_1)$, the error probability P_E can be written as

$$P_E(\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}) = \frac{1}{2}[P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}) + P_M(\boldsymbol{p}_1, \boldsymbol{B}, \boldsymbol{t})]. \tag{2.2}$$

In the following, to simplify the notation, we will drop $\boldsymbol{p}_0, \boldsymbol{p}_1$ from the expression of P_E and will simply write it as $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t})$.

The goal of the attacker is to choose the attack rule $(\boldsymbol{A}, \boldsymbol{B})$ to maximize the error probability (2.2), while the goal of the defender is to choose the decision rule $\boldsymbol{t}$ to minimize the error probability (2.2). In this chapter, we seek to characterize the optimal $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ and $\boldsymbol{t}^*$ by solving the minimax problem

$$\min_{\boldsymbol{t}\in\mathcal{T}} \max_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{A}\times\mathcal{B}} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}). \tag{2.3}$$

2.1.2 Hypothesis-unaware adversary

We also consider a more practical scenario, in which the attacker does not know the true underlying hypothesis when it sees a sample. In this hypothesis-unaware adversary case, there is only one attack matrix $\mathbf{A}$, with $A_{i,j} = \Pr(X' = x_j|X = x_i)$ being the probability that the attacker will change x_i to x_j.

Correspondingly, for a decision rule $\boldsymbol{t}$, the probability of false alarm and miss detection are

$$P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}) = \boldsymbol{p}_0 \boldsymbol{A} \boldsymbol{t}^T, P_M(\boldsymbol{p}_1, \boldsymbol{A}, \boldsymbol{t}) = \boldsymbol{p}_1 \boldsymbol{A}(\mathbf{1} - \boldsymbol{t})^T. \tag{2.4}$$

And the error probability P_E can be written as

$$P_E(\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{A}, \boldsymbol{t}) = \frac{1}{2}[P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}) + P_M(\boldsymbol{p}_1, \boldsymbol{A}, \boldsymbol{t})]. \tag{2.5}$$

Similarly, we will drop $\boldsymbol{p}_0, \boldsymbol{p}_1$ from the expression of P_E and will simply write it as $P_E(\boldsymbol{A}, \boldsymbol{t})$.

Moreover, we seek to characterize the optimal $\boldsymbol{A}^*$ and $\boldsymbol{t}^*$ by solving the minimax problem

$$\min_{\boldsymbol{t} \in \mathcal{T}} \max_{\boldsymbol{A} \in \mathcal{A}} P_E(\boldsymbol{A}, \boldsymbol{t}). \tag{2.6}$$

In the problem formulations (4.2) and (2.6) discussed above, the distributions under $\mathcal{H}_0$ and $\mathcal{H}_1$, i.e. $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$, are known to the attacker and decision maker. These problem formulations can be generalized to the scenario where there are uncertainties about the distributions. Suppose the actual distribution $\boldsymbol{p}_t, t = 0, 1$ under $\mathcal{H}_t$ belongs to the neighborhood of a nominal distribution. The neighborhood, denoted by $\mathcal{P}_t$ can be defined by KL-divergence [57], α-divergence [53], etc. The optimal $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ and $\boldsymbol{t}^*$ for the hypothesis-aware case can be found by solving the complex optimization problem

$$\min_{\boldsymbol{t} \in \mathcal{T}} \max_{(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}} \min_{(\boldsymbol{p}_0, \boldsymbol{p}_1) \in \mathcal{P}_0 \times \mathcal{P}_1} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}, \boldsymbol{p}_0, \boldsymbol{p}_1).$$

Similarly, the optimal $\boldsymbol{A}^*$ and $\boldsymbol{t}^*$ for the hypothesis-unaware case can be found by solving the optimization problem

$$\min_{\boldsymbol{t} \in \mathcal{T}} \max_{\boldsymbol{A} \in \mathcal{A}} \min_{(\boldsymbol{p}_0, \boldsymbol{p}_1) \in \mathcal{P}_0 \times \mathcal{P}_1} P_E(\boldsymbol{A}, \boldsymbol{t}, \boldsymbol{p}_0, \boldsymbol{p}_1).$$

These problem formulations are much more complex than (4.2) and (2.6), and are left as future work.

2.2 Optimal hypothesis-aware adversary

In this section, we focus on the hypothesis-aware case and characterize the optimal solution to the complicated minimax optimization problem (4.2). To achieve this, we will first conduct a saddle-point analysis to reveal the structure of the optimal solution. Based on this, we will derive an upper-bound on the error probability. We will then develop an attack strategy to achieve this bound.

2.2.1 Saddle-point Analysis

In this subsection, we characterize the structure of the optimal decision rules by analyzing the saddle-point of the minimax problem (4.2).

Note that, given $\boldsymbol{t}$, $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t})$ is continuous and linear, and therefore is both convex and concave in $(\boldsymbol{A}, \boldsymbol{B})$. Similarly, given $(\boldsymbol{A}, \boldsymbol{B})$, $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t})$ is continuous and linear, and therefore is both convex and concave in $\boldsymbol{t}$. Furthermore, sets $\mathcal{A} \times \mathcal{B}$ and $\mathcal{T}$ are both compact and convex. Therefore, using Von Neumann minimax theorem [165] (which allows the swapping of the $\min$ and $\max$ operators under certain conditions), we have

$$\min_{\boldsymbol{t}\in\mathcal{T}} \max_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{A}\times\mathcal{B}} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}) = \max_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{A}\times\mathcal{B}} \min_{\boldsymbol{t}\in\mathcal{T}} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}). \tag{2.7}$$

This implies that the solution $(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{t}^*)$ to this minimax problem satisfies the saddle-point property

$$P_E(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{t}) \geq P_E(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{t}^*) \geq P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}^*). \tag{2.8}$$

From these two inequalities, we can characterize the structure of the optimal attack and decision strategies.

The first inequality in (2.8) indicates that the best decision rule must be the Bayesian test with respect to the best adversary $(\boldsymbol{A}^*, \boldsymbol{B}^*)$. It is well known that, for a given arbitrary adversary attack

rule $(\boldsymbol{A}, \boldsymbol{B})$, the optimal detection rule, denoted as $\boldsymbol{t}^*(\boldsymbol{A}, \boldsymbol{B})$, is simply a threshold rule

$$t_i^*(\boldsymbol{A}, \boldsymbol{B}) = \begin{cases} 0 & q_{0,i} > q_{1,i}, \\ \text{arbitrary} & q_{0,i} = q_{1,i}, \\ 1 & q_{0,i} < q_{1,i}, \end{cases} \tag{2.9}$$

where $\boldsymbol{q}_0 = \boldsymbol{p}_0\boldsymbol{A}$, $\boldsymbol{q}_1 = \boldsymbol{p}_1\boldsymbol{B}$. For the optimal adversary $(\boldsymbol{A}^*, \boldsymbol{B}^*)$, the optimal decision rule is $\boldsymbol{t}^* = \boldsymbol{t}^*(\boldsymbol{A^*}, \boldsymbol{B^*})$.

With the optimal form of t^* in terms of $(\boldsymbol{A}, \boldsymbol{B})$ characterized in (2.9), we can then use the second inequality in (2.8) to characterize the optimal $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ by solving

$$\max_{\boldsymbol{A},\boldsymbol{B}} \quad \frac{1}{2}[\boldsymbol{p}_0\boldsymbol{A}(\boldsymbol{t}^*(\boldsymbol{A}, \boldsymbol{B}))^T + \boldsymbol{p}_1\boldsymbol{B}(1 - (\boldsymbol{t}^*(\boldsymbol{A}, \boldsymbol{B}))^T)], \tag{2.10}$$

$$\text{s.t.} \quad A_{i,j} \geq 0, B_{i,j} \geq 0, i, j = 1, .., n, \tag{2.11}$$

$$\sum_{j=1}^{n} A_{i,j} = 1, \sum_{j=1}^{n} B_{i,j} = 1, i = 1, .., n, \tag{2.12}$$

$$\mathbf{1}_{|i-j|>\delta} A_{i,j} = \mathbf{1}_{|i-j|>\delta} B_{i,j} = 0, i, j = 1, .., n, \tag{2.13}$$

in which $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Here, constraints (2.11) and (2.12) guarantee that each row of $\boldsymbol{A}$ and $\boldsymbol{B}$ is a conditional PMF, while constraint (2.13) makes sure that the changes introduced by the attacker has a limited amplitude.

Once we solve (2.10) and obtain $(\boldsymbol{A}^*, \boldsymbol{B}^*)$, the optimal $t^*(\boldsymbol{A}^*, \boldsymbol{B}^*)$ can be obtained by using (2.9). Due to the decision rule in (2.9), the objective function in (2.10) is a complicated function of $(\boldsymbol{A}, \boldsymbol{B})$. In the following, we will characterize the optimal solution to this challenging optimization problem under the following assumptions on $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$. Let $R_0 = \{i | p_{0,i} \geq p_{1,i}\}$ and $R_1 = \{i | p_{0,i} < p_{1,i}\}$. We will assume that R_0 (and hence R_1) is a consecutive region in $[1, n]$. Without loss of generality, we write $R_0 = \{i | 1 \leq i \leq m\}$ and $R_1 = \{i | m + 1 \leq i \leq n\}$.

We now compare this assumption with the assumptions used in the study of classic robust hypothesis testing [57], in which the nominal PMF is assumed to satisfy certain monotonicity

and symmetry properties. Specifically, in [57], monotonicity means that $\frac{p_{1,i}}{p_{0,i}}$ is a monotonically increasing function of i and symmetry means $p_{1,n-i+1} = p_{0,i}, 1 \leq i \leq n$. It is easy to check that the monotonicity assumption implies the assumption made in this chapter. Moreover, our assumption does not require the symmetry condition. Hence, our assumption is significantly weaker than the assumptions in [57].

2.2.2 Upper-bound for P_E

In this section, we develop an upper-bound on the objective function (2.10) that holds for any attack strategy.

We first present a lemma that simplifies $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}^*)$ into two equivalent forms, both of which will be used in the sequel.

Lemma 1. *$P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}^*)$ can be written as*

$$P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}^*) = \frac{1}{2} - \frac{1}{4}\sum_{i=1}^{n} |q_{0,i} - q_{1,i}| \tag{2.14}$$

$$= \frac{1}{2}\sum_{i=1}^{n} \min\{q_{0,i}, q_{1,i}\}. \tag{2.15}$$

Proof. Please see Appendix A.1. □

From (2.14), we can see that the most powerful attacker is the one that minimizes the ℓ_1 distance between $\boldsymbol{q}_0$ and $\boldsymbol{q}_1$, which inspires us to optimize the error probability by components.

To proceed further, we denote the mass moved into $[1, i]$ as $I_{t,i}$ for $t = 0$ (i.e., under hypothesis $\mathcal{H}_0$) and $t = 1$ (i.e., under hypothesis $\mathcal{H}_1$) respectively. Similarly, define the mass moved out from $[1, i]$ as $K_{t,i}$. For example, for region $[1, m]$, we have

$$I_{1,m} = \sum_{j=m+1}^{m+\delta} p_{1,j} \left(\sum_{i=j-\delta}^{m} B_{j,i} \right),$$

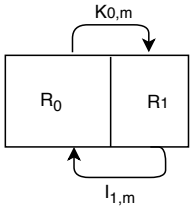


Figure 2.1: Mass moved between two regions

$$K_{0,m} = \sum_{j=m+1-\delta}^{m} p_{0,j} \left(\sum_{i=m+1}^{j+\delta} A_{j,i} \right),$$

as shown in Fig. 2.1.

Define

$$F_0 = F_0(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{n} q_{0,i}, F_j(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{j} \min\{q_{0,i}, q_{1,i}\} + \sum_{i=j+1}^{n} q_{0,i}.$$

Then we can see that $F_{j+1}(\boldsymbol{A}, \boldsymbol{B}) = F_j(\boldsymbol{A}, \boldsymbol{B}) + \min\{q_{1,j+1} - q_{0,j+1}, 0\}$, and thus $2P_E(\boldsymbol{A}, \boldsymbol{B}) = F_n(\boldsymbol{A}, \boldsymbol{B}) \leq ... \leq F_m(\boldsymbol{A}, \boldsymbol{B}) \leq ... \leq F_0$.

We are now ready to derive an upper-bound on the error probability P_E that holds for any attack strategy $(\boldsymbol{A}, \boldsymbol{B})$.

Theorem 2. *For* $\forall(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}$*,*

$$F_m(\boldsymbol{A}, \boldsymbol{B}) \leq \min\left\{1, \min_{1+\delta \leq j \leq m} \{G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}\right\}, \tag{2.16}$$

$$2P_E = F_n(\boldsymbol{A}, \boldsymbol{B}) \leq \min\left\{1, \min_{1+\delta \leq j \leq n} \{G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}\right\}, \tag{2.17}$$

in which

$$G_j(\boldsymbol{p}_0, \boldsymbol{p}_1) = 1 - \sum_{i=1}^{j-\delta} p_{0,i} + \sum_{i=1}^{\min\{n, j+\delta\}} p_{1,i}. \tag{2.18}$$

Furthermore, for $j^* = \arg\min_{1+\delta \leq j \leq n} \{G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}$*, if* $G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq 1$*, the equality in* (2.17) *holds when there exists* $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}$ *such that:*

(i) $q_{1,i} \leq q_{0,i}, 1 \leq i \leq j^*$;

(ii)

$$K_{0,j^*} - I_{0,j^*} = \sum_{i=j^*-\delta+1}^{j^*} p_{0,i}, \tag{2.19}$$

$$I_{1,j^*} - K_{1,j^*} = \sum_{i=j^*+1}^{\min\{n,j^*+\delta\}} p_{1,i}; \tag{2.20}$$

(iii) $F_k(\boldsymbol{A}, \boldsymbol{B}) = F_{j^*}(\boldsymbol{A}, \boldsymbol{B}), j^* < k \leq n$.

If $G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) > 1$*, the equality is achieved when*

$$F_i(\boldsymbol{A}, \boldsymbol{B}) = 1, 1 \leq i \leq n. \tag{2.21}$$

Proof. Please see Appendix A.2. □

We note that the bound in Theorem 2 depends only on $(\boldsymbol{p}_0, \boldsymbol{p}_1)$, the original PMFs before attack.

2.2.3 Optimal Adversary Design

In this section, we design the attack matrix $(\boldsymbol{A}, \boldsymbol{B})$ to achieve the upper-bound in (2.17). As the designed attack matrix achieves the upper-bound, it is an optimal solution to (2.10).

The construction process is motivated by the form in (2.14), which shows that the component-wise absolute difference (ℓ_1 distance) between $\boldsymbol{q}_0$ and $\boldsymbol{q}_1$ needs to be minimized. To minimize the ℓ_1 distance, we find the optimal $(\boldsymbol{A}, \boldsymbol{B})$ column by column. In particular, at the first step, we determine $\boldsymbol{A}_{:,1}, \boldsymbol{B}_{:,1}$ (based on some criteria to be detailed in the sequel). Once $\boldsymbol{A}_{:,1}, \boldsymbol{B}_{:,1}$ are determined, $q_{t,1}$ and F_1 are also determined. We denote these values as $\hat{q}_{t,1}$ and $\hat{F}_1$ respectively. We also have the constrained attack set $\mathcal{A}_1 \times \mathcal{B}_1 = \{(\boldsymbol{A}, \boldsymbol{B}) | \hat{q}_{0,1} \text{ and } \hat{q}_{1,1} \text{ are obtained}\}$. After step $j-1$, the first $j-1$ columns have been determined, and the constrained set is $\mathcal{A}_{j-1} \times \mathcal{B}_{j-1}$. Then

at step j, among all valid attack matrices in $\mathcal{A}_{j-1} \times \mathcal{B}_{j-1}$, we determine $\boldsymbol{A}_{:,j}$, $\boldsymbol{B}_{:,j}$ (based on a process to be detailed in the sequel) and obtain $\hat{q}_{t,j}, \hat{F}_j$. The constrained set is further refined to be $\mathcal{A}_j \times \mathcal{B}_j = \{(\boldsymbol{A}, \boldsymbol{B}) | \hat{q}_{0,j} \text{ and } \hat{q}_{1,j} \text{ are obtained}\} \subset \mathcal{A}_{j-1} \times \mathcal{B}_{j-1}$. The process ends at step n.

In the following, we describe our design of $(\boldsymbol{A}, \boldsymbol{B})$ to achieve the upper-bound. We will first focus on $1 \leq j \leq m$, i.e., $j \in R_0$, to obtain the equality in (2.16). Then focus on $m+1 \leq j \leq n$, i.e., $j \in R_1$, to achieve the equality in (2.17).

Column design for $j \in R_0$:

In R_0, we design the columns of $(\boldsymbol{A}, \boldsymbol{B})$ to satisfy

1) $\hat{q}_{1,1} = \sum_{i=1}^{1+\delta} p_{1,i}$;

2) $\hat{q}_{1,j} = p_{1,j+\delta}, 2 \leq j \leq m$;

3) $\hat{q}_{0,j} = \hat{q}_{1,j}, 1 \leq j \leq \delta$;

4) $\hat{q}_{0,j} = \max\{p_{0,j-\delta} A_{j-\delta,j}, \hat{q}_{1,j}\}, \delta+1 \leq j \leq m$,

which will then be shown to achieve the optimal value of $F_m(\boldsymbol{A}, \boldsymbol{B})$ in (2.16). These conditions are also listed in Table 2.1.

	$j=1$	$2 \leq j \leq \delta$	$\delta+1 \leq j \leq m$
$\mathcal{H}_0 : \hat{q}_{0,j}$	$\sum_{i=1}^{1+\delta} p_{1,i}$	$p_{1,j+\delta}$	$\max\{p_{0,j-\delta} A_{j-\delta,j}, p_{1,j+\delta}\}$
$\mathcal{H}_1 : \hat{q}_{1,j}$	$\sum_{i=1}^{1+\delta} p_{1,i}$	$p_{1,j+\delta}$	$p_{1,j+\delta}$

Table 2.1: PMF design in R_0

First, we specify how to design each element of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ so that $\hat{q}_{0,j}$s and $\hat{q}_{1,j}$s are set to be these values.

For the first step, by $1), 3)$, we have $\hat{q}_{0,1} = \hat{q}_{1,1} = \sum_{i=1}^{1+\delta} p_{1,i}$, and thus $\hat{F}_1 = F_0 = 1$. Moreover,

we can achieve this by setting the first column of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ as

$$\begin{aligned}
&\hat{A}_{1,1} = \min\left\{1, \frac{\hat{q}_{0,1}}{p_{0,1}}\right\}, \\
&\hat{A}_{i,1} = \min\left\{1, \frac{\max\left\{0, \hat{q}_{0,1} - \sum_{k=1}^{i-1} p_{0,k}\right\}}{p_{0,i}}\right\}, 2 \le i \le n, \\
&\hat{B}_{i,1} = 1, 1 \le i \le 1+\delta, B_{i,1} = 0, 2+\delta \le i \le n.
\end{aligned}$$

We continue to next columns. For columns $2 \le j \le \delta$, from 2) and 3), we have $\hat{q}_{0,j} = \hat{q}_{1,j} = p_{1,j+\delta}$, then $\hat{F}_j = \hat{F}_{j-1}$. We can achieve this by setting the j-th columns of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ as

$$\forall 1 \le i \le n, \hat{A}_{i,j} = \min\left\{1 - \sum_{k=1}^{j-1} \hat{A}_{i,k}, \frac{\max\left\{\hat{q}_{0,j} - \sum_{k=1}^{i-1} p_{0,k}\left(1 - \sum_{t=1}^{j-1} \hat{A}_{k,t}\right), 0\right\}}{p_{0,i}}\right\}, \tag{2.22}$$

$$\hat{B}_{j+\delta,j} = 1, \hat{B}_{i,j} = 0, \forall i \ne j+\delta. \tag{2.23}$$

For columns $\delta + 1 \le j \le m$, $\hat{\boldsymbol{A}}_{:,j}$ and $\hat{\boldsymbol{B}}_{:,j}$ are also designed by (2.22) and (2.23).

In Appendix A.3, we show that, with this design of $\hat{\boldsymbol{A}}_{:,j}$ and $\hat{\boldsymbol{B}}_{:,j}$, the requirements in $1), 2), 3), 4)$ are satisfied.

Remark 1. *The column design for $\boldsymbol{A}$ in* (2.22) *indicates that*

$$\forall 1 \le i \le n, \hat{A}_{i,j} = \min\left\{A_{i,j}^{(1)}, \max\left\{A_{i,j}^{(2)}, A_{i,j}^{(3)}\right\}\right\},$$

in which

$$\begin{aligned}
A_{i,j}^{(1)} &= 1 - \sum_{k=1}^{j-1} \hat{A}_{i,k} = \max_{\boldsymbol{A} \in \mathcal{A}_{j-1}} A_{i,j}, \\
A_{i,j}^{(2)} &= \frac{\hat{q}_{0,j} - \sum_{k=1}^{i-1} p_{0,k}\left(1 - \sum_{t=1}^{j-1} \hat{A}_{k,t}\right)}{p_{0,i}}, \\
A_{i,j}^{(3)} &= 0 = \min_{\boldsymbol{A} \in \mathcal{A}_{j-1}} A_{i,j}.
\end{aligned}$$

For a given component j, looking at i which starts from 1 and goes to n, we notice that the value of $\hat{A}_{i,j}$ will goes from $A^{(1)}_{i,j}$, to $A^{(2)}_{i,j}$ and then $A^{(3)}_{i,j}$.

Second, we show that $\hat{F}_m$ achieves the equality in (2.16) by checking the values of $\hat{F}_j$ one by one from $j = \delta + 1$ to $j = m$. We have three cases that will occur in order as j increases.

Case 1: $\hat{q}_{0,j} = \hat{q}_{1,j}$, then $\hat{F}_j = \hat{F}_{j-1}$.

Case 2: j is the first component such that $\hat{q}_{0,j} > \hat{q}_{1,j}$, or equivalently, j is the smallest component satisfying

$$\sum_{i=1}^{j} \hat{q}_{0,i} > \sum_{i=1}^{j} \hat{q}_{1,i} = \sum_{i=1}^{j+\delta} p_{1,i}.$$

This means that we have $\hat{F}_{j-1} = \hat{F}_{j-2} = ... = 1$. As for $\hat{q}_{0,j}$, if $\hat{q}_{0,j} \neq \hat{q}_{1,j}$, then $\hat{q}_{0,j} = p_{0,j-\delta}\hat{A}_{j-\delta,j} > 0$ and thus

$$\sum_{i=1}^{j} \hat{q}_{0,i} \overset{(a)}{=} \sum_{i=1}^{j-\delta} p_{0,i} > \sum_{i=1}^{j} \hat{q}_{1,i} = \sum_{i=1}^{j+\delta} p_{1,i}. \tag{2.24}$$

To derive (a), as discussed above, we have $\hat{A}_{j-\delta,j} > 0$, which indicates $\hat{A}_{j-\delta,j-1} \neq A^{(1)}_{j-\delta,j-1}$. Then $K_{0,j} = \sum_{i=j-\delta+1}^{j} p_{0,i}$ and (a) is true. Therefore,

$$\begin{aligned} \hat{F}_j &= 1 + \hat{q}_{1,j} - \hat{q}_{0,j} = 1 + \sum_{i=1}^{j} (\hat{q}_{1,j} - \hat{q}_{0,j}) \\ &= 1 + \sum_{i=1}^{j+\delta} p_{1,i} - \sum_{i=1}^{j-\delta} p_{0,i} = G_j(\boldsymbol{p}_0, \boldsymbol{p}_1). \end{aligned} \tag{2.25}$$

Case 3: Suppose k is the largest component with

$$\hat{F}_k = G_k(\boldsymbol{p}_0, \boldsymbol{p}_1) = 1 + \sum_{i=1}^{k+\delta} p_{1,i} - \sum_{i=1}^{k-\delta} p_{0,i}. \tag{2.26}$$

Similar to Case 2, we have

$$\sum_{i=k+1}^{j} \hat{q}_{0,i} = \sum_{i=k-\delta+1}^{j-\delta} p_{0,i} > \sum_{i=k+1}^{j} \hat{q}_{1,i} = \sum_{i=k+\delta+1}^{j+\delta} p_{1,i}. \tag{2.27}$$

Therefore,

$$\begin{aligned}
\hat{F}_i &= \hat{F}_k, k+1 \le i \le j-1, \\
\hat{F}_j &= \hat{F}_k + \sum_{i=k+1}^{j} (\hat{q}_{1,i} - \hat{q}_{0,i}) \\
&\overset{(b)}{=} \hat{F}_k + \sum_{i=k+\delta+1}^{j+\delta} p_{1,i} - \sum_{i=k-\delta+1}^{j-\delta} p_{0,i} \\
&\overset{(c)}{=} 1 + \sum_{i=1}^{j+\delta} p_{1,i} - \sum_{i=1}^{j-\delta} p_{0,i} = G_j(\boldsymbol{p}_0, \boldsymbol{p}_1),
\end{aligned} \tag{2.28}$$

where (b) is from (2.27) and (c) is true due to (2.26).

Taking all three cases into consideration, we have

$$\hat{F}_j = \min\left\{\hat{F}_{j-1}, G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\}, \tag{2.29}$$

and thus $\hat{F}_m = \min\left\{1, \min_{1\le j\le m} G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\}$, which achieves the equality in (2.16).

Column design for $j \in R_1$:

In R_1, we design the columns of $(\boldsymbol{A}, \boldsymbol{B})$ to satisfy

$$1) \quad \hat{q}_{0,j} = \max\left\{\min_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j}, \min\left\{\max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j}, \max_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j}\right\}\right\}, \tag{2.30}$$

$$2) \quad \hat{q}_{1,j} = \max\left\{\min_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j}, \min\left\{\hat{q}_{0,j}, \max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j}\right\}\right\}, \tag{2.31}$$

where

$$\begin{aligned}
&\min_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j} = p_{0,j-\delta}A_{j-\delta,j},\\
&\max_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j} = K_{0,j-1} - I_{0,j-1} + \sum_{i=j}^{j+\delta} p_{0,i},\\
&\max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j} = \sum_{i=j}^{j+\delta} p_{1,i} - I_{1,j-1} + K_{1,j-1},\\
&\min_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j} = p_{1,j-\delta}B_{j-\delta,j}.
\end{aligned}$$

First, we describe the construction of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$. Note that, the first m columns of $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ have already been selected in R_0. For columns from $m+1$ to n, $\hat{\boldsymbol{A}}$ is constructed by (2.22) and $\hat{\boldsymbol{B}}$ is constructed by

$$\hat{B}_{i,j} = \min\left\{1 - \sum_{k=1}^{j-1}\hat{B}_{i,k}, \frac{\hat{q}_{1,j} - \sum_{k=1}^{i-1} p_{1,k}\hat{B}_{k,j}}{p_{1,i}}\right\}. \tag{2.32}$$

In Appendix A.3, we show that such $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ design satisfies the conditions on $\hat{\boldsymbol{q}}_0, \hat{\boldsymbol{q}}_1$ in $1), 2)$.

Second, we verify that $\hat{F}_n$ achieves the equality in (2.17) if the conditions on $\hat{\boldsymbol{q}}_0, \hat{\boldsymbol{q}}_1$ in $1), 2)$ are satisfied. The main idea is to derive the value of $\hat{F}_j$ based on the value of $\hat{F}_{j-1}$ by calculating $\hat{q}_{0,j} - \hat{q}_{1,j}$. According to the previously designed columns, the relationship between $\hat{q}_{0,j}$ and $\hat{q}_{1,j}$ has three different cases.

Case 1: $\hat{q}_{0,j} \le \hat{q}_{1,j}$, then $\hat{F}_j = \hat{F}_{j-1}$. Moreover, we have $\hat{q}_{0,j} \neq p_{0,j-\delta}A_{j-\delta,j}$ in this case.

Case 2: Assume that j is the smallest component in R_1 with $\hat{q}_{0,j} > \hat{q}_{1,j}$. Specifically, by $1), 2)$, for this component, we have

$$\begin{aligned}
\hat{q}_{0,j} &= \min_{\boldsymbol{A}\in\mathcal{A}_{j-1}} q_{0,j} = p_{0,j-\delta}A_{j-\delta,j} = K_{0,j-1} - I_{0,j-1} - \sum_{i=j-\delta+1}^{j-1} p_{0,i},\\
\hat{q}_{1,j} &= \max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j} = \sum_{i=j}^{\min\{j+\delta,n\}} p_{1,i} - I_{1,j-1} + K_{1,j-1}.
\end{aligned}$$

Then

$$\begin{aligned}
\hat{q}_{0,j} - \hat{q}_{1,j} &= p_{0,j-\delta} A_{j-\delta,j} - \max_{\boldsymbol{B}\in\mathcal{B}_{j-1}} q_{1,j} \\
&= (K_{0,j-1} - I_{0,j-1} - \sum_{i=j-\delta+1}^{j-1} p_{0,i}) - \left(\sum_{i=j}^{\min\{j+\delta,n\}} p_{1,i} - I_{1,j-1} + K_{1,j-1} \right) \\
&\overset{(a)}{=} \hat{F}_{j-1} - 1 + \sum_{i=1}^{j-1}(p_{0,i} - p_{1,i}) - \sum_{i=j-\delta+1}^{j-1} p_{0,i} - \sum_{i=j}^{\min\{j+\delta,n\}} p_{1,i} \\
&= \hat{F}_{j-1} - G_j(\boldsymbol{p}_0, \boldsymbol{p}_1),
\end{aligned}$$

in which (a) comes from the following fact,

$$\begin{aligned}
F_j(\boldsymbol{A}, \boldsymbol{B}) &= 1 + \sum_{i=1}^{j} \min\{(q_{1,i} - q_{0,i}), 0\} \\
&\overset{(b)}{\leq} 1 + \sum_{i=1}^{j}(q_{1,i} - q_{0,i}) = 1 + \sum_{i=1}^{j} q_{1,i} - \sum_{i=1}^{j} q_{0,j} \\
&= 1 - \sum_{i=1}^{j}(p_{0,i} - p_{1,i}) + K_{0,j} - K_{1,j} + I_{1,j} - I_{0,j},
\end{aligned}$$

where the equality in (b) is attained when $q_{1,i} \leq q_{0,i}, 1 \leq i \leq j$. Recall that for $i \in R_0$, we have $\hat{q}_{0,i} \geq \hat{q}_{1,i}$. Then based on the assumption, we have $\hat{q}_{1,i} \leq \hat{q}_{0,i}, 1 \leq i \leq j$ and hence (a) is true.

Recall that $j^* = \arg\min_{1\leq j\leq n} \{G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}$ and we prove that $j^* \in R_1$ by contradiction. Suppose $j^* \in R_0$. Then note that $\hat{F}_m = G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq G_j(\boldsymbol{p}_0, \boldsymbol{p}_1), \forall j \in R_1$ and thus

$$\hat{q}_{0,j} - \hat{q}_{1,j} = \hat{F}_{j-1} - G_j(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq \hat{F}_m - G_j(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq 0,$$

which contradicts with the assumption that $\hat{q}_{0,j} > \hat{q}_{1,j}$. Hence, $j^* \in R_1$ and we have $\hat{F}_j = G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)$.

Case 3: For $k > j$ such that $\hat{q}_{0,j} = p_{0,j-\delta} A_{j-\delta,j} > \hat{q}_{1,j}$, by the similar idea of proving (2.28) in R_0, we will also have $\hat{F}_j = G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)$.

Taking all three cases into consideration, in R_1, (2.29) also holds, which indicates that the

equality in Theorem 2 is obtained for the designed adversary.

2.3 Optimal hypothesis-unaware adversary

In Section 2.2, we have considered a powerful hypothesis-aware adversary who knows the true underlying hypothesis before attack. In this section, we consider a more practical scenario with a hypothesis-unaware adversary who does not know the true underlying hypothesis that generates the observed data. In this section, we will investigate the optimal solution to the minimax problem characterized in (2.6).

Under the hypothesis-unaware setting, as the adversary has less information, the attack is more difficult to carry out. However, the approach in Section 2.2 can be modified and applied here.

First, the saddle point analysis in Section 2.2.1 can be easily extended to the hypothesis-unaware case to simplify (2.6). In particular, following a similar saddle-point analysis, for any given attack matrix $\boldsymbol{A}$, we have that the optimal form of the decision rule is

$$t_i^*(\boldsymbol{A}) = \begin{cases} 0 & q_{0,i} > q_{1,i}, \\ \text{arbitrary} & q_{0,i} = q_{1,i}, \\ 1 & q_{0,i} < q_{1,i}, \end{cases} \tag{2.33}$$

where $\boldsymbol{q}_0 = \boldsymbol{p}_0\boldsymbol{A}$, $\boldsymbol{q}_1 = \boldsymbol{p}_1\boldsymbol{A}$. The optimal attack matrix $\boldsymbol{A}^*$ is the solution to

$$\max_{\boldsymbol{A}} \quad \frac{1}{2}[\boldsymbol{p}_0\boldsymbol{A}(\boldsymbol{t}^*(\boldsymbol{A})^T + \boldsymbol{p}_1\boldsymbol{A}(1 - \boldsymbol{t}^*(\boldsymbol{A})].$$

This can be further rewritten as

$$\begin{aligned} \max_{\boldsymbol{A}} \quad & \frac{1}{2}[1+(\boldsymbol{p}_0-\boldsymbol{p}_1)\boldsymbol{A}\boldsymbol{t}^*(\boldsymbol{A})^T], \\ \text{subject to} \quad & A_{i,j}\geq 0, i,j=1,..,n, \\ & \sum_{j=1}^{n} A_{i,j}=1, \\ & 1_{|i-j|>\delta}A_{i,j}=0, i,j=1,..,n. \end{aligned} \tag{2.34}$$

In the following, we will generalize the approach in Section 2.2 to characterize the optimal solution to (2.34).

2.3.1 Upper-bound for P_E

Let $F_{m-\delta}(\boldsymbol{A})=\sum_{i=1}^{m-\delta} q_{1,i}+\sum_{i=m-\delta+1}^{n} q_{0,i}$ and

$$f(j,\boldsymbol{A})=\sum_{i=1}^{m-\delta} q_{1,i}+\sum_{i=m-\delta+1}^{j}\min\{q_{0,i},q_{1,i}\}+\sum_{i=j+1}^{n} q_{0,i}.$$

Define

$$F_j(\boldsymbol{A})=\begin{cases} F_{m-\delta}(\boldsymbol{A}) & 1\leq j\leq m-\delta, \\ f(j,\boldsymbol{A}) & m-\delta+1\leq j\leq m+\delta, \\ f(m+\delta,\boldsymbol{A}) & m+\delta+1\leq j\leq n. \end{cases}$$

Then from the definition, we have

$$F_{j+1}(\boldsymbol{A},\boldsymbol{B})=F_j(\boldsymbol{A},\boldsymbol{B})+\min\{q_{1,j+1}-q_{0,j+1},0\},$$

and thus

$$\begin{aligned} 2P_E(\boldsymbol{A},\boldsymbol{B}) &\overset{(a)}{=} F_n(\boldsymbol{A},\boldsymbol{B})=...=F_{m+\delta}(\boldsymbol{A},\boldsymbol{B})\leq ... \\ &\leq F_m(\boldsymbol{A},\boldsymbol{B})\leq ...\leq F_{m-\delta}=F_{m-\delta-1}=...=F_0, \end{aligned}$$

where (a) is due to the fact that

$$\begin{aligned}&\sum_{i=1}^{m-\delta} q_{1,i} + \sum_{i=m-\delta+1}^{m+\delta} \min\{q_{0,i}, q_{1,i}\} + \sum_{i=m+\delta+1}^{n} q_{0,i} \\ =& \sum_{i=1}^{n} \min\{q_{0,i}, q_{1,i}\}.\end{aligned}$$

Similar to Theorem 2, we have the following bound.

Theorem 3. *For $\forall \boldsymbol{A} \in \mathcal{A}$,*

$$F_m(\boldsymbol{A}) \leq \min_{m-\delta \leq j \leq m} \left\{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\}, \tag{2.35}$$

$$2P_E(\boldsymbol{A}) = F_{m+\delta}(\boldsymbol{A}) \leq \min_{m-\delta \leq j \leq m+\delta} \left\{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\} \coloneqq E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1), \tag{2.36}$$

in which

$$\begin{aligned}&E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1) = 1 - \sum_{i=1}^{m-2\delta} (p_{0,i} - p_{1,i}), \\ &E_j(\boldsymbol{p}_0, \boldsymbol{p}_1) = 1 - \sum_{i=1}^{j-\delta} (p_{0,i} - p_{1,i}) + \sum_{i=m+1}^{\min\{n, j+\delta\}} (p_{1,i} - p_{0,i}), \\ &j^* = \arg \min_{m-\delta \leq j \leq m+\delta} \left\{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\}.\end{aligned} \tag{2.37}$$

If $E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1)$, the equality in (2.36) *holds when there exists $\boldsymbol{A} \in \mathcal{A}$ such that:*

(i) $K_{0,m-\delta} - K_{1,m-\delta} = \sum_{i=m-2\delta+1}^{m-\delta} (p_{0,i} - p_{1,i})$;

(ii) $q_{1,i} \leq q_{0,i}, m - \delta + 1 \leq i \leq j^*$;

(iii) $K_{0,j^*} - K_{1,j^*} = \sum_{i=j^*-\delta+1}^{\min\{j^*, m\}} (p_{0,i} - p_{1,i})$,
$I_{1,j^*} - I_{0,j^*} = \sum_{i=\max\{m+1, j^*+1\}}^{\min\{n, j^*+\delta\}} (p_{1,i} - p_{0,i})$;

(iv) $F_k(\boldsymbol{A}) = F_{j^*}(\boldsymbol{A}), j^* < k \leq m + \delta$.

If $E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) > E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1)$, the equality is achieved when

$$F_i(\boldsymbol{A}) = E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1), m - \delta \leq i \leq m + \delta.$$

Proof. Please see Appendix A.4. □

2.3.2 Attack strategy design

In this section, we design an attack matrix $\hat{\boldsymbol{A}}$ to achieve the upper-bound in (2.36). As the designed matrix achieves the upper-bound, it is an optimal solution to (2.34). Similar to the design of hypothesis-aware attack matrix, we construct the optimal $\boldsymbol{A}$ column by column.

Before proceeding further, we need to define quantities related to mass moving between different regions. In particular, for $t = 0, 1$, define

- $a_{t,j}$: $[1, j-1] \to j$,
- $b_{t,j}$: $[1, j-1] \to [j+1, n]$,
- $c_{t,j}$: $j \to [j+1, n]$,
- $d_{t,j}$: $[j+1, n] \to j$,
- $e_{t,j}$: $[j+1, n] \to [1, j-1]$,
- $f_{t,j}$: $j \to [1, j-1]$.

These quantities are illustrated in Fig. 2.2.

Moreover, we will use $\hat{a}, \hat{b}, \hat{c}, \hat{d}, \hat{e}, \hat{f}$ to denote the value of a, b, c, d, e, f determined by $\hat{\boldsymbol{A}}$ while using $\hat{F}_j$ to denote the value of F_j achieved by $\hat{\boldsymbol{A}}$.

Column design for $j \in R_0$:

In R_0, for $t = 0, 1$, we design columns of attack matrix $\hat{\boldsymbol{A}}$ to achieve

(1) $\hat{q}_{t,j} = p_{t,j}, 1 \leq j \leq m - 2\delta$;

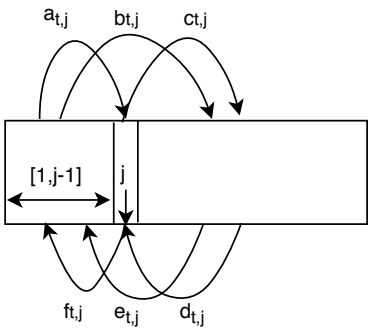


Figure 2.2: Moved mass between different regions at component j

(2) $\hat{q}_{t,j} = 0, m - 2\delta + 1 \leq j \leq m - \delta$;

(3) $\hat{q}_{t,j} = p_{t,j-\delta} + \hat{d}_{t,j}, m - \delta + 1 \leq j \leq m$, where $\hat{d}_{t,j}$ is selected to satisfy

$$\hat{d}_{1,j} - \hat{d}_{0,j} = \min\{p_{0,j-\delta} - p_{1,j-\delta}, \hat{F}_{m-\delta} - \hat{F}_{j-1} + \sum_{i=m+1}^{\min\{n,j+\delta\}} (p_{1,i} - p_{0,i}) + \sum_{i=m-2\delta+1}^{j-\delta-1} (p_{1,i} - p_{0,i})\}.$$

To summarize, these conditions are listed in Table 2.2.

	$\mathcal{H}_t : \hat{q}_{t,j}$
$1 \leq j \leq m - 2\delta$	$p_{t,j}$
$m - 2\delta + 1 \leq j \leq m - \delta$	0
$m - \delta + 1 \leq j \leq m$	$p_{t,j-\delta} + \hat{d}_{t,j}$

Table 2.2: PMF design in R_0 for the hypothesis-unaware adversary

Here, again, we will first describe how to design $\hat{\boldsymbol{A}}$ so that 1), 2) and 3) are satisfied. We will then show that, once these conditions are satisfied, the equality in (2.35) is achieved. Hence, the designed $\hat{\boldsymbol{A}}$ is optimal.

In particular, we set columns 1 to m of $\hat{\boldsymbol{A}}$ to be

a) $1 \leq j \leq m - 2\delta, \hat{A}_{j,j} = 1, \hat{A}_{i,j} = 0, i \neq j$;

b) $m - \delta + 1 \leq j \leq m, \hat{A}_{j-\delta,j} = 1$,
$\hat{A}_{i,j} = \min\left\{1, \max\left\{\frac{\hat{d}_{1,j} - \hat{d}_{0,j} - \sum_{k=m+1}^{i-1}(p_{1,k} - p_{0,k})}{p_{1,i} - p_{0,i}}, 0\right\}\right\}, m + 1 \leq i \leq n.$

Following the same proof in Appendix A.3, we can show that using design specified in a), b), the equalities in $1), 2), 3)$ are satisfied for $1 \leq j \leq m$. Details of the proofs are omitted for brevity.

We now verify that we can achieve the equality in the upper-bound (2.35) once conditions 1), 2) and 3) are satisfied.

$$\begin{aligned}\hat{F}_{m-\delta} &= \sum_{i=1}^{m-\delta}(p_{1,i}-p_{0,i}) + I_{1,m-\delta} - I_{0,m-\delta} - K_{1,m-\delta} + K_{0,m-\delta} + 1 \\ &= 1 - \sum_{i=1}^{m-2\delta}(p_{0,i}-p_{1,i}) \coloneqq E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1).\end{aligned}$$

For $\forall m-\delta \leq j \leq m$,

$$\begin{aligned}\hat{F}_j &= \hat{F}_{j-1} + \min\{0, \hat{q}_{1,j} - \hat{q}_{0,i}\} \\ &= \min\{\hat{F}_{j-1}, \hat{F}_{j-1} + \hat{q}_{1,j} - \hat{q}_{0,i}\} \\ &= \min\{\hat{F}_{j-1}, \hat{F}_{j-1} + p_{1,j-\delta} - p_{0,j-\delta} + \hat{d}_{1,j} - \hat{d}_{0,i}\} \\ &= \min\left\{\hat{F}_{j-1}, 1 - \sum_{i=1}^{j-\delta}(p_{0,i}-p_{1,i}) + \sum_{i=m+1}^{\min\{n,j+\delta\}}(p_{1,i}-p_{0,i})\right\} \\ &\coloneqq \min\left\{\hat{F}_{j-1}, E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\},\end{aligned}$$

and thus

$$\hat{F}_m = \min_{m-\delta \leq j \leq m}\left\{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\right\}, \tag{2.38}$$

which reaches the equality in (2.35).

Column design for $j \in R_1$:

In R_1, the first m columns of $\hat{\boldsymbol{A}}$ have been determined in R_0 and for $j \in R_1$, we further design $\boldsymbol{A}_{:,m+1:n}$ to achieve

$$\hat{q}_{1,j} - \hat{q}_{0,j} = \max\{\min_{\boldsymbol{A}\in\mathcal{A}_{j-1}}(q_{1,j}-q_{0,j}), \min\{0, \max_{\boldsymbol{A}\in\mathcal{A}_{j-1}}(q_{1,j}-q_{0,j})\}\}.$$

We will design $\boldsymbol{A}_{:,m+1:n}$ in two cases. For the first case, we always have $j^* \in R_0$ and $\boldsymbol{A}_{:,m+1:n}$ can be designed in a simple way. For the second case, similar procedure in Section 2.2.3 **Case 2** can be applied. In the following part, we will provide the assumptions of two cases and analyze the first scenario in detail while skip the details for the second scenario.

Case 1:

$$\min_{m-\delta\le j\le m-1}\{\sum_{i=j-\delta+1}^{m}(p_{0,i}-p_{1,i})-\sum_{i=j+\delta+1}^{m+\delta}(p_{1,i}-p_{0,i})\}\le 0.$$

By applying (2.38), this condition is equivalent to

$$\hat{F}_m \le 1+\sum_{i=1}^{m+\delta}(p_{1,i}-p_{0,1}),$$

and thus $\forall j \in [m+1, m+\delta]$,

$$E_j(\boldsymbol{p}_0,\boldsymbol{p}_1) = 1-\sum_{i=1}^{j-\delta}(p_{0,i}-p_{1,i})+\sum_{i=m+1}^{j+\delta}(p_{1,i}-p_{0,i})$$
$$\ge 1-\sum_{i=1}^{m}(p_{0,i}-p_{1,i})+\sum_{i=m+1}^{m+\delta}(p_{1,i}-p_{0,1}) \ge \hat{F}_m.$$

Therefore, we will be able to find an $\hat{\boldsymbol{A}} \in \mathcal{A}_m$, such that $\hat{F}_{m+\delta} = \hat{F}_m$.

The desired $\boldsymbol{A}_{:,m+1:n}$ is designed by

(1) $\forall m-\delta+1 \le j \le m, \hat{A}_{j,m+1} = 1$;

(2) $\forall m+1 \le j \le m+\delta, \hat{A}_{j,m+1} = 1-\sum_{i=1}^{m}\hat{A}_{j,i}$;

(3) $\forall m+\delta+1 \le j \le n, \hat{A}_{j,j} = 1$.

Then we have

$$\begin{aligned}
\hat{q}_{1,m+1} - \hat{q}_{0,m+1} &= \sum_{k=m-\delta+1}^{m+\delta+1} (p_{1,k} - p_{0,k}) \hat{A}_{k,m+1} \\
&= K_{1,m} - K_{0,m} + \sum_{k=m+1}^{m+\delta+1} (p_{1,k} - p_{0,k})(1 - \sum_{i=1}^{m} \hat{A}_{k,i}) \\
&= K_{1,m} - K_{0,m} + \sum_{k=m+1}^{m+\delta+1} (p_{1,k} - p_{0,k}) - I_{1,m} + K_{0,m} \\
&\overset{(a)}{=} 1 + \sum_{i=1}^{m+\delta} (p_{1,i} - p_{0,1}) - \hat{F}_m \geq 0,
\end{aligned}$$

where (a) is because $\forall m - \delta + 1 \leq j \leq m + \delta, \forall \boldsymbol{A} \in \mathcal{A}$,

$$\begin{aligned}
F_j(\boldsymbol{A}) &= F_{m-\delta}(\boldsymbol{A}) + \sum_{i=m-\delta+1}^{j} \min\{(q_{1,i} - q_{0,i}), 0\} \\
&\overset{(b)}{\leq} F_{m-\delta}(\boldsymbol{A}) + \sum_{i=m-\delta+1}^{j} (q_{1,i} - q_{0,i}) \\
&= \sum_{i=1}^{j} q_{1,i} + \sum_{i=j+1}^{n} q_{0,i} \\
&= 1 + \sum_{i=1}^{j} (q_{1,i} - q_{0,i}) \\
&= 1 - \sum_{i=1}^{j} (p_{0,i} - p_{1,i}) + K_{0,j} - K_{1,j} + I_{1,j} - I_{0,j},
\end{aligned}$$

and the equality in (b) holds when $q_{1,i} - q_{0,i} \leq 0, m - \delta + 1 \leq i \leq j$. For here, $j = m$ and we have $q_{1,i} - q_{0,i} \leq 0$ in R_0 and (a) is true.

Furthermore, we have $\hat{q}_{1,j} = \hat{q}_{0,j} = 0, m + 2 \leq j \leq m + \delta$. Therefore, for the designed $\hat{\boldsymbol{A}}$, we have

$$\hat{F}_{m+\delta}(\hat{\boldsymbol{A}}) = \hat{F}_{m+1}(\hat{\boldsymbol{A}}) = \hat{F}_m(\hat{\boldsymbol{A}}) + \min\{\hat{q}_{1,m+1} - \hat{q}_{0,m+1}, 0\} = \hat{F}_m(\hat{\boldsymbol{A}}),$$

and thus the equality in Theorem 3 is achieved.

Case 2:

$$\min_{m-\delta\leq j\leq m-1}\{\sum_{i=j-\delta+1}^{m}(p_{0,i}-p_{1,i})-\sum_{i=j+\delta+1}^{m+\delta}(p_{1,i}-p_{0,i})\}>0.$$

Under this condition, by the same idea in 2.2.3 **Case 2**, we will have $\hat{F}_j = \min\{\hat{F}_{j-1}, E_j(\boldsymbol{p}_0,\boldsymbol{p}_1)\}$. Therefore, the equality in Theorem 3 is attained.

2.4 Numerical Results

In this section, we provide numerical examples to illustrate results obtained in this chapter.

In the first example, we give two specific PMFs with a few components and perform hypothesis-aware and hypothesis-unaware attacks to show how the adversary works. In this example, the PMF before attack is provided in (2.39) and Fig. 2.3. It is easy to calculate that for this PMF, if there is no adversary, the error probability corresponding to the optimal Bayesian detection rule is $P_E = \frac{11}{32}$. Assume that the attack amplitude is $\delta = 1$. Following the design process in Section 2.2.3 and 2.3.2, the optimal hypothesis-aware attack strategy $\hat{\boldsymbol{A}}_a, \hat{\boldsymbol{B}}_a$ and the optimal hypothesis-unaware attack strategy $\hat{\boldsymbol{A}}_u$ are

$$\hat{\boldsymbol{A}}_a = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{3}{7} & \frac{4}{7} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

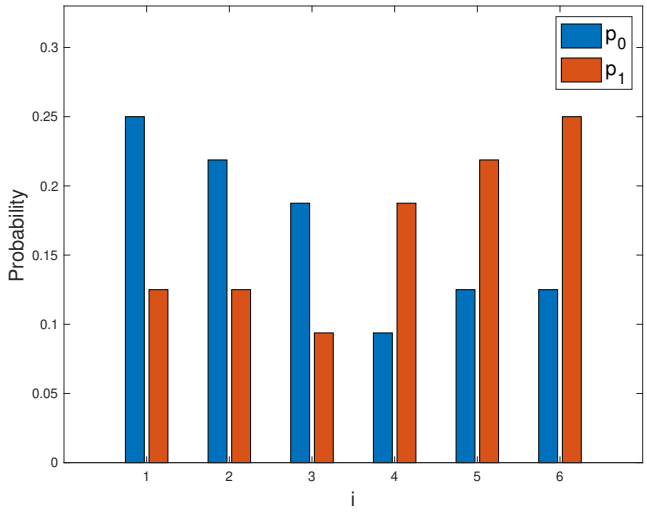


Figure 2.3: PMFs $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$

$$\hat{\boldsymbol{B}}_a = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \hat{\boldsymbol{A}}_u = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Thus, the PMFs after attack can be calculated and are provided in (2.40) and Fig. 2.4 for the hypothesis-aware model and the PMFs of hypothesis-unaware model are provided in (2.41) and Fig. 2.5. It is easy to check that, for the constructed adversary, the error probabilities are $P_E(\hat{\boldsymbol{A}}_a, \hat{\boldsymbol{B}}_a, t^*(\hat{\boldsymbol{A}}_a, \hat{\boldsymbol{B}}_a)) = \frac{1}{2}$ and $P_E(\hat{\boldsymbol{A}}_u, t^*(\hat{\boldsymbol{A}}_u)) = \frac{7}{16}$ correspondingly. Since the error probability is 1/2 (the largest possible value) for the hypothesis-aware attack, the designed attack matrix is clearly optimal. For the hypothesis-unaware attack, the error probability under $\hat{\boldsymbol{A}}_u$ is less than $\frac{1}{2}$. This already achieves the maximal value of $P_E(\boldsymbol{A}_u, t^*(\boldsymbol{A}_u))$ by Theorem 3, $2P_E(\boldsymbol{A}_u, t^*(\boldsymbol{A}_u)) \leq E_4(\boldsymbol{p}_0, \boldsymbol{p}_1) = \frac{7}{8}$. Therefore, for this particular example, the hypothesis-unaware attacker is not as

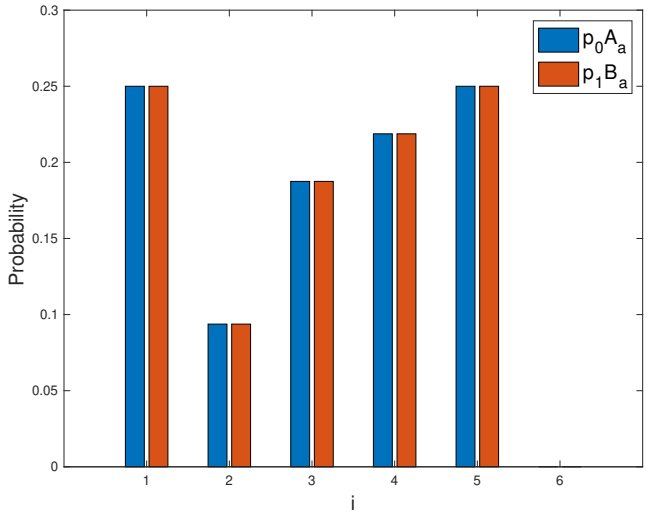


Figure 2.4: PMFs $\boldsymbol{p}_0\hat{\boldsymbol{A}}_a$ and $\boldsymbol{p}_1\hat{\boldsymbol{B}}_a$ for the hypothesis-aware case

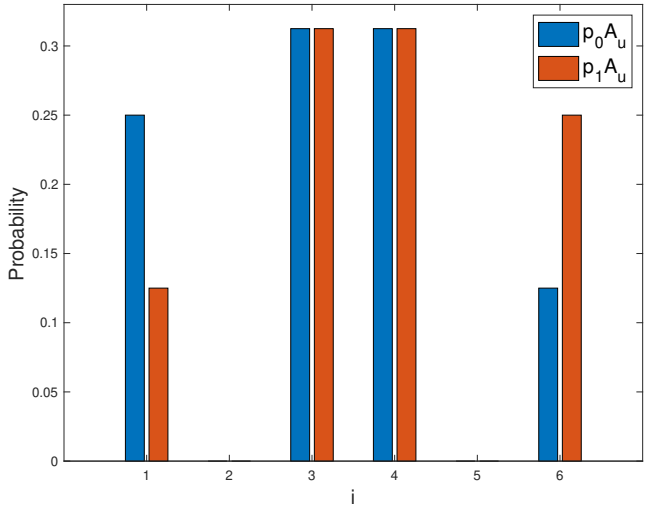


Figure 2.5: PMFs $\boldsymbol{p}_0\hat{\boldsymbol{A}}_u$ and $\boldsymbol{p}_1\hat{\boldsymbol{A}}_u$ for the hypothesis-unaware case

powerful as the hypothesis-aware attacker.

$$\begin{array}{lllllll} \boldsymbol{p}_0 & 8/32 & 7/32 & 6/32 & 3/32 & 4/32 & 4/32 \\ \boldsymbol{p}_1 & 4/32 & 4/32 & 3/32 & 6/32 & 7/32 & 8/32 \end{array} \tag{2.39}$$

$$\begin{array}{lllllll} \boldsymbol{p}_0\hat{\boldsymbol{A}}_a & 8/32 & 3/32 & 6/32 & 7/32 & 8/32 & 0 \\ \boldsymbol{p}_1\hat{\boldsymbol{B}}_a & 8/32 & 3/32 & 6/32 & 7/32 & 8/32 & 0 \end{array} \tag{2.40}$$

$$\begin{array}{lllllll} \boldsymbol{p}_0\hat{\boldsymbol{A}}_u & 8/32 & 0 & 10/32 & 10/32 & 0 & 4/32 \\ \boldsymbol{p}_1\hat{\boldsymbol{A}}_u & 4/32 & 0 & 10/32 & 10/32 & 0 & 8/32 \end{array} \tag{2.41}$$

In the second example, we explore how δ affects the prediction error for a randomly selected $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ under two attack models. In our simulation, we generate $2n$ random numbers in $[0, 1]$ by uniform distribution, divide them into two sequences and normalize each sequence to make it a PMF while maintaining two consecutive regions to meet the assumption made in Section 2.2.1. After $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are generated, they are fixed throughout the experiment. We then apply the proposed attack schemes to find one of the optimal attackers and calculate its prediction error under the Bayesian test. The results are shown in Fig. 2.6, where both the upper-bounds for the error probability and the error probability under constructed optimal attackers are presented. There are only two lines in Fig. 2.6 since the upper-bounds are achieved by the designed adversary and they overlap each other, which verifies the correctness of the construction process. From Fig. 2.6, we can see that, for each adversary, the attacker becomes more powerful as δ increases. In particular, for the hypothesis-aware case, when δ is large enough, the prediction error will reach $\frac{1}{2}$, the largest possible value.

In the third example, we investigate the impact of the alphabet size n on the prediction error. The PMFs before the attack are generated in the same manner as the second example. From Fig. 2.7, we have that, for a fixed attack amplitude $\delta = 50$, the prediction error decreases as the alphabet size n increases. The reason is that, as n increases, the relative attack strength δ/n decreases, and hence the impact of the attack on the error probability also decreases. However, if

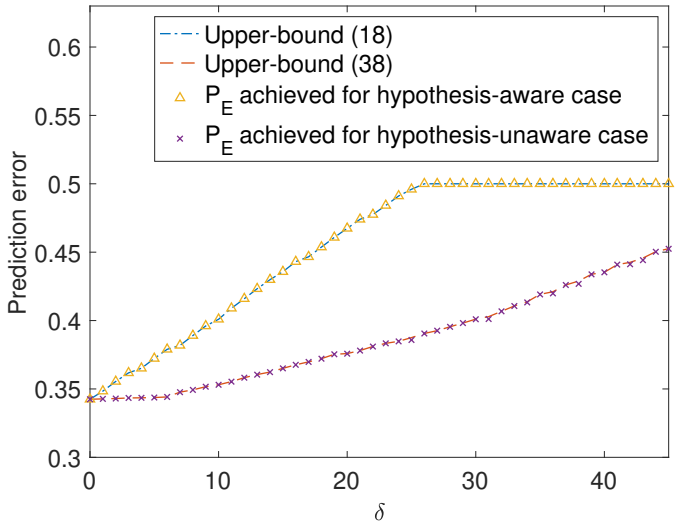


Figure 2.6: Prediction error v.s. $\delta, n = 200, m = 97$

the ratio between δ and n is fixed (for example, $\delta/n = 0.03, 0.06, 0.1$ as shown in Fig. 2.8 and $\delta/n = 0.1, 0.15, 0.2$ as shown in Fig. 2.9), there is no significant change in the prediction error as the alphabet size increases. In particular, from the hypothesis-aware result given in Fig. 2.8, we see that the prediction error reaches 0.5 when $\delta = 0.1n$ for n varies from 400 to 1000, indicating that even a relatively small perturbation could have a big impact on the prediction accuracy. On the other hand, for the hypothesis-unaware model, from Fig. 2.9, we see that it is harder for the prediction error to reach $\frac{1}{2}$, indicating that the strength of attack has been highly restricted if the hypothesis information is hidden from the adversary.

In the fourth example, we illustrate the characteristic of PMFs before and after attack. First, we generate the PMFs by truncating a Poisson distribution with parameter $\lambda_t, t = 0, 1$, since the normal Poisson distribution is defined on an infinite set. To normalize the distribution, we then move the mass on the tails to the finite alphabet equally and name the distribution as truncated Poisson distribution. Thus, the PMFs can be written as $\boldsymbol{p}_{t,i} = \boldsymbol{p}^0_{t,i} + d, t = 0, 1, 1 \leq i \leq n$, where $\boldsymbol{p}^0_{t,i} = \frac{(\lambda_t^i e^{-\lambda_t})}{i!}$ and $d = \frac{1-\sum_{i=1}^n \boldsymbol{p}^0_{t,i}}{n}$. By setting $n = 110$ and $\lambda_0 = 35, \lambda_1 = 75$ for $\mathcal{H}_0, \mathcal{H}_1$ respectively, the PMFs before attack are shown in Fig. 2.10. Under this setup, the PMFs after attack are shown in Fig. 2.10 and Fig. 2.11 for the hypothesis-aware and hypothesis-unaware attackers

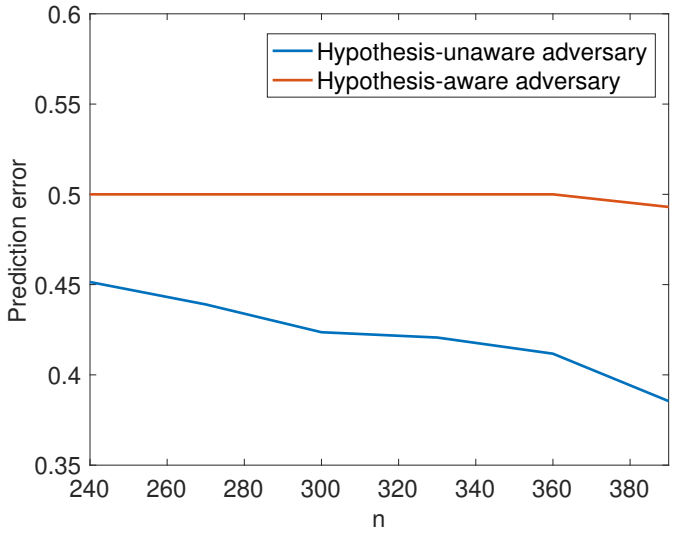


Figure 2.7: Prediction error v.s. alphabet size n for $\delta = 50$

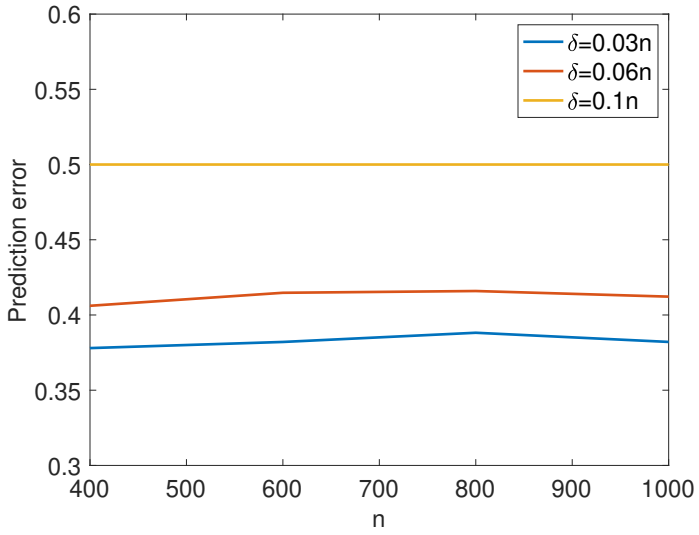


Figure 2.8: Prediction error v.s. alphabet size n (hypothesis-aware)

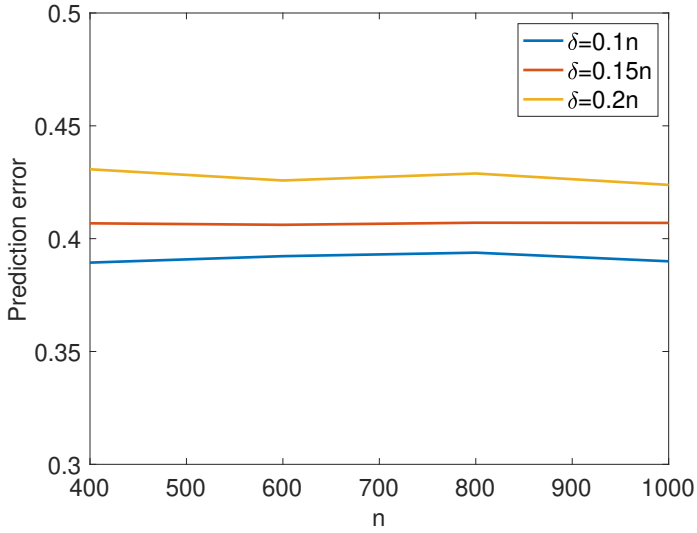


Figure 2.9: Prediction error v.s. alphabet size n (hypothesis-unaware)

respectively. In these figures, we set $\delta = 24$. The results show that, for both hypothesis-aware and hypothesis-unaware adversary, the PMFs after attack can be made the same under two hypotheses. As the result, for both adversary models, $P_E = \frac{1}{2}$ after the attack.

Fig. 2.12 illustrates the PMFs before and after attack for the hypothesis-unaware case when $\delta = 20$. From this figure, we can see that, $\boldsymbol{q}_0$ and $\boldsymbol{q}_1$, the PMFs after attack for different hypotheses, are not the same under the optimal hypothesis-unaware adversary. On the other hand, for the hypothesis-aware attacker, the error probability is equal to 1/2.

Fig. 2.13 illustrates how P_E increases as the attack amplitude δ increases. From this figure, we can see that, for both attack models, P_E increases with δ. Furthermore, the prediction error in the hypothesis-aware case is always larger than hypothesis-unaware case and reaches 1/2 earlier than the hypothesis-unaware case. This is consistent with the simulation result in the previous random distribution scenario.

2.5 Conclusion

In this chapter, we have investigated the adversarial robustness of hypothesis testing rules. We have formulated this as a minimax hypothesis testing problem. We have characterized the optimal attack

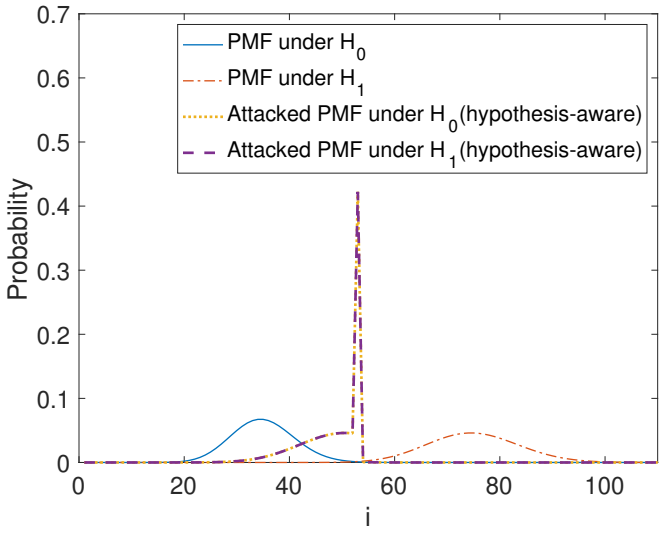


Figure 2.10: The PMF before and after attack (hypothesis-aware)

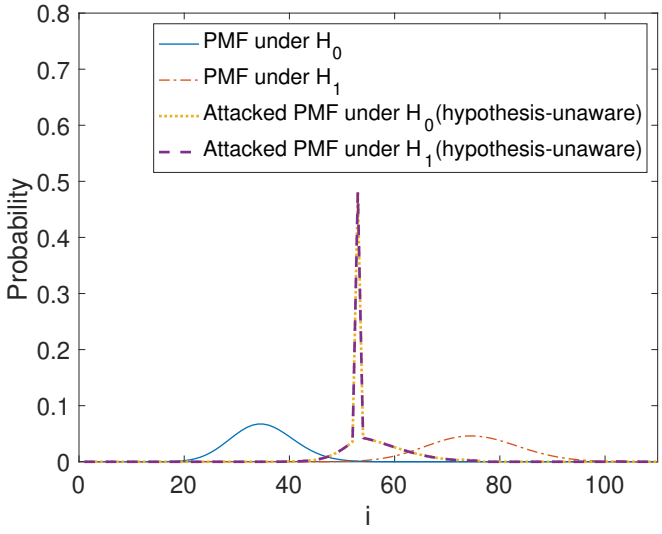


Figure 2.11: The PMF before and after attack (hypothesis-unaware)

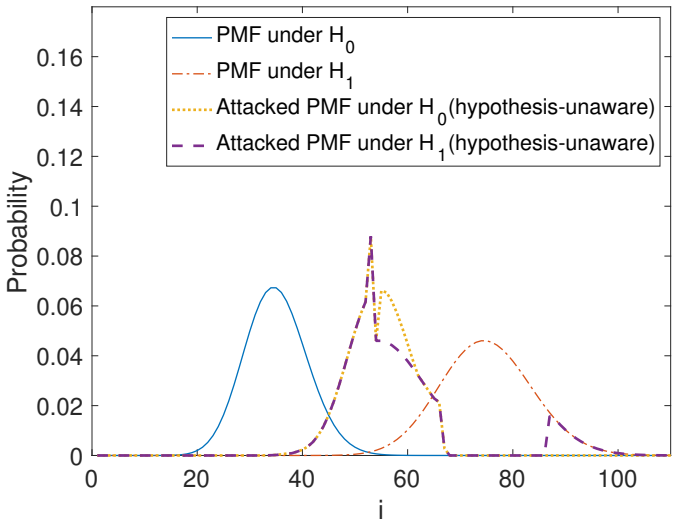


Figure 2.12: The PMF before and after attack (hypothesis-unaware) when $\delta = 20$

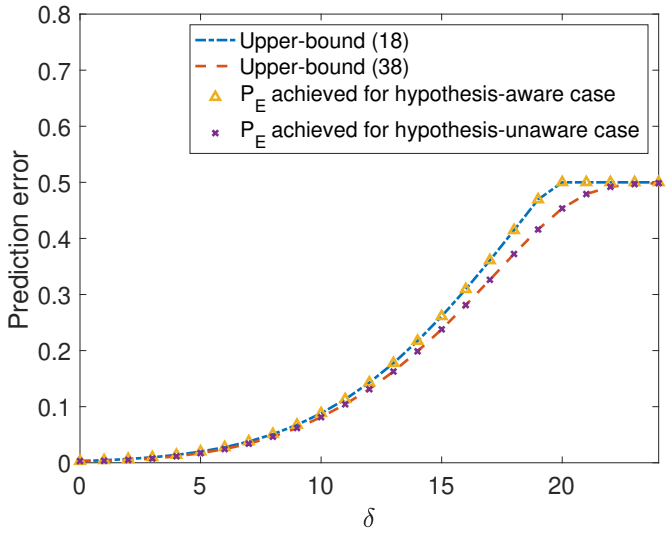


Figure 2.13: Prediction error v.s. δ for truncated Poisson distribution

and the corresponding optimal decision rules for both hypothesis-aware and hypothesis-unaware adversary models. We have also provided numerical examples to illustrate the analytical results obtained in this chapter.

Building on the problem formulation and analysis in this chapter, there are several interesting future research directions. Firstly, it is important to extend the analysis to the scenario where the true underlying distributions are unknown to both the attacker and decision-maker. Secondly, the application to steganography and steganalysis [166], in which steganography aims to hide secret messages in the cover media while steganalysis tries to detect hidden secret information embedded in the cover media, is another interesting research direction. Thirdly, our work can be applied to the decentralized detection setup [167–169], with a fusion center and distributed nodes, some of which might be compromised. The compromised nodes may send fake messages to the fusion center, and the goal of the fusion center is to make correct decisions in spite of the presence of misbehaving nodes. Finally, other than the amplitude constraint considered in this chapter, it is important to investigate other types of constraints on the adversary.

Chapter 3

Privacy-accuracy Trade-off of Inference as Service

In this chapter, we propose a general framework to provide a desirable trade-off between inference accuracy and privacy protection in the inference as service scenario (IAS). Instead of sending data directly to the server, the user will preprocess the data through a privacy-preserving mapping. This privacy-preserving mapping has two opposing effects. On one hand, it will prevent the server from observing the data directly and hence enhance the privacy protection. On the other hand, this might reduce the inference accuracy. To properly address the trade-off between these two competing goals, we formulate an optimization problem to find the optimal privacy-preserving mapping.

Particularly, in Section 3.1, we introduce the problem formulation. In Section 3.2, we present the proposed algorithm and provide the convergence analysis to find the local optimal privacy-mapping. In Section 3.3, we present numerical results. In Section 3.4, we offer concluding remarks.

3.1 Problem formulation

Consider an inference problem, in which one would like to infer the parameter $S \in \mathcal{S}$ of data $Y \in \mathcal{Y}$, in which $\mathcal{Y}$ has a finite alphabet. In the inference as service scenario, one would send Y to the server who will determine the parameter S using its sophisticated models and powerful

computing capabilities. However, directly sending data Y to the server brings the privacy issue, as now the server knows Y perfectly. To reduce the privacy leakage, instead of sending Y directly, one can employ a privacy-preserving mapping to transform data Y to $U \in \mathcal{U}$ and send U to the server. Here, $\mathcal{U}$ also has a finite alphabet and is allowed to be different from $\mathcal{Y}$. Without loss of generality, we will employ a randomized privacy-preserving mapping and use $p(u|y)$ to denote the probability that data $Y = y$ will be mapped to $U = u$ and the whole mapping is denoted as $P_{U|Y}$. Furthermore, we use P_S to denote the prior distribution of S and $P_{Y|S}$ to denote the conditional distribution Y given S, while the lower-case letter p is used to denote the component-wise probability (e.g., $p(s), p(y), p(y|s)$ will be used in the sequel).

To measure the inference accuracy, note that the distributional difference between P_S and $P_{S|U}$ characterizes the information about S contained in U. Since the inference at the server side is solely based on U, such information determines the inference accuracy. As $I(S;U)$ is the averaged Kullback–Leibler (KL) divergence between P_S and $P_{S|U}$, we use it to measure the inference accuracy. We would like to make $I(S;U)$ as large as possible, which means that we would like to retain as much information about the parameter of interest S in U as possible so that the server can make a more accurate inference.

To measure the privacy leakage, instead of choosing one particular privacy metric, we intend to investigate a general form $\mathbb{E}_{Y,U}[d(y,u)]$ that is applicable for different privacy metrics. Here, $d(y,u) = f(\frac{p(y)}{p(y|u)})$ and f is a continuous function defined on $(0, +\infty)$. We note that $\mathbb{E}_{Y,U}[d(y,u)] = \mathbb{E}_{Y,U}[f(\frac{p(y)}{p(y|u)})]$ measures the distributional distance between P_Y and $P_{Y|U}$, where P_Y is the prior distribution of Y and $P_{Y|U}$ is the posterior distribution of Y after observing U. Hence, the smaller the distance, the less information U can provide about Y and the better the privacy protection. Note that $\frac{p(y)}{p(y|u)} = \frac{p(u)}{p(u|y)}$. Hence we will also use $\frac{p(u)}{p(u|y)}$ as the argument to f in the sequel. Since $p(u|y)$ shows in the denominator, we assume that $\epsilon \leq p(u|y) \leq 1, \forall y, u$, where $\epsilon > 0$.

To balance the inference accuracy and privacy protection, we propose to find the privacy-

preserving mapping $P_{U|Y}$ by solving the following optimization problem

$$\max_{P_{U|Y}} \quad \mathcal{F}[P_{U|Y}] \triangleq I(S;U) - \beta \mathbb{E}_{Y,U}\left[f\left(\frac{p(y)}{p(y|u)}\right)\right], \tag{3.1}$$

$$\text{s.t.} \quad p(u|y) \geq \epsilon, \forall y, u,$$

$$\sum_u p(u|y) = 1, \forall y. \tag{3.2}$$

Here, $\beta \in (0, \infty)$ is a trade-off parameter that indicates the relative importance of maximizing $I(S;U)$ (i.e., maximizing inference accuracy) and minimizing the distance $\mathbb{E}_{Y,U}[d(y,u)]$ between P_Y and $P_{Y|U}$ (i.e., maximizing the privacy).

Another possible problem formulation is to maximize the inference accuracy under the constraint that the privacy leakage is less than certain threshold δ:

$$\max_{P_{U|Y}} \quad I(S;U) \tag{3.3}$$

$$\text{s.t.} \quad \mathbb{E}_{Y,U}\left[f\left(\frac{p(y)}{p(y|u)}\right)\right] \leq \delta, p(u|y) \geq \epsilon, \forall y, u, \sum_u p(u|y) = 1, \forall y.$$

However, directly solving such constrained optimization problems is very challenging. A typical way to solve this kind of problems is to form the Lagrangian of the maximization problem, whose objective is written as the weighted sum of the original objective and the constraints. Hence, our problem formulation can be viewed as the Lagrangian of the problem formulation (3.3). The trade-off parameter β can be treated as the Lagrangian multiplier. Different value of β corresponds to different privacy constraint δ in (3.3), whose value depends on different applications. In particular, using the proposed algorithms, solutions can be computed for a broad range of β. We can then obtain the Pareto optimal curve for accuracy and privacy leakage, where each point corresponds to one sub-problem solved to maximize the inference performance subject to a certain upper bound of privacy leakage. Then the user can select an operating point from the Pareto optimal curve depending on the user's preference and the constraint imposed by the applications.

For the privacy measure function f, we assume that

(a) $f(\cdot)$ is strictly convex;

(b) $f(\cdot)$ is twice-differentiable;

(c) $f^{'}(t)$ is l_f- Lipschitz continuous of t.

Here we provide some comments about these assumptions. (a) guarantees certain convexity of the problem. In particular, under (a), the sub-problems are shown to be convex, which ensures the feasibility and simplification of the proposed method. (b) and (c) are needed to ensure the convergence of the proposed method. These assumptions are fairly weak. As will be discussed in Section 3.3, most of the widely used distance measures satisfy these assumptions.

The proposed framework in (3.1) is very general. Different choices of f will lead to different privacy measures. For example, if we choose f to be $-\log(\cdot)$, then we have

$$\mathbb{E}_{Y,U}[d(y,u)] = -\sum_{y,u} p(y)p(u|y)\log\left(\frac{p(u)}{p(u|y)}\right) = \sum_{y} p(y)D_{KL}[P_{U|y} \parallel P_U] = I[U;Y],$$

in which $D_{KL}(\cdot \parallel \cdot)$ is the KL divergence. As the result, choosing f to be the $-\log$ function means we will use mutual information between U and Y to measure the information leakage, a very common choice in information theory study. More examples will be provided in Section 3.3.

3.2 Algorithms and Convergence Proof

In this section, we discuss how to solve the optimization problem defined in (3.1) for general f. One natural approach to solving (3.1) is to apply the gradient ascent (GA) algorithm. However, GA faces several challenges such as proper step size, computation complexity, convergence speed and the quality of the optimal point found etc. To overcome these challenges, we propose a new algorithm that transforms the maximization problem over single argument to an alternative maximization problem over multiple arguments and then employ ideas from ADMM to solve the transformed problem.

3.2.1 Algorithm

We first have the following lemma that are useful for transforming the objective function.

Lemma 2.

$$I(S;U) = I(S;Y) - \sum_{u,y} p(y)p(u|y)D_{KL}[P_{S|y} \parallel P_{S|u}].$$

Proof. Please refer to Appendix B.1. □

By Lemma 2, the objective function defined in (3.1) can be written as

$$\mathcal{F}[P_{U|Y}, P_U, P_{S|U}] \;\; = \;\; I(S;Y) - \beta\mathbb{E}_{Y,U}[d(y,u)] - \sum_{u,y} p(y)p(u|y)D_{KL}[P_{S|y} \parallel P_{S|u}].$$

Note that $I(S;Y)$, $p(y)$ and $p(s|y)$ are fixed, hence the cost function can be viewed as a function of three arguments $P_{U|Y}$, P_U and $P_{S|U}$. For consistency, we require the following equations to be satisfied simultaneously

$$p(u) = \sum_y p(u|y)p(y), \forall u, \tag{3.4}$$

$$p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}, \forall u, \forall s. \tag{3.5}$$

By (3.5), we further require that $p(u) > 0, \forall u$. As the result, we can reformulate (3.1) as the following alternative optimization problem

$$
\begin{aligned}
\max_{P_{S|U},P_U,P_{U|Y}} \quad & \mathcal{F}[P_{U|Y}, P_U, P_{S|U}]. \qquad (3.6)\\
\text{s.t.} \quad & p(u|y) \geq \epsilon, \forall y, \forall u, \ \sum_u p(u|y) = 1, \forall y,\\
& p(u) > 0, \forall u, \ \sum_u p(u) = 1,\\
& p(u) = \sum_y p(u|y)p(y), \forall u,\\
& p(s|u) \geq 0, \forall u, \forall s, \ \sum_s p(s|u) = 1, \forall u,\\
& p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}, \forall u, \forall s.
\end{aligned}
$$

The following lemma illustrates the nice property of the alternative formulation (3.6): the alternative optimization problem is convex in each argument given the other two arguments.

Lemma 3. *Suppose that $f(\cdot)$ is a strictly convex function. Then for given $P_U, P_{S|U}$, $\mathcal{F}[P_{U|Y}, P_U, P_{S|U}]$ is concave in each $P_{U|y_i}, \forall y_i \in \mathcal{Y}$. Similarly, for given $P_{U|Y}, P_{S|U}$, $\mathcal{F}[P_{U|Y}, P_U, P_{S|U}]$ is concave in P_U. For given $P_{U|Y}, P_U$, $\mathcal{F}[P_{U|Y}, P_U, P_{S|U}]$ is concave in $P_{S|U}$.*

Proof. Please refer to Appendix B.2. □

Using this lemma, a natural approach to maximizing the objective function in (3.6) is to alternately iterate between $P_{U|Y}$, P_U and $P_{S|U}$ until reaching convergence. In particular, we propose an iterative algorithm with two blocks to obtain a solution to (3.6): update of $P_{S|U}$ and update of $P_{U|Y}, P_U$. Firstly, for a given P_U and $P_{U|Y}$, we update $P_{S|U}$ by solving the maximization on $P_{S|U}$ and derive an analytical result as a function of P_U and $P_{U|Y}$. Secondly, for the derived $P_{S|U}$, we update P_U and $P_{U|Y}$ by using the ADMM scheme to solve the maximization on P_U and $P_{U|Y}$. In the following, we show that the proposed algorithm will converge. We would like to note that, however, as the problem in (3.6) is non-convex in the product space of $\{P_{U|Y}, P_U, P_{S|U}\}$, the derived limit point is not expected to be the global optimal solution of (3.6). In the following, we provide details for each iteration. The convergence proof of the proposed algorithm will be presented in

Section 3.2.2.

Updating $P_{S|U}$

For the $P_{S|U}$ subproblem, the maximization problem is

$$\begin{aligned} \max_{P_{S|U}} \quad & \mathcal{F}[P_{S|U}|P_{U|Y}, P_U], \\ \text{s.t.} \quad & p(s|u) \geq 0, \forall u, \forall s, && (3.7) \\ & \sum_s p(s|u) = 1, \forall u, && (3.8) \\ & p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}, \forall u, \forall s. && (3.9) \end{aligned}$$

Lemma 4. *The solution to the $P_{S|U}$ subproblem is*

$$p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}. \tag{3.10}$$

Proof. Please refer to Appendix B.3. □

Updating $P_{U|Y}$ and P_U

Now, for a given $P_{S|U}$, we discuss how to update $P_{U|Y}$ and P_U by solving

$$\max_{P_{U|Y} \in \mathcal{P}_{U|Y}, P_U \in \mathcal{P}_U} \quad \mathcal{F}[P_{U|Y}, P_U|P_{S|U}], \tag{3.11}$$

$$\text{s.t. } \delta(u) = p(u) - \sum_y p(u|y)p(y) = 0, \forall u, \tag{3.12}$$

where

$$\mathcal{P}_{U|Y} = \{P_{U|Y} : p(u|y) \geq \epsilon, \sum_u p(u|y) = 1\}, \tag{3.13}$$

$$\mathcal{P}_U = \{P_U : p(u) > 0, \sum_u p(u) = 1\}, \tag{3.14}$$

and (3.12) corresponds to the consistency requirement (3.4).

Moreover, note that each row in the matrix $P_{U|Y}$ is independent and we further show that the objective function in (3.11) can be written as the sum of $|\mathcal{Y}|$ terms, each of which depends only on one row of $P_{U|Y}$.

$$\begin{aligned}\mathcal{F}[P_{U|Y}, P_U|P_{S|U}] &= -\beta \sum_{i=1}^{|\mathcal{Y}|} \left[p(y_i) \sum_u p(u|y_i) d\left(\frac{p(u)}{p(u|y_i)}\right)\right] \\ &\quad - \sum_{i=1}^{|\mathcal{Y}|} \left[p(y_i) \sum_u p(u|y_i) D_{KL}[P_{S|y_i} \parallel P_{S|u}]\right] + I(S;Y) \\ &= \sum_{i=1}^{|\mathcal{Y}|} \mathcal{F}_i' \left[P_{U|Y}, P_U|P_{S|U}\right] + I(S;Y), \end{aligned} \tag{3.15}$$

where

$$\mathcal{F}_i' \left[P_{U|Y}, P_U|P_{S|U}\right] = p(y_i) \left[-\beta \sum_u p(u|y_i) f\left(\frac{p(u)}{p(u|y_i)}\right) - \sum_u p(u|y_i) D_{KL}[P_{S|y_i} \parallel P_{S|u}]\right]. \tag{3.16}$$

Thus, the optimization on $P_{U|Y}$ can be divided into $|\mathcal{Y}|$-problems, each of which corresponds to one row in $P_{U|Y}$.

As the result, although (3.11) is a non-convex problem in $(P_{U|Y}, P_U)$ jointly, it is a convex problem of one argument given the others, as shown in Lemma 3. This motivates us to apply the ADMM approach to solve the problem.

The augmented Lagrangian for the above problem is

$$\mathcal{L}[P_{U|Y}, P_U, P_{S|U}; \Lambda]\mathcal{F}[P_{U|Y}, P_U|P_{S|U}] + \sum_u \lambda(u)\delta(u) - \frac{\rho}{2}\sum_u \delta^2(u), \tag{3.17}$$

where Λ is a vector of size $|\mathcal{U}|$ and each component is denoted as $\lambda(u)$. Since $P_{S|U}$ is given, we will omit it from the expression of $\mathcal{L}$.

In the ADMM approach, there are updates of $P_{U|Y}$, P_U and Λ respectively. Exploiting the

structure in (3.15), we can solve (3.11) using the following iterative procedure

$$P_{U|y_i}^{t+1} = \arg \max_{P_{U|y_i} \in \mathcal{P}_{U|y_i}} \mathcal{L}[P_{U|y_i}, P_{U|Y^{(i-)}}^{t+1}, P_{U|Y^{(i+)}}^{t}, P_U^t; \Lambda^t], i = 1, 2, \cdots, |\mathcal{Y}|, \tag{3.18}$$

$$P_U^{t+1} = \arg \max_{P_U \in \mathcal{P}_U} \mathcal{L}[P_{U|Y}^{t+1}, P_U; \Lambda^t], \tag{3.19}$$

$$\Lambda^{t+1} = \Lambda^t - \rho(P_U^{t+1} - (P_{U|Y}^{t+1})^T P_Y), \tag{3.20}$$

$$\text{or} \quad \lambda^{t+1}(u) = \lambda^t(u) - \rho[p^{t+1}(u) - \sum_y p^{t+1}(u|y)p(y)] = \lambda^t(u) - \rho\delta^{t+1}(u),$$

where $\mathcal{P}_{U|y_i} = \{P_{U|y_i} : p(u|y) \geq \epsilon, \sum_u p(u|y_i) = 1\}$, $P_{U|Y^{(i-)}}$ denotes all rows before the i-th row in the matrix $P_{U|Y}$ and $P_{U|Y^{(i+)}}$ denotes all rows after the i-th row. Note that here we use Gauss–Seidel ADMM where the local variables are updated sequentially in the Gauss–Seidel order and current conditional distributions ($P_{U|Y^{i-}}^{t+1}$ and $P_{U|Y^{i+}}^{t}$) are used to obtain $P_{U|y_i}^{t+1}$. Another update approach is to use $P_{U|Y^{i-}}^{t}$ to update $P_{U|y_i}$ in the $(t+1)$-th iteration. It has been shown that for multi-block problems, Gauss–Seidel ADMM often performs numerically better in practice than the directly extended ADMM [96, 97, 170–172], as the updated information $P_{U|Y^{i-}}^{t+1}$ is immediately utilized.

For $P_{U|y_i}$, the optimization problem is

$$\begin{aligned} \max_{P_{U|y_i}} \quad & \mathcal{L}[P_{U|y_i}, P_{U|Y^{(i-)}}^{t+1}, P_{U|Y^{(i+)}}^{t}, P_U^t; \Lambda^t], \\ \text{s.t.} \quad & p(u|y_i) \geq \epsilon, \forall u, \sum_u p(u|y_i) = 1. \end{aligned} \tag{3.21}$$

We have the following lemma regarding the objective function in (3.21). The proof follows similar steps to the proof of Lemma 3.

Lemma 5. *The objective function in* (3.21) *is a strictly concave function.*

Proof. Please refer to Appendix B.4. □

Hence, each sub-problem is a convex optimization problem with $|\mathcal{U}|$ inequality constraints and one equality constraint. In practice, under a specified $f(\cdot)$, the sub-problem can be solved

numerically.

The sub-problem with respect to P_U is

$$\max_{P_U} \mathcal{L}[P_{U|Y}^{t+1}, P_U; \Lambda^t], \text{s.t. } p(u) > 0, \forall u, \ \sum_u p(u) = 1.$$

Following similar steps of Lemma 3, we can prove the following lemma.

Lemma 6. *The objective function in* (3.22) *is a strictly concave function.*

Proof. Please refer to Appendix B.4. □

Although there is a constraint, $P_U \in \mathcal{P}_U$, in this sub-problem, we can ignore it first and in the convergence proof, we will show that for the limit point, the constraint is naturally satisfied. We represent the solution to the unconstrained problem as $P_U^{t+1} = \arg\max_{P_U} \mathcal{L}[P_{U|Y}^{t+1}, P_U; \Lambda^t]$.

After solving two sub-problems on $P_{U|Y}$ and P_U respectively, we update the value of Λ.

In summary, we employ two nested loops to find the privacy-preserving mapping. In the outer loop, there are two update steps: update of $P_{S|U}$ and update of $(P_{U|Y}, P_U)$, where the update of $(P_{U|Y}, P_U)$ is performed by ADMM (which will be referred to as the inner loop). In the inner loop, we update $P_{U|Y}$ and P_U by going through the process of (3.18), (3.19), (3.20). We will use (j) to denote the j-th outer iteration and use $(j), t$ to denote the arguments at the t-th inner iteration of the j-th outer iteration. The algorithm is summarized in Algorithm 3.1. To quantify the matrix differences, we use the Frobenius norm [173], where for an $m \times n$ matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2}$. To quantify the vector differences, we use the ℓ_2 norm, where for vector $\boldsymbol{b} = (b_1, b_2, \cdots, b_n)$, $\|\boldsymbol{b}\|_2^2 = \sum_{i=1}^n b_i^2$. For the thresholds, η is chosen to be a small value such that the function value is converged and η_p is chosen to be a small value such that $\mathcal{L}[P_{U|Y}^{t+1}, P_U^{t+1}; \Lambda^t] \geq \mathcal{L}[P_{U|Y}^{t+1}, P_U^t; \Lambda^t] \geq \mathcal{L}[P_{U|Y}^t, P_U^t, \Lambda^t]$ is true.

3.2.2 Convergence Analysis

In this section, we provide the convergence proof for Algorithm 3.1. To prove the convergence of the proposed iterative algorithm, we need to verify that the value of the functional $\mathcal{F}$ does not

Algorithm 3.1 Design the privacy-preserving mapping

Input:

Prior distribution P_S and conditional distribution $P_{Y|S}$.

Trade-off parameter β.

Converge parameter η, η_p, η_d.

Output:

A mapping $P_{U|Y}$ from $Y \in \mathcal{Y}$ to $U \in \mathcal{U}$.

Initialization:

Randomly initiate $P_{U|Y}$ and calculate $P_U, P_{S|U}$ by (3.4) and (3.5).

$j = 1$.
while $\left\| P_{S|U}^{(j)} - P_{S|U}^{(j-1)} \right\|_F > \eta$ **do**
  $P_U^{(j),1} = P_U^{(j-1)}$.
  $P_{U|Y}^{(j),1} = P_{U|Y}^{(j-1)}$.
  $t = 1$.
  while $t = 1$ or $\left\| P_U^{(j),t} - P_U^{(j),t-1} \right\|_2^2 > \eta_p$ **do**
    Update $P_{U|y_i}$ by solving (3.21).
    Update P_U by solving (3.22).
    Update Λ by (3.20).
    $t = t + 1$.
  Update $P_{S|U}^{(i)}$ by (3.10).
  $j = j + 1$.
return $P_{U|Y}$

decrease while iterating, and that this functional is bounded from above.

The following lemma shows that $\mathcal{F}$ is upper-bounded.

Lemma 7. *For a continuous function $f(\cdot)$, $\mathcal{F}[P_{U|Y}, P_U, P_{S|U}]$ is bounded from above.*

Proof. Please refer to Appendix B.5. □

Then we prove that the value of $\mathcal{F}$ is non-decreasing between two iterations of the outer loop. There are two steps in the outer loop, updating $P_{S|U}$ by (3.10), and updating $(P_{U|Y}, P_U)$ by applying ADMM. For the update of $P_{S|U}$, since the optimization with respect to $P_{S|U}$ is a convex optimization problem and has a closed-form solution as the update function, the objective function $\mathcal{F}$ is non-decreasing in this step. To show that the value of $\mathcal{F}$ is non-decreasing for the limit point found by ADMM, it is necessary to prove that the proposed ADMM procedure converges subsequently. Otherwise, the consistency requirement between P_U and $P_{U|Y}$ may not be satisfied.

In particular, in the following we prove that any sequence generated by the proposed ADMM procedure is bounded and has a limit point that is also the stationary point of (3.11), and the value of $\mathcal{F}$ is upper-bounded and non-decreasing between iterations of ADMM.

We note that the convergence proof of the proposed ADMM procedure for our problem setup is non-trivial, as the considered objective function has more than 2 local variables and is non-separable with respect to these local variables. Directly using multi-block ADMM may be non-convergent, even if the functions are separable with respect to these blocks of variables [94], and numerous research efforts have been devoted to analyzing the convergence of multi-block ADMM under certain assumptions [96, 97, 174]. In contrast to the separable case, studies on the convergence properties of n-block ADMM with non-separable objective, even for $n = 2$, are limited [175, 176], and the convergence is not guaranteed and has to be handled differently.

To make the presentation clear, in the following, we consider the case $|\mathcal{Y}| = 2$ and the proof can be easily generalized to the case when $\mathcal{Y}$ has a finite alphabet. For $|\mathcal{Y}| = 2$, the optimization problem in (3.1) can be further represented as

$$\begin{aligned}
\max \quad & -\left[p(y_1)\sum_u p\left(u|y_1\right) D_{KL}[P_{S|y_1} \parallel P_{S|u}] + p(y_2)\sum_u p\left(u|y_2\right) D_{KL}[P_{S|y_2} \parallel P_{S|u}]\right] \\
& -\beta \sum_u \left[p\left(u|y_1\right) p(y_1) f\left(\frac{p(u)}{p(u|y_1)}\right) + p(u|y_2)p(y_2) f\left(\frac{p(u)}{p(u|y_2)}\right)\right], \\
\text{s. t} \quad & p(u|y_i) \geq \epsilon, \forall u, \sum_u p(u|y_i) = 1, i = 1, 2, \\
& p(u) > 0, \forall u, \sum_u p(u) = 1, \\
& -p(u|y_1)p(y_1) - p(u|y_2)p(y_2) + p(u) = 0, \forall u,
\end{aligned}$$

where the last constraint can also be written in the vector form, $-p(y_1)P_{U|y_1} - p(y_2)P_{U|y_2} + P_U = \mathbf{0}$.

For presentation convenience, we denote

$$h_i(P_{U|y_i}) = -p(y_i)\sum_u p(u|y_i)D_{KL}[P_{S|y_i} \parallel P_{S|u}], i = 1, 2,$$
$$g(P_{U|y_1}, P_{U|y_2}, P_U) = -\beta\sum_u \left[p(u|y_1)p(y_1)f\left(\frac{p(u)}{p(u|y_1)}\right) + p(u|y_2)p(y_2)f\left(\frac{p(u)}{p(u|y_2)}\right)\right].$$

Thus, the objective function is $h_1(P_{U|y_1}) + h_2(P_{U|y_2}) + g(P_{U|y_1}, P_{U|y_2}, P_U)$, and the augmented Lagrangian is

$$\begin{aligned} & \mathcal{L}[P_{U|Y}, P_U, P_{S|U}; \Lambda] \\ = \; & \mathcal{F}[P_{U|Y}, P_U|P_{S|U}] + \sum_u \lambda(u)\delta(u) - \frac{\rho}{2}\sum_u \delta(u)^2 \\ = \; & h_1(P_{U|y_1}) + h_2(P_{U|y_2}) + g(P_{U|y_1}, P_{U|y_2}, P_U) + \sum_u \lambda(u)\delta(u) - \frac{\rho}{2}\sum_u \delta(u)^2. \end{aligned}$$

For the update of the dual variable Λ, we have the following lemma which characterizes the relationship between the dual variable Λ and the primal variables.

Lemma 8. *Suppose that $f(\cdot)$ is twice-differentiable and $f^{'}(t)$ is l_f- Lipschitz continuous of t. We have*

$$\left\|\Lambda^{t+1} - \Lambda^t\right\|_2^2 \leq l_\Lambda\left(\left\|P_{U|y_1}^{t+1} - P_{U|y_1}^t\right\|_2^2 + \left\|P_{U|y_2}^{t+1} - P_{U|y_2}^t\right\|_2^2 + \left\|P_U^{t+1} - P_U^t\right\|_2^2\right), \tag{3.22}$$

with $l_\Lambda = \frac{16\beta^2 l_f^2}{\epsilon^4}$.

Proof. Please refer to Appendix B.6. □

For the ascent of $\mathcal{L}$ between two iterations, we have the following lemma.

Lemma 9. *Suppose that $f(\cdot)$ is twice-differentiable and $f^{'}(t)$ is l_f- Lipschitz continuous of t. We*

have

$$\begin{aligned}
&\mathcal{L}\left[P_{U|Y}^{t+1}, P_U^{t+1}; \Lambda^{t+1}\right] - \mathcal{L}\left[P_{U|Y}^{t}, P_U^{t}; \Lambda^{t}\right] \\
\geq\ & \left[\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}\right] \|P_{U|y_1}^{t+1} - P_{U|y_1}^{t}\|_2^2 + \left[\frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}\right] \left\|P_{U|y_2}^{t+1} - P_{U|y_2}^{t}\right\|_2^2 \\
&+ \left(\frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\right) \left\|P_U^{t+1} - P_U^{t}\right\|_2^2, \qquad (3.23)
\end{aligned}$$

where $l_{y_1} = l_{y_2} = \frac{\beta l_f}{\epsilon^3}, l_u = \frac{\beta l_f}{\epsilon}$.

Proof. Please refer to Appendix B.7. □

With these supporting results, we now analyze the convergence of the proposed ADMM procedure. We first show that $\mathcal{L}$ is monotonic and upper-bounded, and the sequence $\{P_{U|Y}, P_U, \Lambda\}^t$ generated by ADMM is bounded.

Proposition 2. *Suppose that $f(\cdot)$ is twice-differentiable and $f^{'}(t)$ is l_f-Lipschitz continuous of t. We have that*

(1) if $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$, $\mathcal{L}[P_{U|Y}^{t+1}, P_U^{t+1}; \Lambda^{t+1}] \geq \mathcal{L}[P_{U|Y}^{t}, P_U^{t}, \Lambda^{t}]$;

(2) $\forall t \in \mathbb{N}$, $\mathcal{L}[P_{U|Y}^{t}, P_U^{t}; \Lambda^{t}]$ *is upper-bounded;*

(3) $\{P_{U|Y}, P_U, \Lambda\}^t$ *is bounded.*

Proof. Please refer to Appendix B.8. □

We then show the asymptotic regularity of the sequence $\{P_{U|Y}, P_U, \Lambda\}^t$.

Proposition 3. *Suppose that $f(\cdot)$ is twice-differentiable and $f^{'}(t)$ is l_f- Lipschitz continuous of t. When ρ is sufficiently large such that* $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$, *as $t \to \infty$, we have*

(1) $\left\|P_{U|y_1}^{t+1} - P_{U|y_1}^{t}\right\|_2^2 \to 0$,

(2) $\left\| P_{U|y_2}^{t+1} - P_{U|y_2}^{t} \right\|_2^2 \to 0$,

(3) $\left\| P_{U}^{t+1} - P_{U}^{t} \right\|_2^2 \to 0$.

(4) $\|\Lambda^{t+1} - \Lambda^{t}\|_2^2 \to 0$,

(5) $P_{U}^{t+1} - p\left(y_1\right) P_{U|y_1}^{t+1} - p\left(y_2\right) P_{U|y_2}^{t+1} \to 0$.

Proof. Please refer to Appendix B.9. □

Proposition 4. *The sequence* $\{P_{U|Y}, P_U, \Lambda\}^t$ *has a limit point* $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$*, which is also a stationary point of* (3.11)*.*

Proof. Please refer to Appendix B.10. □

We now summarize the convergence results in the following theorem.

Theorem 4. *Suppose that* $f(\cdot)$ *is twice-differentiable and* $f^{'}(t)$ *is* l_f*-Lipschitz continuous of* t*. Choose* ρ *such that* $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$*. The proposed ADMM procedure could converge subsequently, that is, staring from any* $(P_{U|Y}^0, P_U^0, \Lambda^0)$*, it generates a sequence that is bounded, has a limit point* $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$*, and the limit point is a stationary point of* (3.11)*.*

Proof. Please refer to Appendix B.11. □

Therefore, for the limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$, the value of $\mathcal{F}$ is non-decreasing after the ADMM procedure. Then $\mathcal{F}$ is also non-decreasing between two iterations of the outer loop, which indicates that the proposed algorithm will converge.

For the case $|Y| = k$, there will be $(k+1)$ terms on the right hand side of (3.22) and (3.23). Then Propositions 2, 3, 4 and Theorem 4 still hold in a similar manner and the convergence analysis also applies.

3.2.3 Stronger Convergence for f with More Assumptions

In Section 3.2.2, for the convergence analysis of ADMM, the value of ρ should be chosen large enough such that $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$. Thus, the feasible set of ρ will depend on the choice of ϵ. In this subsection, we propose another ADMM procedure with Bregman distance and make stronger assumptions on f to provide a convergence analysis with weaker constraints on ρ.

First we introduce the definition of Bregman distance. Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable and strictly convex function. Denote $\nabla\phi(y)$ as the gradient of ϕ on y. Then the Bregman distance induced by ϕ is defined as

$$\Delta_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle, \tag{3.24}$$

where ϕ is called the kernel function or distance-generating function. From the property of Bregman distance, we have that $\Delta_\phi(x, y)$ is convex in x for fixed y [177]. The Bregman distance plays an important role in iterative algorithms. In particular, Bregman divergences are used to replace the quadratic penalty term in the standard ADMM (see $\delta^2(u)$ in (3.17)). Then we can choose a suitable Bregman divergence so that the sub-problems can be solved more efficiently [177].

To solve the optimization problem in (3.11), for notation simplicity, we denote $x_1 : P_{U|y_1}$, $x_2 : P_{U|y_2}$, and $v : P_U$.

Recall the definition of $h_1(\cdot), h_2(\cdot), g(\cdot)$ in Section 3.2.2. We propose an algorithm starting with (x_1^0, x_2^0, v^0) and Λ^0. Suppose that $\varphi_1, \varphi_2, \phi$ are differentiable and strictly convex functions. Then with the given iteration point $w^k = \left(x_1^k, x_2^k, v^k, \Lambda^k\right)$, the new iteration point $w^{k+1} =$

$\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^{k+1}\right)$ is given as:

$$
\begin{aligned}
x_1^{k+1} = \arg\max \Big\{ & h_1(x_1) + \left(x_1 - x_1^k\right)^T \nabla_{x_1} g\left(x_1^k, x_2^k, v^k\right) \\
& - \frac{\rho}{2} \left\| p(y_1) x_1 + p(y_2) x_2^k - v^k - \frac{\Lambda^k}{\rho} \right\|_2^2 - \Delta_{\varphi_1}\left(x_1, x_1^k\right) \Big\}, \\
x_2^{k+1} = \arg\max \Big\{ & h_2(x_2) + \left(x_2 - x_2^k\right)^T \nabla_{x_2} g\left(x_1^k, x_2^k, v^k\right) \\
& - \frac{\rho}{2} \left\| p(y_1) x_1^{k+1} + p(y_2) x_2 - v^k - \frac{\Lambda^k}{\rho} \right\|_2^2 - \Delta_{\varphi_2}\left(x_2, x_2^k\right) \Big\}, \\
v^{k+1} = \arg\max \Big\{ & g\left(x_1^{k+1}, x_2^{k+1}, v\right) - \frac{\rho}{2} \left\| p(y_1) x_1^{k+1} + p(y_2) x_2^{k+1} - v - \frac{\Lambda^k}{\rho} \right\|_2^2 - \Delta_{\phi}\left(v, v^k\right) \Big\}, \\
\Lambda^{k+1} = & \Lambda^k - \rho\left(p(y_1) x_1^{k+1} + p(y_2) x_2^{k+1} - v^{k+1}\right),
\end{aligned} \tag{3.25}
$$

where $\Delta_{\varphi_1}\left(x_1, x_1^k\right), \Delta_{\varphi_2}\left(x_2, x_2^k\right)$, and $\Delta_{\phi}\left(v, v^k\right)$ are the Bregman distances associated with φ_1, φ_2, and ϕ respectively. Here, φ_1, φ_2, and ϕ should be properly chosen with respect to different $f(\cdot)$ adopted in the privacy measure.

To guarantee that the algorithm converges, we assume that

(i) ∇g is l_g-Lipschitz continuous;

(ii) $\nabla\varphi_1, \nabla\varphi_2, \nabla\phi$ are Lipshitz continuous with the modulus $l_{\varphi_1}, l_{\varphi_2}, l_{\phi}$, respectively;

(iii) $\varphi_1, \varphi_2, \phi$ are strongly convex with the modulus $\delta_{\varphi_1}, \delta_{\varphi_2}, \delta_{\phi}$, and $\delta_{\varphi_1}, \delta_{\phi_2} > l_g$.

Then we have

Lemma 10.

$$
\begin{aligned}
& \left\|\Lambda^{k+1} - \Lambda^k\right\|_2^2 \\
\leq\ & 3l_g^2 \left(\left\|x_1^{k+1} - x_1^k\right\|_2^2 + \left\|x_2^{k+1} - x_2^k\right\|_2^2 \right) + 3\left(l_g^2 + l_{\phi}^2\right) \left\|v^{k+1} - v^k\right\|_2^2 + 3l_{\phi}^2 \left\|v^k - v^{k-1}\right\|_2^2.
\end{aligned}
$$

Proof. Please refer to Appendix B.12. □

By considering the updates of 3 primal variables, we have

Lemma 11.

$$\begin{aligned}
&\left(\mathcal{L}\left(w^{k+1}\right)-\frac{3l_\phi^2}{\rho}\left\|v^{k+1}-v^k\right\|_2^2\right)-\left(\mathcal{L}\left(w^k\right)-\frac{3l_\phi^2}{\rho}\left\|v^k-v^{k-1}\right\|_2^2\right)\\
\geq\ &\left(\frac{\delta_{\varphi_1}-l_g}{2}-\frac{3l_g^2}{\rho}\right)\left\|x_1^{k+1}-x_1^k\right\|_2^2+\left(\frac{\delta_{\varphi_2}-l_g}{2}-\frac{3l_g^2}{\rho}\right)\left\|x_2^{k+1}-x_2^k\right\|_2^2\\
&+\left(\frac{\delta_\phi}{2}-\frac{3l_g^2+6l_\phi^2}{\rho}\right)\left\|v^{k+1}-v^k\right\|_2^2.
\end{aligned}$$

Proof. Please refer to Appendix B.13. □

Proposition 5. *Under assumptions (i), (ii), (iii), we have*

(1) if $\rho \geq \max\{\frac{6l_g^2}{\delta_{\varphi_1}-l_g}, \frac{6l_g^2}{\delta_{\varphi_2}-l_g}, \frac{6l_g^2+12l_\phi^2}{\delta_\phi}\}$ *(feasible under assumption (iii)),*

$$\left(\mathcal{L}\left(w^{k+1}\right)-\frac{3l_\phi^2}{\rho}\left\|v^{k+1}-v^k\right\|_2^2\right)-\left(\mathcal{L}\left(w^k\right)-\frac{3l_\phi^2}{\rho}\left\|v^k-v^{k-1}\right\|_2^2\right)\geq 0;$$

(2) $\forall k \in \mathbb{N}$, $\mathcal{L}[w^k]$ *is upper-bounded;*

(3) $\{w\}^k$ *is bounded.*

Then following similar analysis in Section 3.2.2, when ρ is chosen properly such that $\rho \geq \max\{\frac{6l_g^2}{\delta_{\varphi_1}-l_g}, \frac{6l_g^2}{\delta_{\varphi_2}-l_g}, \frac{6l_g^2+12l_\phi^2}{\delta_\phi}\}$, we have $\left\|x_1^{k+1}-x_1^k\right\|_2^2 \to 0$, $\left\|x_2^{k+1}-x_2^k\right\|_2^2 \to 0$, and $\left\|v^{k+1}-v^k\right\|_2^2 \to 0$. By Lemma 10, we have $\left\|\Lambda^{k+1}-\Lambda^k\right\|_2^2 \to 0$. Moreover, the limit point of $\{w\}^k$ can also be shown to be the stationary point of (3.11). Thus, when replacing the ADMM procedure in Section 3.2.1 with this ADMM procedure with Bregman distance, Algorithm 3.1 converges in a similar manner.

3.3 Examples and Numerical results

In this section, we first give examples of different choices of f and then provide numerical results with specific f to show the performance of the proposed method.

3.3.1 Examples of f

We now provide examples of f, each of which leads to a well-known and widely used divergence measure.

In the first example, we consider $f(t) = -\log(t)$. As shown in Section 3.1, if $f(t) = -\log(t)$, the privacy measure is then the mutual information. For the algorithm proposed in this chapter, we check whether all the assumptions are satisfied. The constraints defined in the maximization problem for $P_{U|Y}$ stipulate that $\epsilon \leq p(u|y) \leq 1$. Consequently, this infers that $\epsilon \leq \frac{p(u)}{p(u|y)} \leq \frac{1}{\epsilon}$. Then we first have that $-\log(\cdot)$ is strictly convex on $\left[\epsilon, \frac{1}{\epsilon}\right]$. Secondly, we have that $f'(t) = -\frac{1}{t}$ is Lipschitz continuous since it is everywhere differentiable on $\left[\epsilon, \frac{1}{\epsilon}\right]$ and the absolute value of the derivative is bounded above by $\frac{1}{\epsilon^2}$.

In the second example, we consider the following strictly convex function $f(t) = t\log\frac{2t}{t+1} + \log\frac{2}{t+1}$. This choice leads to the Jensen-Shannon divergence [178]: $\mathbb{E}_{Y,U}[d(y,u)] = \sum_y p(y) JS[P_{U|y}, P_U]$, in which $JS[P_{U|y}, P_U] = D_{KL}\left[P_{U|y} \parallel \frac{P_{U|y}+P_U}{2}\right] + D_{KL}\left[P_U \parallel \frac{P_{U|y}+P_U}{2}\right]$. To check the assumption (b), we have $f'(t) = \log\frac{2t}{t+1}$, $f''(t) = \frac{1}{t(t+1)} \leq \frac{1}{\epsilon(\epsilon+1)}$, and thus it is Lipschitz continuous.

In the third example, consider the strictly convex function $f(t) = (1-t)^2/(2t+2)$, which leads to the Le Cam divergence [179] as the privacy measure, $\mathbb{E}_{Y,U}[d(y,u)] = \sum_y p(y) LC[P_{U|y} \parallel P_U]$, in which

$$LC[P_{U|y} \parallel P_U] = \frac{1}{2}\sum_u \frac{[p(u) - p(u|y)]^2}{p(u|y) + p(u)}. \tag{3.26}$$

For this choice of f, again, assumptions (b) and (c) are satisfied.

In the fourth example, we consider the following function $f(t) = (1-\sqrt{t})^2$, which corresponds to the squared Hellinger distance [180]. It is easy to check that the assumptions are satisfied.

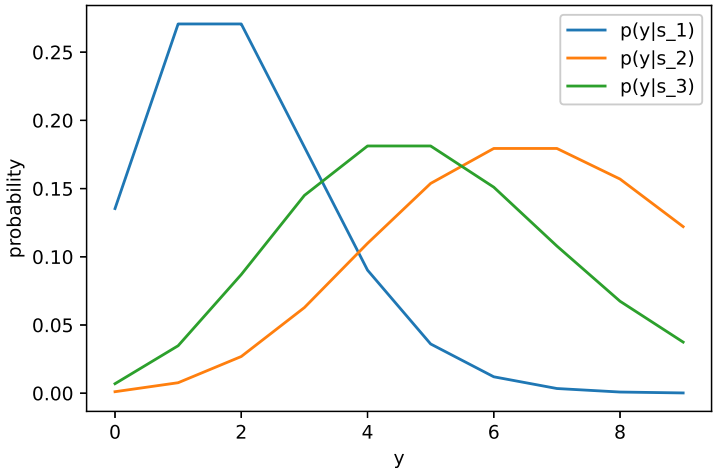


Figure 3.1: Conditional distribution $p(y|s)$

3.3.2 Numerical results

In this subsection, we provide numerical examples to show that our methods converge much faster than GA, and the solution found by our methods has much better quality than the one found by GA. Moreover, we explore how the weight parameter β and the alphabet size of $\mathcal{U}$ affects the privacy protection as well as the inference accuracy.

In the first example, we set the prior distribution $P_S = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ and let $|\mathcal{Y}| = 10, |\mathcal{U}| = 12$. The conditional distributions $P_{Y|S}$ under each s are shown in Fig. 3.1. Under this setup, we will perform both Algorithm 3.1 and GA to find the transition mapping $P_{U|Y}$ that maximizes the functional defined in (3.1). Suppose that the trade-off parameter $\beta = 2$ and Jensen-Shannon divergence is used as the privacy metric. The initial mapping $P_{U|Y}$ is obtained by selecting random numbers conforming to uniform distribution and normalizing them.

For the convergence speed, we investigate the relationship between $\mathcal{F}$ and the outer iteration, which is illustrated in Fig 3.2. We notice that the function value is increasing and converges as the iterative process progresses. For comparison purposes, we also plot the corresponding figures for GA in Fig. 3.3 (with step size 0.0001) and Fig. 3.4 (with step size 0.00005). From these figures, we can see that Algorithm 1 converges within 20 iterations. On the other hand, for gradient ascent

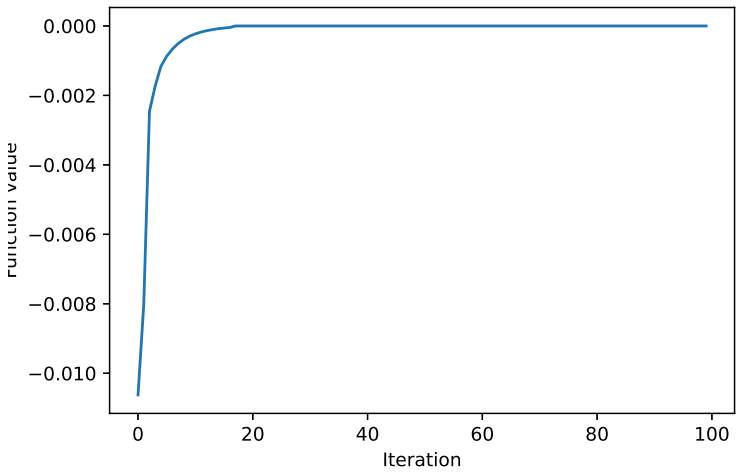


Figure 3.2: Function value v.s. iteration (Algorithm 3.1)

algorithm, even for a pretty small step size 0.0001, the function value fails to keep increasing, which indicates that the step size is too large. Then for a smaller step size 0.00005, the function value converges as shown in Fig. 3.4. However, the value of the objective function found by GA is smaller than the value found by Algorithm 3.1.

For the relationship between β and the privacy protection, after random initialization, we run Algorithm 3.1 and GA until they terminate. The stopping criterion is either $||P_{U|Y}^{t+1} - P_{U|Y}^{t}||_F < 10^{-5}$ (convergence case) or a maximum number of iterations is reached (divergence case). We repeat this procedure 100 times for each β. Recall that the smaller the term $\mathbb{E}[d(y,u)]$, the better the privacy protection. In particular, we set $\mathbb{E}[d(y,u)]$ to be 1 for divergence cases since the maximum $\mathbb{E}[d(y,u)]$ under the converge scenario is smaller than 1. As shown in Fig. 3.5, we notice that $\mathbb{E}[d(y,u)]$ decreases as β increases for our proposed method while it is non-decreasing for GA. By setting the maximum number of iterations to be 3000, GA diverges under many choices of β. Even for the scenarios where GA converges, compared with Algorithm 3.1, the privacy protection obtained by GA is weaker. Therefore, the privacy-preserving mapping designed by GA could hardly guarantee the protection of privacy. In addition, we also explore the relationship between β and the information accuracy. As shown in Fig. 3.6, the inference accuracy measure $I(S;U)$ decreases as β increases, which indicates that the predictive ability becomes weaker. The reason is that as

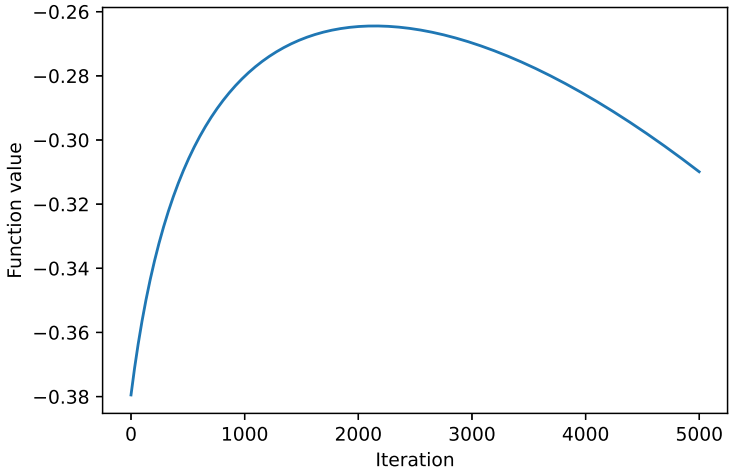


Figure 3.3: Function value v.s. iteration (GA)

U leaks less information about Y when β increases, it also provides less information about the parameter of interest, which will reduce the predictive performance. However, Fig. 3.6 shows that the reduction of $I(S;U)$ is not very large, which implies that the model still has good predictive ability when there are stronger protections for privacy.

To explore other privacy measures, we now set f as $f(t) = (1-t)^2/(2t+2)$, which corresponds to the Le Cam divergence as discussed in Section 3.3.1. We again compare Algorithm 3.1 and GA. The results are shown in Table 3.1. From the table, we can see that the maximum function value found by our method is greater than those found by GA.

Methods	**Convergent value**
Algorithm 3.1	-6.697e-14
Gradient ascent($\alpha = 0.05$)	-0.251
Gradient ascent($\alpha = 0.07$)	-0.245
Gradient ascent($\alpha = 0.1$)	-0.317
Gradient ascent($\alpha = 0.15$)	-0.235
Gradient ascent($\alpha = 0.2$)	Diverge

Table 3.1: Convergent value of Algorithm 3.1 and GA

To compare different privacy measures, we set the trade-off parameter $\beta = 8$, which indicates that the privacy term is dominant in the objective function. As shown in Fig. 3.7, although the function values under JS-divergence and LC-divergence are different, the convergence speed and

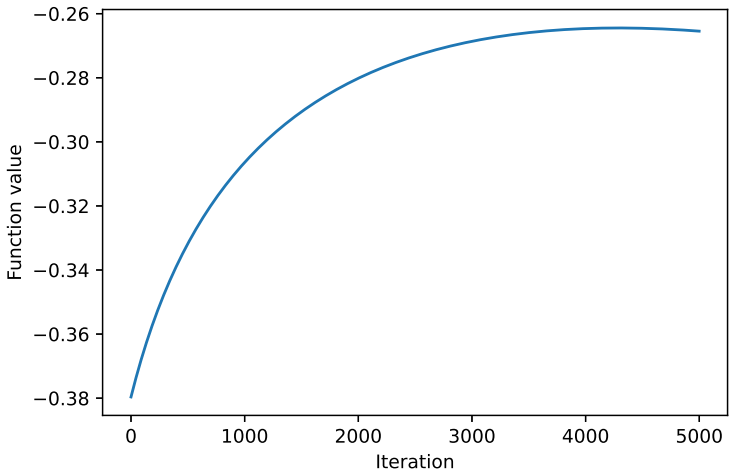


Figure 3.4: Function value v.s. iteration (GA)

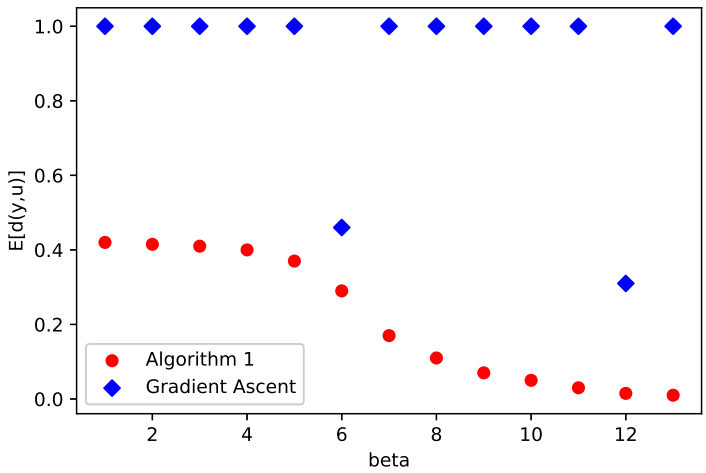


Figure 3.5: β v.s. privacy protection (Algorithm 3.1 and GA)

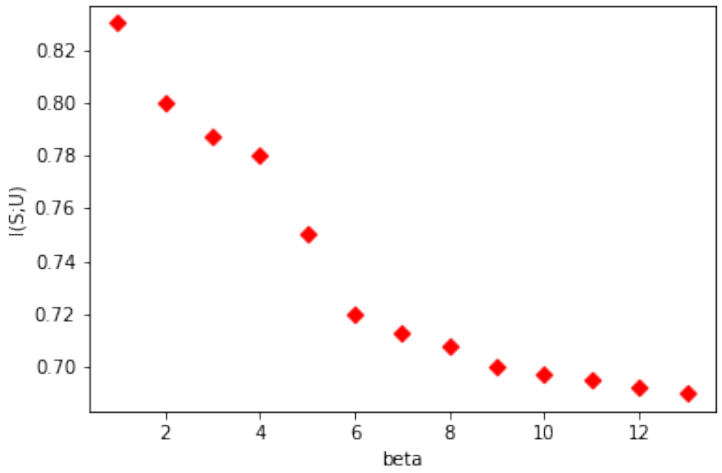


Figure 3.6: β v.s. inference accuracy (Algorithm 3.1)

convergence curve are almost the same, which shows that the proposed algorithm can converge in a similar manner under different metrics. However, the optimal privacy-preserving mapping $P_{U|Y}$ found by those two privacy measures are different. Therefore, in practical applications, an appropriate task-oriented privacy measure needs to be chosen.

Finally, we explore the relationship between $|\mathcal{U}|$ and the privacy protection. Note that in the proposed method, the alphabet sizes of $\mathcal{Y}$ and $\mathcal{U}$ are not necessarily equal. Thus, for $|\mathcal{Y}| = 10$, we explore how $|\mathcal{U}|$ affects the convergent function value. Here, we set $\beta = 8$ and use the LC-divergence to measure the privacy leakage. From Fig. 3.8, it is shown that although the function value is increasing as $|\mathcal{U}|$ increases, the alphabet size $|\mathcal{U}|$ has limited effects on the function value when $|\mathcal{U}| \geq 7$, which indicates that a large alphabet size of $\mathcal{U}$ is not necessary to derive a satisfactory privacy-preserving mapping. By setting $|\mathcal{Y}|$ to different values, we notice that when $\frac{|\mathcal{U}|}{|\mathcal{Y}|} \geq 0.8$, the convergent function value is relatively large.

3.4 Conclusion

We have proposed a general framework to design privacy-preserving mapping to achieve privacy-accuracy trade-off in the IAS scenarios. We have formulated optimization problems to find the

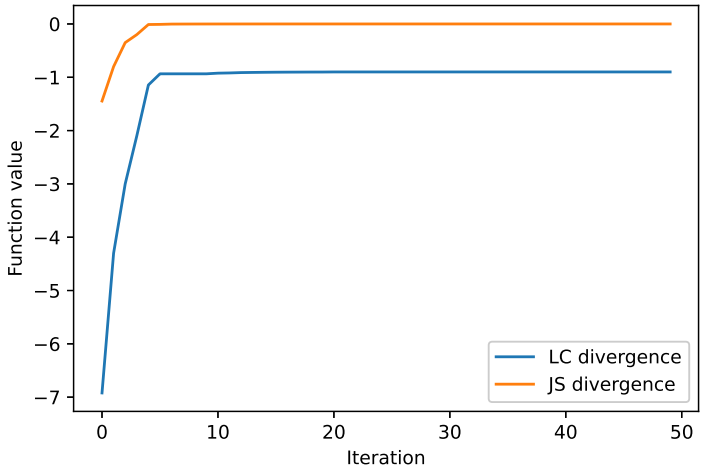


Figure 3.7: Convergence process for JS and LC divergences (Algorithm 3.1)

desirable mapping. We have discussed the structure of the formulated problems and designed an iterative method to solve these complicated optimization problems. We have also proved the convergence of the proposed method under certain assumptions. Moreover, we have provided numerical results showing that this method has better performance than GA in the convergence speed, solution quality and algorithm stability.

In terms of future work, we will address the limitations of the currently work along the following lines. Firstly, we have several technical assumptions on the function f. In the future, we will try to weaken those assumptions. Secondly, for the proposed algorithm, we are only able to show the convergence, but we have not characterized the convergence rate, of the proposed algorithms. Moreover, the proposed method is only guaranteed to converge but not to the global optima. Thus, it is of interest to further modify the proposed method to find the global optimal solution and determine the corresponding convergence rate. Thirdly, we are also interested in comparing our proposed privacy protection scheme with other existing private mechanisms. Finally, in this work, we only consider the case when Y is discrete and generate the privacy-preserving mapping $P_{U|Y}$. In the future, we will consider the continuous case and find the optimal conditional pdf $f_{U|Y}$.

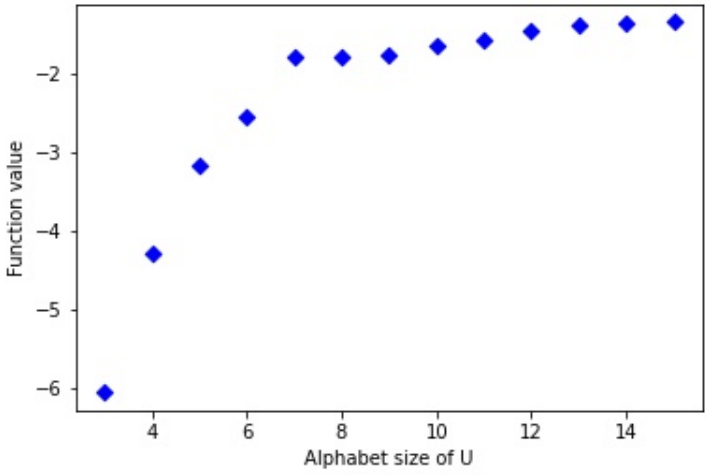


Figure 3.8: Function value v.s. Alphabet size of $\mathcal{U}$ (Algorithm 3.1)

Chapter 4

Fairness-aware Regression Robust to Adversarial Attacks

In this chapter, we take a first step towards answering the question of how to design fair machine learning algorithms that are robust to adversarial attacks. Using a minimax framework, we aim to design an adversarially robust fair regression model that achieves optimal performance in the presence of an attacker who is able to add a carefully designed adversarial data point to the dataset or perform a rank-one attack on the dataset. By solving the proposed nonsmooth nonconvex-nonconcave minimax problem, the optimal adversary as well as the robust fairness-aware regression model are obtained. For both synthetic data and real-world datasets, numerical results illustrate that the proposed adversarially robust fair models have better performance on poisoned datasets than other fair machine learning models in both prediction accuracy and group-based fairness measure.

In Section 4.1, we summarize the related work. In Section 4.2, we investigate the case when the adversary is allowed to add a poisoned data point into the dataset. In Section 4.3, we consider a more powerful adversary who is able to perform a rank-one attack on the dataset. In Section 4.4, we present numerical results. Finally, we offer concluding remarks in Section 4.5.

The main notations used in the chapter are listed in Table 4.1.

Notation	Description
$\{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{G}\}$	Clean dataset
$\{\hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}}\}$	Poisoned dataset
$(\boldsymbol{x}_0, y_0, G_0)$	Adversarial data point
$\boldsymbol{\Delta}$	rank-one feature modification matrix
β	Regression coefficient
η	Energy constraint parameter
λ	Trade-off parameter
n	Number of training samples
m	Number of training samples from group 1

Table 4.1: Main notations

4.1 Related work

Adversarial attacks on Fair Machine Learning (FML). There are many research works exploring the design of adversarial examples to reduce the testing accuracy and fairness of FML models. For example, [83] develops a gradient-based poisoning attack, [84] presents anchoring attack and influence attack, [85] provides three online attacks based on different group-based fairness measures, and [86] shows that adversarial attacks can worsen the model's fairness gap on test data while satisfying the fairness constraint on training data.

Adversarial robustness. A large variety of methods have been proposed to improve the model robustness against adversarial attacks [181–184]. Although promising to improve the model's robustness, those adversarial training algorithms have been observed to result in a large disparity of accuracy and robustness among different classes while natural training does not [185].

Intersection of fairness and robustness. Fairness and robustness are critical elements of trustworthy AI that need to be addressed together [147]. Firstly, in the field of adversarial training, several research works are proposed to interpret the accuracy/robustness disparity phenomenon and to mitigate the fairness issue [147–149]. For example, [148] presents an adversarially-trained neural network that is closer to achieve some fairness measures than the standard model on the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset. Secondly, a class-wise loss re-weighting method is shown to obtain more fair standard and robust classifiers [150]. Moreover, [151] and [152] argue that traditional notions of fairness are not sufficient when

the model is vulnerable to adversarial attacks, investigate the class-wise robustness and propose methods to improve the robustness of the most vulnerable class, so as to obtain a fairer robust model.

4.2 Attack with one adversarial data point

In this section, we consider the scenario where the attacker can add one carefully designed adversarial data point to the existing dataset.

4.2.1 Problem formulation

Using a set of training samples $\{\boldsymbol{x}_i, y_i, G_i\}_{i=1}^n \coloneqq \{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{G}\}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the feature vector, y_i is the response variable and G_i indicates the group membership or sensitive status (for example, race, gender), we aim to develop a model that can predict the value of a target variable Y from the input variables X. In this chapter, we consider the case when there are only two groups, i.e., $G_i \in \{1, 2\}$ and assume that the first m training samples are from group 1 and the remaining samples are from group 2. For simplification, we denote $\boldsymbol{X} = [\boldsymbol{X}_1; \boldsymbol{X}_2], \boldsymbol{y} = [\boldsymbol{y}_1; \boldsymbol{y}_2]$.

To build a robust model, we assume that there is an adversary who can observe the whole training dataset and then carefully design an adversarial data point, $\{\boldsymbol{x}_0, y_0, G_0\}$, and add it into the existing dataset. After inserting this poisoned data point, we have the poisoned dataset $\{\hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}}\}$, where $\hat{\boldsymbol{X}} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T, \hat{\boldsymbol{y}} = [y_0, y_1, \cdots, y_n]^T, \hat{\boldsymbol{G}} = [G_0, G_1, \cdots, G_n]^T$. From this poisoned dataset, we aim to design a robust fairness-aware regression model.

In order to characterize both prediction and fairness performance, we consider the following objective function

$$L = f(\boldsymbol{\beta}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}}) + \lambda F(\boldsymbol{\beta}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}}), \tag{4.1}$$

where $\boldsymbol{\beta}$ is the regression coefficient, $f(\boldsymbol{\beta}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}})$ corresponds to the prediction accuracy loss,

$F(\boldsymbol{\beta}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}})$ corresponds to the group fairness gap and λ is the trade-off parameter. The goal of the adversary is to maximize (4.1) to make the model less fair and less accurate, while the robust fairness-aware regression model aims at minimizing (4.1). To make the problem meaningful, we introduce an energy constraint on the adversarial data point and use ℓ_2 norm to measure the energy. The energy constraint serves two main purposes. Firstly, it helps to prevent certain kinds of easily detectable adversarial data points, which have a large energy and can be identified as outliers. Secondly, the energy constraint is essential for the MSE-based accuracy and fairness metrics since the absence of energy constraints can significantly affect the MSE value, thereby reducing the significance of the analysis. Thus, we have the minimax problem

$$\min_{\boldsymbol{\beta}} \max_{\substack{(\boldsymbol{x}_0, y_0, G_0), \\ \text{s.t. } \|[\boldsymbol{x}_0^T, y_0]\|_2 \leq \eta}} L = f(\boldsymbol{\beta}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}}) + \lambda F(\boldsymbol{\beta}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}}). \tag{4.2}$$

Given that Mean Squared Error (MSE) is the standard error metric for regression tasks, we leverage it to quantify the predictive accuracy of our model, and have $f(\boldsymbol{\beta}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}}) = \mathbb{E}[(Y - \hat{Y})^2]$, where $\hat{Y}$ is the prediction result. For the group fairness gap, various metrics have been proposed, including demographic parity [116], equality of opportunity [82], equalized odds [82], and accuracy parity [186]. While many of these metrics are well-suited for classification problems, they may not be directly applicable to regression problems. However, accuracy parity stands out as a fairness criterion that remains relevant across both classification and regression contexts. In particular, accuracy parity focuses on achieving equal accuracy losses among different groups [186], $\mathbb{E}[(Y - \hat{Y})^2 | G = 1] = \mathbb{E}[(Y - \hat{Y})^2 | G = 2]$. Then the absolute difference between two groups can be used to measure the severity of violations [187] and we have $F(\boldsymbol{\beta}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}}) = |\mathbb{E}[(Y - \hat{Y})^2 | G = 1] - \mathbb{E}[(Y - \hat{Y})^2 | G = 2]|$.

4.2.2 Proposed method

To solve the minimax problem in (4.2), we will first solve the inner maximization problem with respect to the adversary to design the optimal adversarial data point $\{\boldsymbol{x}_0, y_0, G_0\}$ under the energy

constraint. Then we will solve the outer minimization problem to find a robust fairness-aware model that can optimize both prediction accuracy and the group fairness guarantee.

Maximization Problem

We first note that there are two choices of G_0, and the form of the objective function L under different choices of G_0 is different. For $G_0 = 1$, the objective function L can be written as

$$L_1 = \frac{1}{n+1}\left(\|y_0 - \boldsymbol{x}_0^T\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2\right) + \lambda\left|\frac{1}{m+1}\|y_0 - \boldsymbol{x}_0^T\boldsymbol{\beta}\|_2^2 + \frac{1}{m+1}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 - \frac{1}{n-m}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2\right|.$$

For $G_0 = 2$, the objective function L can be written as

$$L_2 = \frac{1}{n+1}\left(\|y_0 - \boldsymbol{x}_0^T\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2\right) + \lambda\left|\frac{1}{m}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 - \frac{1}{n-m+1}\|y_0 - \boldsymbol{x}_0^T\boldsymbol{\beta}\|_2^2 - \frac{1}{n-m+1}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2\right|.$$

It is worth noting that for either L_1 or L_2, the objective function of the minimax problem (4.2) is non-smooth noncovex-nonconcave. However, we observe that by exploring four different cases depending on the value of G_0 and the signs of the terms inside $|\cdot|$, the maximization problem can be solved exactly as shown in the following theorem.

Theorem 5. *For any given $\boldsymbol{\beta}$, we have*

$$\max_{\substack{(\boldsymbol{x}_0, y_0, G_0),\\ \text{s.t. } \|[\boldsymbol{x}_0^T, y_0]\|_2 \le \eta}} L \overset{(a)}{=} \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta}), g_2(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\},$$

where

$$\begin{aligned} g_1(\boldsymbol{\beta}) &= C_{g_1}\eta^2(1+\|\boldsymbol{\beta}\|_2^2) + C_{g_1}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 + D_{g_1}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2, \\ h_1(\boldsymbol{\beta}) &= \max\{0, C_{h_1}\}\eta^2(1+\|\boldsymbol{\beta}\|_2^2) + C_{h_1}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 + D_{h_1}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2, \\ g_2(\boldsymbol{\beta}) &= \max\{0, D_{g_2}\}\eta^2(1+\|\boldsymbol{\beta}\|_2^2) + C_{g_2}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 + D_{g_2}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2, \\ h_2(\boldsymbol{\beta}) &= D_{h_2}\eta^2(1+\|\boldsymbol{\beta}\|_2^2) + C_{h_2}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 + D_{h_2}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2, \end{aligned}$$

with $C_{g_1} = \frac{\lambda}{m+1} + \frac{1}{n+1}, D_{g_1} = -\frac{\lambda}{n-m} + \frac{1}{n+1}, C_{h_1} = -\frac{\lambda}{m+1} + \frac{1}{n+1}, D_{h_1} = \frac{\lambda}{n-m} + \frac{1}{n+1}, C_{g_2} = \frac{\lambda}{m} + \frac{1}{n+1}, D_{g_2} = -\frac{\lambda}{n-m+1} + \frac{1}{n+1}, C_{h_2} = -\frac{\lambda}{m} + \frac{1}{n+1}, D_{h_2} = \frac{\lambda}{n-m+1} + \frac{1}{n+1}$. *Denote* $\tilde{\boldsymbol{x}}_0 = [\boldsymbol{x}_0^T, y_0]^T, \boldsymbol{b} = [\boldsymbol{\beta}^T, -1]^T$. *Then we have*

- *when either of the following occurs: 1)* $g_1(\boldsymbol{\beta}) \geq \max\{h_1(\boldsymbol{\beta}), g_2(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}$*, 2)* $h_1(\boldsymbol{\beta}) \geq \max\{g_1(\boldsymbol{\beta}), g_2(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}$ *and* $C_{h_1} \geq 0$*, the maximum value of* L *(equality (a)) is achieved if* $\tilde{\boldsymbol{x}}_0^*(\boldsymbol{\beta}) = \eta\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_2}$ *and* $G_0 = 1$*;*

- *when* $h_1(\boldsymbol{\beta}) \geq \max\{g_1(\boldsymbol{\beta}), g_2(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}$ *and* $C_{h_1} < 0$*, (a) is attained as long as* $\tilde{\boldsymbol{x}}_0^*(\boldsymbol{\beta}) \perp \boldsymbol{b}$ *and* $G_0 = 1$*;*

- *when either of the following occurs: 1)* $g_2(\boldsymbol{\beta}) \geq \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}$ *and* $D_{g_2} \geq 0$*, 2)* $h_2(\boldsymbol{\beta}) \geq \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta}), g_2(\boldsymbol{\beta})\}$*, (a) is attained if* $\tilde{\boldsymbol{x}}_0^*(\boldsymbol{\beta}) = \eta\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_2}$ *and* $G_0 = 2$*;*

- *when* $g_2(\boldsymbol{\beta}) \geq \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}$ *and* $D_{g_2} < 0$*, (a) is attained if* $\tilde{\boldsymbol{x}}_0^*(\boldsymbol{\beta}) \perp \boldsymbol{b}$ *and* $G_0 = 2$*.*

Proof. Please refer to Appendix C.1. □

Remark 2. $g_1(\boldsymbol{\beta})$*,* $h_1(\boldsymbol{\beta})$*,* $g_2(\boldsymbol{\beta})$ *and* $h_2(\boldsymbol{\beta})$ *involve* $\boldsymbol{\beta}$ *only through* $\|\boldsymbol{\beta}\|_2^2$*,* $\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2$ *and* $\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2$*. Furthermore, from Theorem 5, for* $G_0 = 1$*, we have*

$$\max_{(\boldsymbol{x}_0, y_0, 1), s.t. \|[\boldsymbol{x}_0^T, y_0]\|_2 \leq \eta} L_1 = \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta})\},$$

where $g_1(\boldsymbol{\beta})$ corresponds to the case in which the terms inside $|\cdot|$ of L_1 is non-negative and $h_1(\boldsymbol{\beta})$ corresponds to the case in which the terms inside $|\cdot|$ is negative. Subsequently, for the conditions of equality, we discuss two cases $L_1 = g_1(\boldsymbol{\beta}) \geq h_1(\boldsymbol{\beta})$ and $L_1 = h_1(\boldsymbol{\beta}) > g_1(\boldsymbol{\beta})$, where there are two sub-cases for $L_1 = h_1(\boldsymbol{\beta})$ based on the value of C_{h_1}. There are similar observations for $G_0 = 2$.

Minimization Problem

Using Theorem 5, the original minmax problem is converted to the following problem

$$\min_{\boldsymbol{\beta}} \max_{(\boldsymbol{x}_0, y_0, G_0)} L = \min_{\boldsymbol{\beta}} \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta}), g_2(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}. \tag{4.3}$$

As we seek to minimize the largest of four functions, (4.3) can be separated into four sub-problems. One of them is

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \ g_1(\boldsymbol{\beta}), \\ &\text{s.t.} \ \ g_1(\boldsymbol{\beta}) \geq g_2(\boldsymbol{\beta}), g_1(\boldsymbol{\beta}) \geq h_1(\boldsymbol{\beta}), g_1(\boldsymbol{\beta}) \geq h_2(\boldsymbol{\beta}), \end{aligned} \tag{4.4}$$

and other sub-problems can be written in a similar manner. Once these sub-problems are solved, the solution to (4.3) can be obtained.

For notation simplicity, we denote $\frac{1}{2}\frac{\partial^2 g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = C_{g_1}(\eta^2 \boldsymbol{I} + \boldsymbol{X}_1^T \boldsymbol{X}_1) + D_{g_1} \boldsymbol{X}_2^T \boldsymbol{X}_2 \coloneqq \boldsymbol{M_{g_1}}$, $\frac{1}{2}\frac{\partial^2 h_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \max\{0, C_{h_1}\}\eta^2 \boldsymbol{I} + C_{h_1} \boldsymbol{X}_1^T \boldsymbol{X}_1 + D_{h_1} \boldsymbol{X}_2^T \boldsymbol{X}_2 \coloneqq \boldsymbol{M_{h_1}}$, $\frac{1}{2}\frac{\partial^2 g_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \max\{0, D_{g_2}\}\eta^2 \boldsymbol{I} + C_{g_2} \boldsymbol{X}_1^T \boldsymbol{X}_1 + D_{g_2} \boldsymbol{X}_2^T \boldsymbol{X}_2 \coloneqq \boldsymbol{M_{g_2}}$, $\frac{1}{2}\frac{\partial^2 h_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = C_{h_2} \boldsymbol{X}_1^T \boldsymbol{X}_1 + D_{h_2}(\eta^2 \boldsymbol{I} + \boldsymbol{X}_2^T \boldsymbol{X}_2) \coloneqq \boldsymbol{M_{h_2}}$.

In the following, we focus on solving (4.4). The analysis of other sub-problems can be done similarly. Specifically, (4.4) can be further written as

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \quad g_1(\boldsymbol{\beta}) = C_{g_1}\eta^2(1 + \|\boldsymbol{\beta}\|_2^2) + C_{g_1}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 + D_{g_1}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2, \\ &\text{s.t.} \qquad C_1(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - h_1(\boldsymbol{\beta}) \geq 0, C_2(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - g_2(\boldsymbol{\beta}) \geq 0, C_3(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - h_2(\boldsymbol{\beta}) \geq 0. \end{aligned} \tag{4.5}$$

For the objective function in (4.5), since D_{g_1} can be negative, $\boldsymbol{M_{g_1}}$ is not necessarily positive-

semidefinite. Hence, (4.5) is a non-convex quadratic minimization problem with several quadratic constraints (QCQP), which is NP hard in general [158]. Despite this challenge, we are able to solve this problem by exploiting the structure inherent to our problem. The following proposition gives us sufficient conditions for global minimizers of QCQP, following from *Proposition 3.2* in [188].

Proposition 6. *If* $\exists \alpha_i \geq 0, i = 1, 2, 3$ *such that for* $\boldsymbol{\beta} = \boldsymbol{\beta}^*$*,*

$$\begin{aligned}
&\boldsymbol{M}_{\boldsymbol{g_1}} - \sum_{i=1}^{3} \alpha_i \frac{\partial^2 C_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \succeq \mathbf{0}, \\
&\frac{\partial g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}^*} - \sum_{i=1}^{3} \alpha_i \frac{\partial C_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}^*} = \mathbf{0}, \\
&\sum_{i=1}^{3} \alpha_i C_i(\boldsymbol{\beta}^*) = 0, \\
&C_i(\boldsymbol{\beta}^*) \geq 0, i = 1, 2, 3,
\end{aligned} \tag{4.6}$$

then $\boldsymbol{\beta}^*$ *is a global minimizer of QCQP* (4.5)*.*

Remark 3. *From* (4.6)*, we have that for each constraint* $C_i(\boldsymbol{\beta})$*, there are two possible cases: 1)* $\alpha_i = 0, C_i(\boldsymbol{\beta}^*) \geq 0$*; 2)* $\alpha_i > 0, C_i(\boldsymbol{\beta}^*) = 0$*. In total, there will be* 2^3 *cases of different combinations of* α_i*s. By examining these* 8 *different cases, we can obtain the optimal regression coefficient* $\boldsymbol{\beta}^*$ *of the sub-problem* (4.5)*.*

In the following, we will analyze four types of cases sequentially: 1) $\alpha_1 = \alpha_2 = \alpha_3 = 0$; 2) the case with only one non-zero α_i, i.e. $\exists! \alpha_i > 0$ and $\alpha_k = 0, \forall k \neq i$; 3) the case with two non-zero α_is, i.e. $\exists i, j, i \neq j, \alpha_i > 0, \alpha_j > 0$ and $\alpha_k = 0, \forall k \notin \{i, j\}$; 4) $\alpha_i > 0, i = 1, 2, 3$.

Case 1: $\alpha_1 = \alpha_2 = \alpha_3 = 0$

By Proposition 6, if there exists $\tilde{\boldsymbol{\beta}}$, such that

$$\boldsymbol{M}_{\boldsymbol{g_1}} \succeq \mathbf{0}, \tag{4.7}$$

$$\frac{\partial g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\tilde{\boldsymbol{\beta}}} = \mathbf{0}, \tag{4.8}$$

$$C_i(\tilde{\boldsymbol{\beta}}) \geq 0, i = 1, 2, 3, \tag{4.9}$$

then $\tilde{\boldsymbol{\beta}}$ is a global minimizer of (4.5). From (4.7), we require that $\boldsymbol{M_{g_1}}$ is positive-semidefinite, which can be true when λ is small, e.g. when $D_{g_1} \geq 0$. From (4.8), when $\boldsymbol{M_{g_1}}$ is invertible, we have

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{M}_{\boldsymbol{g_1}}^{-1} \left[C_{g_1} \boldsymbol{X}_1^T \boldsymbol{Y}_1 + D_{g_1} \boldsymbol{X}_2^T \boldsymbol{Y}_2 \right]. \tag{4.10}$$

If (4.9) is satisfied at (4.10), then $\tilde{\boldsymbol{\beta}}$ is a global minimizer of (4.5). Otherwise, there does not exist a global minimizer in *Case 1* and we will consider *Case 2*.

Case 2: $\exists! \alpha_i > 0$ *and* $\alpha_k = 0, \forall k \neq i$

We will consider the particular case $\alpha_1 > 0, \alpha_2 = \alpha_3 = 0$ and other cases can be analyzed similarly.

By Proposition 6, if there exists $\bar{\boldsymbol{\beta}}$ and $\alpha_1 > 0$, such that

$$\boldsymbol{M_{g_1}} - \alpha_1 (\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_1}}) \succeq 0, \tag{4.11}$$

$$\frac{\partial g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\bar{\boldsymbol{\beta}}} - \alpha_1 \frac{\partial C_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\bar{\boldsymbol{\beta}}} = \mathbf{0}, \tag{4.12}$$

$$C_1(\bar{\boldsymbol{\beta}}) = 0, C_2(\bar{\boldsymbol{\beta}}) \geq 0, C_3(\bar{\boldsymbol{\beta}}) \geq 0, \tag{4.13}$$

then $\bar{\boldsymbol{\beta}}$ is a global minimizer of (4.5).

Proposition 7. *Denote the largest eigenvalue of* $\boldsymbol{X}_1^T \boldsymbol{X}_1$ *as* $v_{X_1,p}$ *and the largest eigenvalue of* $\boldsymbol{X}_2^T \boldsymbol{X}_2$ *as* $v_{X_2,p}$*. Assuming that* $\eta^2 \geq \eta_{\min}^2 = \max\left\{ \frac{(n+1)v_{X_1,p}}{m(m+1)}, \frac{(n+1)v_{X_2,p}}{(n-m+1)(n-m)} \right\}$*, we have* $A_{g_1 h_1} = \{\alpha : \boldsymbol{M_{g_1}} - \alpha(\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_1}}) \succ 0\} \neq \emptyset$*. By randomly selecting an* $\alpha_1^* \in A_{g_1 h_1}$*, for*

$$\check{\boldsymbol{\beta}} = \left[(1 - \alpha_1^* - \gamma^*) \boldsymbol{M}_{g_1} + (\alpha_1^* + \gamma^*) \boldsymbol{M}_{h_1} \right]^{-1} \cdot \left[(1 - \alpha_1^* - \gamma^*) \boldsymbol{E}_{g_1} - (\alpha_1^* + \gamma^*) \boldsymbol{E}_{h_1} \right],$$

where γ^* *is a certain Lagrangian multiplier, and* $\boldsymbol{E}_{g_1} = C_{g_1} \boldsymbol{X}_1^T \boldsymbol{y_1} + D_{g_1} \boldsymbol{X}_2^T \boldsymbol{y_2}$, $\boldsymbol{E}_{h_1} = C_{h_1} \boldsymbol{X}_1^T \boldsymbol{y_1} + D_{h_1} \boldsymbol{X}_2^T \boldsymbol{y_2}$*, if we have* $C_2(\check{\boldsymbol{\beta}}) \geq 0, C_3(\check{\boldsymbol{\beta}}) \geq 0$*, then* $\check{\boldsymbol{\beta}}$ *satisfies* (4.11)*,* (4.12)*,* (4.13) *and is a global minimizer of* (4.5)*.*

Proof. Please refer to Appendix C.2. □

Case 3: $\exists i, j, i \neq j, \alpha_i > 0, \alpha_j > 0$ *and* $\alpha_k = 0, \forall k \notin \{i, j\}$

We will consider the particular case $\alpha_1 > 0, \alpha_2 > 0, \alpha_3 = 0$ and other cases can be analyzed in a similar manner. By Proposition 6, if there exists $\hat{\boldsymbol{\beta}}$ and $\alpha_1 > 0, \alpha_2 > 0$, such that

$$\boldsymbol{M_{g_1}} - \alpha_1(\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_1}}) - \alpha_2(\boldsymbol{M_{g_1}} - \boldsymbol{M_{g_2}}) \succeq 0, \tag{4.14}$$

$$\frac{\partial g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}} - \alpha_1 \frac{\partial C_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}} - \alpha_2 \frac{\partial C_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}} = \mathbf{0}, \tag{4.15}$$

$$C_1(\hat{\boldsymbol{\beta}}) = 0, C_2(\hat{\boldsymbol{\beta}}) = 0, \tag{4.16}$$

$$C_3(\hat{\boldsymbol{\beta}}) \geq 0, \tag{4.17}$$

then $\hat{\boldsymbol{\beta}}$ is a global minimizer of (4.5).

Proposition 8. *For*

$$\check{\boldsymbol{\beta}} = [(1 - \alpha_1^* - \gamma_1^* - \gamma_2^*)\boldsymbol{M}_{g_1} + (\alpha_1^* + \gamma_1^*)\boldsymbol{M}_{h_1} + \gamma_2^* \boldsymbol{M}_{g_2}]^{-1}$$
$$\cdot [(1 - \alpha_1^* - \gamma_1^* - \gamma_2^*)\boldsymbol{E}_{g_1} + (\alpha_1^* + \gamma_1^*)\boldsymbol{E}_{h_1} + \gamma_2^* \boldsymbol{E}_{g_2}],$$

where γ_1^*, γ_2^* *are certain Lagrangian multipliers, and* $\boldsymbol{E}_{g_2} = C_{g_2}\boldsymbol{X}_1^T\boldsymbol{y_1} + D_{g_2}\boldsymbol{X}_2^T\boldsymbol{y_2}$*, if* $C_3(\check{\boldsymbol{\beta}}) \geq 0$*, then* $\check{\boldsymbol{\beta}}$ *satisfies* (4.14)*,* (4.15)*,* (4.16)*,* (4.17) *and is a global minimizer of* (4.5)*.*

Proof. Please refer to Appendix C.3. □

Case 4: $\alpha_i > 0, i = 1, 2, 3$

By Proposition 6, if there exists $\acute{\boldsymbol{\beta}}$ and $\alpha_i > 0, i = 1, 2, 3$, such that

$$\boldsymbol{M_{g_1}} - \alpha_1(\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_1}}) - \alpha_2(\boldsymbol{M_{g_1}} - \boldsymbol{M_{g_2}}) - \alpha_3(\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_2}}) \succeq 0, \tag{4.18}$$

$$\frac{\partial g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\acute{\boldsymbol{\beta}}} - \alpha_1 \frac{\partial C_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\acute{\boldsymbol{\beta}}} - \alpha_2 \frac{\partial C_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\acute{\boldsymbol{\beta}}} - \alpha_3 \frac{\partial C_3(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\acute{\boldsymbol{\beta}}} = \mathbf{0}, \tag{4.19}$$

$$C_1(\acute{\boldsymbol{\beta}}) = 0, C_2(\acute{\boldsymbol{\beta}}) = 0, C_3(\acute{\boldsymbol{\beta}}) = 0, \tag{4.20}$$

then $\acute{\boldsymbol{\beta}}$ is a global minimizer of (4.5). From Remark 2, we note that with (4.20), there are three equations on $\|\boldsymbol{\beta}\|_2^2$, $\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2$ and $\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2$, which indicates that there will be deterministic

Algorithm 4.1 Attack with one adversarial point

Input: $\{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{G}\}, \lambda, \eta$.
Output: Optimal adversarial point $(\boldsymbol{x}_0^*, y_0^*, G_0^*)$. Robust fairness-aware regression coefficient $\boldsymbol{\beta}_{rob}^*$.
Procedure:

1: Separate (4.2) into 4 sub-problems.
2: **for** each sub-problem **do**
3: step through *Case 1* to *Case 4* until a $\boldsymbol{\beta}$ that satisfies the sufficient conditions in Proposition 6 is found.
4: Select $\boldsymbol{\beta}$ that minimizes $\max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta}), g_2(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}$ from the solution set as $\boldsymbol{\beta}_{rob}^*$.
5: Plug $\boldsymbol{\beta}_{rob}^*$ into the optimal attack strategy to obtain $(\boldsymbol{x}_0^*, y_0^*, G_0^*)$.

solutions for them or the feasible set is empty.

When the feasible set of (4.20) is nonempty (for example, when $\lambda > \max\{\frac{m+1}{n+1}, \frac{n-m+1}{n+1}\}$), the value of $g_1(\boldsymbol{\beta}), C_1(\boldsymbol{\beta}), C_2(\boldsymbol{\beta}), C_3(\boldsymbol{\beta})$ is determined as there have been deterministic solutions for $\|\boldsymbol{\beta}\|_2^2$, $\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2$ and $\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2$. Then the process of finding $\acute{\boldsymbol{\beta}}$ is

1. Solve (4.20) and derive the solution for $\|\boldsymbol{\beta}\|_2^2$, $\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2$ and $\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2$.

2. Calculate the value of $g_1(\boldsymbol{\beta}), C_1(\boldsymbol{\beta}), C_2(\boldsymbol{\beta}), C_3(\boldsymbol{\beta})$.

3. Select $\alpha_1, \alpha_2, \alpha_3$ such that (4.18) is satisfied. Then (4.19) is satisfied naturally as $g_1(\boldsymbol{\beta}), C_1(\boldsymbol{\beta}), C_2(\boldsymbol{\beta}), C_3(\boldsymbol{\beta})$ are constants.

Algorithm 4.1 summarizes the process of finding the robust fairness-aware model and the optimal adversary data point, which does not involve any transformation gap.

In comparison to [158], our proposed QCQP (4.5) is different from the setting considered in [158]. Notably, the objective function in (4.5) is nonconvex, and the constraints are not concave, which distinguishes our problem from the one in [158]. Moreover, while [158] provides approximation bounds for a norm minimization problem with multiple concave quadratic constraints, our objective is to find a global minimizer of (4.5) based on the sufficient conditions of global minimizers to QCQP, as stated in Proposition 6. Consequently, the methods proposed in [158] are not applicable to our problem.

4.3 Rank-one attack

In Section 4.2, we have discussed how to design one adversarial point to attack the fair regression model. In this section, we consider a more powerful adversary who can observe the whole training dataset and then perform a rank-one attack on the feature matrix. This type of attack covers many practical scenarios, for example, modifying one entry of the feature matrix, deleting one feature, changing one feature, replacing one feature, etc [184]. In particular, the attacker will carefully design a rank-one feature modification matrix $\boldsymbol{\Delta}$ and add it to the original feature matrix $\boldsymbol{X}$, so as to obtain the modified feature matrix $\hat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta}$. Since $\boldsymbol{\Delta}$ is of rank 1, we can write $\boldsymbol{\Delta} = \boldsymbol{c}\boldsymbol{d}^T$, where $\boldsymbol{c} \in \mathbb{R}^n$ and $\boldsymbol{d} \in \mathbb{R}^p$. Moreover, recall that there are samples from two groups, we denote the modification matrix of the first group as $\boldsymbol{\Delta}_1$, i.e., the first m rows of $\boldsymbol{\Delta}$, and assume that $\boldsymbol{\Delta}_1 = \boldsymbol{c}_1\boldsymbol{d}^T$, where $\boldsymbol{c}_1$ consists of the first m components of $\boldsymbol{c}$. Similarly, for the second group, the modification matrix is $\boldsymbol{\Delta}_2 = \boldsymbol{c}_2\boldsymbol{d}^T$. Then the modified feature matrices are $\hat{\boldsymbol{X}}_1 = \boldsymbol{X}_1 + \boldsymbol{\Delta}_1$ and $\hat{\boldsymbol{X}}_2 = \boldsymbol{X}_2 + \boldsymbol{\Delta}_2$.

Similar to Section 4.2, we introduce an energy constraint on the rank-one attack. We use the Frobenius norm to measure the energy of $\boldsymbol{\Delta}$. Recall that $\boldsymbol{y}, \boldsymbol{G}$ remain unchanged in this attack scheme, we have the minimax problem

$$\min_{\boldsymbol{\beta}} \max_{\|\Delta\|_F \leq \eta} f(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) + \lambda F(\boldsymbol{\beta}, \hat{\boldsymbol{X}}). \tag{4.21}$$

To solve (4.21), we will first investigate the inner maximization problem. We will perform various variable augmentations, and convert the maximization problem into a form with five arguments, four of which can be solved exactly. Then we will transform the original nonconvex-nonconcave minimax problem into several weakly-convex-weakly-concave minimax problems.

Maximization problem

For the objective function in (4.21), we have

$$f(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) + \lambda F(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = \frac{1}{n}\|\boldsymbol{y} - \hat{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2 + \lambda \left| \frac{1}{m}\|\boldsymbol{y}_1 - \hat{\boldsymbol{X}}_1\boldsymbol{\beta}\|_2^2 - \frac{1}{n-m}\|\boldsymbol{y}_2 - \hat{\boldsymbol{X}}_2\boldsymbol{\beta}\|_2^2 \right|$$
$$= \max\{g(\boldsymbol{\beta}, \hat{\boldsymbol{X}}), h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})\},$$

in which $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = C_g\|\boldsymbol{y}_1 - \hat{\boldsymbol{X}}_1\boldsymbol{\beta}\|_2^2 + D_g\|\boldsymbol{y}_2 - \hat{\boldsymbol{X}}_2\boldsymbol{\beta}\|_2^2$, $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = C_h\|\boldsymbol{y}_1 - \hat{\boldsymbol{X}}_1\boldsymbol{\beta}\|_2^2 + D_h\|\boldsymbol{y}_2 - \hat{\boldsymbol{X}}_2\boldsymbol{\beta}\|_2^2$, with $C_g = \frac{1}{n} + \frac{\lambda}{m}$, $D_g = \frac{1}{n} - \frac{\lambda}{n-m}$, $C_h = \frac{1}{n} - \frac{\lambda}{m}$, $D_h = \frac{1}{n} + \frac{\lambda}{n-m}$.

Lemma 12. *For $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ and $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$, we have that*

(1) if $D_g \geq 0$, $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ is convex in $\boldsymbol{c}_1$ for any given $\boldsymbol{c}_2, \boldsymbol{d}$, and also convex in $\boldsymbol{c}_2$ for any given $\boldsymbol{c}_1, \boldsymbol{d}$; otherwise, $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ is convex in $\boldsymbol{c}_1$ for any given $\boldsymbol{c}_2, \boldsymbol{d}$, and concave in $\boldsymbol{c}_2$ for any given $\boldsymbol{c}_1, \boldsymbol{d}$;

(2) if $C_h \geq 0$, $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ is convex in $\boldsymbol{c}_1$ for any given $\boldsymbol{c}_2, \boldsymbol{d}$, and also convex in $\boldsymbol{c}_2$ for any given $\boldsymbol{c}_1, \boldsymbol{d}$; otherwise, $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ is concave in $\boldsymbol{c}_1$ for any given $\boldsymbol{c}_2, \boldsymbol{d}$, and convex in $\boldsymbol{c}_2$ for any given $\boldsymbol{c}_1, \boldsymbol{d}$.

Based on Lemma 12, we now solve the maximization problem in (4.21). First, note that

$$\max_{\|\boldsymbol{c}\boldsymbol{d}^T\|_F \leq \eta} \max\{g(\boldsymbol{\beta}, \hat{\boldsymbol{X}}), h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})\} = \max\left\{ \max_{\|\boldsymbol{c}\boldsymbol{d}^T\|_F \leq \eta} g(\boldsymbol{\beta}, \hat{\boldsymbol{X}}), \max_{\|\boldsymbol{c}\boldsymbol{d}^T\|_F \leq \eta} h(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) \right\},$$

which indicates that the maximization problem can be separated into two sub-problems. For simplicity of presentation, we will only explore the sub-problem of $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ in detail and the sub-problem of $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ can be analyzed similarly.

1) Sub-problem of $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$

According to Lemma 12, the value of D_g will affect the property of $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$. In the following, we will first explore the case $D_g \geq 0$ and obtain Lemma 13 as well as Proposition 9, and then explore the case $D_g < 0$ and obtain Lemma 14 as well as Proposition 10.

Lemma 13. *For $D_g \geq 0$, we have*

$$\begin{aligned}
&\max_{\|\boldsymbol{c}\boldsymbol{d}^T\|_F \leq \eta} g(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) \\
&= \max_{0<\eta_c\leq\eta} \max_{0\leq\eta_{c_1}\leq\eta_c} \max_{\|\boldsymbol{d}\|_2\leq 1} \max_{\|\boldsymbol{c}_2\|_2=\sqrt{\eta_c^2-\eta_{c_1}^2}} \max_{\|\boldsymbol{c}_1\|_2=\eta_{c_1}} g(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) \\
&= \max_{0<\eta_c\leq\eta} \max_{0\leq\eta_{c_1}\leq\eta_c} \max_{\|\boldsymbol{d}\|_2\leq 1} g_{m_1}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d}),
\end{aligned}$$

where $g_{m_1}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d}) = C_g(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\boldsymbol{d}^T\boldsymbol{\beta})^2 + D_g(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta_c^2 - \eta_{c_1}^2}\boldsymbol{d}^T\boldsymbol{\beta})^2$.

Proof. Please refer to Appendix C.4. □

Note that $g_{m_1}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d})$ is a quadratic function with respect to $\boldsymbol{d}^T\boldsymbol{\beta}$, we have the following proposition.

Proposition 9.

$$\max_{\|\boldsymbol{c}\boldsymbol{d}^T\|_F \leq \eta} g(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = \max_{0\leq\eta_{c_1}\leq\eta} g_a(\eta_{c_1}, \boldsymbol{\beta}),$$

where $g_a(\eta_{c_1}, \boldsymbol{\beta}) = C_g(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 + D_g(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta^2 - \eta_{c_1}^2}\|\boldsymbol{\beta}\|_2)^2$.

Proof. Please refer to Appendix C.5. □

Lemma 14. *For $D_g < 0$, we have*

$$\max_{\|\boldsymbol{c}\boldsymbol{d}^T\|_F \leq \eta} g(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = \max_{0<\eta_c\leq\eta} \max_{0\leq\eta_{c_1}\leq\eta_c} \max_{\|\boldsymbol{d}\|_2\leq 1} g_{m_2}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d}),$$

where

$$g_{m_2}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d}) = \begin{cases} C_g(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\boldsymbol{d}^T\boldsymbol{\beta})^2, \textit{if } \|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 \leq \eta\|\boldsymbol{\beta}\|_2, \\ C_g(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\boldsymbol{d}^T\boldsymbol{\beta})^2 + D_g(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 - \sqrt{\eta_c^2 - \eta_{c_1}^2}\boldsymbol{d}^T\boldsymbol{\beta})^2, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{otherwise.} \end{cases}$$

Proof. Please refer to Appendix C.6. □

From the above lemma, we have the following proposition.

Proposition 10.

$$\max_{\|\boldsymbol{cd}^T\|_F \le \eta} g(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = \max_{0 \le \eta_{c_1} \le \eta} g_b(\eta_{c_1}, \boldsymbol{\beta}),$$

where

$$g_b(\eta_{c_1}, \boldsymbol{\beta}) = \begin{cases} g_{b_1}(\eta_{c_1}, \boldsymbol{\beta}), \text{ if } \|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 \le \eta\|\boldsymbol{\beta}\|_2, \\ g_{b_2}(\eta_{c_1}, \boldsymbol{\beta}), \text{ otherwise.} \end{cases}$$

$$g_{b_1}(\eta_{c_1}, \boldsymbol{\beta}) = C_g(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2,$$

$$g_{b_2}(\eta_{c_1}, \boldsymbol{\beta}) = \left[C_g(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 + D_g(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 - \sqrt{\eta^2 - \eta_{c_1}^2}\|\boldsymbol{\beta}\|_2)^2\right].$$

Proof. Please refer to Appendix C.7. □

2) Sub-problem of $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$

Following similar process in analyzing the sub-problem of $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$, we have that

- if $C_h \ge 0$, we have

  $$\max_{\|\boldsymbol{cd}^T\|_F \le \eta} h(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = \max_{0 \le \eta_{c_1} \le \eta} h_a(\eta_{c_1}, \boldsymbol{\beta}),$$

  where $h_a(\eta_{c_1}, \boldsymbol{\beta}) = C_h(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 + D_h(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta^2 - \eta_{c_1}^2}\|\boldsymbol{\beta}\|_2)^2$;

- if $C_h < 0$, we have

  $$\max_{\|\boldsymbol{cd}^T\|_F \le \eta} h(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = \max_{0 \le \eta_{c_1} \le \eta} h_b(\eta_{c_1}, \boldsymbol{\beta}),$$

  where

  $$h_b(\eta_{c_1}, \boldsymbol{\beta}) = \begin{cases} h_{b_1}(\eta_{c_1}, \boldsymbol{\beta}), \text{ if } \|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 \le \eta\|\boldsymbol{\beta}\|_2, \\ h_{b_2}(\eta_{c_1}, \boldsymbol{\beta}), \text{ otherwise,} \end{cases}$$

  $$h_{b_1}(\eta_{c_1}, \boldsymbol{\beta}) = D_h(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta^2 - \eta_{c_1}^2}\|\boldsymbol{\beta}\|_2)^2,$$

  $$h_{b_2}(\eta_{c_1}, \boldsymbol{\beta}) = C_h(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 - \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 + D_h(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta^2 - \eta_{c_1}^2}\|\boldsymbol{\beta}\|_2)^2.$$

Transformation of the minimax problem

After solving sub-problems above, the minimax problem (4.21) can be transformed to a minimax problem for one vector and one scalar with a piece-wise max-type objective function. For example, if $D_g \geq 0$ and $C_h < 0$, (4.21) can be represented as

$$\min_{\beta} \max_{0 \leq \eta_{c_1} \leq \eta} \max\{g_a(\eta_{c_1}, \boldsymbol{\beta}), h_b(\eta_{c_1}, \boldsymbol{\beta})\}. \tag{4.22}$$

Then we have the following two lemmas characterizing the nice properties of the sub-functions in the objective function.

Lemma 15. *If the norm of $\boldsymbol{\beta}$ is bounded, i.e. $\|\boldsymbol{\beta}\|_2 \leq B_\beta$, then we have*

(1) g_a is weakly-concave in η_{c_1} for any given $\boldsymbol{\beta}$ and weakly-convex in $\boldsymbol{\beta}$ for any given η_{c_1};

(2) h_b is a piece-wise function and each piece (h_{b_1} or h_{b_2}) is weakly-concave in η_{c_1} for any given $\boldsymbol{\beta}$ and weakly-convex in $\boldsymbol{\beta}$ for any given η_{c_1}.

Proof. Please refer to Appendix C.8. □

Lemma 16. *For any given $\boldsymbol{\beta}$, g_a, g_{b_2}, h_a and h_{b_2} are all unimodal functions with respect to η_{c_1} that increase first and then decrease.*

Proof. Please refer to Appendix C.9. □

Moreover, to deal with the piece-wise structure in the objective function, we further transform the minimax problem to several sub-problems. For example, (4.22) can be transformed to three sub-problems:

(1) $\min_{\beta} \max_{0 \leq \eta_{c_1} \leq \eta} h_{b_1}(\eta_{c_1}, \boldsymbol{\beta})$, s.t. $g_a(\eta_{c_1}, \boldsymbol{\beta}) < h_{b_1}(\eta_{c_1}, \boldsymbol{\beta})$, $\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 \leq \eta\|\boldsymbol{\beta}\|_2$;

(2) $\min_{\beta} \max_{0 \leq \eta_{c_1} \leq \eta} h_{b_2}(\eta_{c_1}, \boldsymbol{\beta})$, s.t. $g_a(\eta_{c_1}, \boldsymbol{\beta}) < h_{b_2}(\eta_{c_1}, \boldsymbol{\beta})$, $\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 > \eta\|\boldsymbol{\beta}\|_2$.

(3) $\min_{\beta} \max_{0 \leq \eta_{c_1} \leq \eta} g_a(\eta_{c_1}, \boldsymbol{\beta})$, s.t. $g_a(\eta_{c_1}, \boldsymbol{\beta}) \geq h_b(\eta_{c_1}, \boldsymbol{\beta})$.

For the sub-problem 1), the maximization on η_{c_1} can be solved exactly and the saddle-point can be easily derived.

For sub-problems 2) and 3), we will ignore the constraints first and derive the saddle-point of the minimax problem, and then check the constraints. For example, for sub-problem 2), we assume that $\|\boldsymbol{\beta}\|_2 \leq B_\beta$, which is reasonable in reality, and have that: the feasible set $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2 \leq B_\beta\} \times [0, \eta]$ is convex and compact; the objective function is weakly-convex-weakly-concave by Lemma 15; the saddle-point exists by Lemma 16. Based on those properties, we are able to apply a first-order algorithms proposed by [160] to solve the non-convex non-concave minimax problem as in sub-problem 2) and derive the nearly ϵ-stationary solution. In particular, define $\mathcal{Z} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2 \leq B_\beta\} \times [0, \eta]$ and the mapping $H(\boldsymbol{z}) \coloneqq (\partial_{\boldsymbol{\beta}} h_{b_2}(\eta_{c_1}, \boldsymbol{\beta}), \partial_{\eta_{c_1}}[-h_{b_2}(\eta_{c_1}, \boldsymbol{\beta})])^T$, where $\boldsymbol{z} = (\boldsymbol{\beta}, \eta_{c_1})$. The minty variational inequality (MVI) problem corresponding to the saddle-point problem in sub-problem 2) is to find $\boldsymbol{z}^* \in \mathcal{Z}$ such that $\langle \boldsymbol{\xi}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0, \forall \boldsymbol{z} \in \mathcal{Z}, \forall \boldsymbol{\xi} \in H(\boldsymbol{z})$. Then the saddle-point problem can be solved through the lens of MVI. In [160], the proposed inexact proximal point method consists of approximately solving a sequence of strongly monotone MVIs constructed by adding a strongly monotone mapping to $H(\boldsymbol{z})$ with a sequentially updated proximal center. Thus, the complex non-convex non-concave minmax problem can be decomposed into a sequence of easier strongly-convex strongly-concave problems.

In comparison to [160], our focus is on investigating fairness issues in predictive models while ensuring robustness, rather than a general analysis of weakly-convex-weakly-concave minimax problems. Consequently, the proposed transformation and corresponding analysis are crucial in our work. Furthermore, the problem setup here differs from that in [160]. In our transformed minimax problem (4.22), the objective function is a max-type function, which is not an exact weakly-convex-weakly-concave minimax problem.

4.4 Numerical Results

In this section, we provide numerical examples to illustrate the results in this chapter. We conduct experiments on a synthetic dataset and two real-world datasets:

1. **Synthetic Dataset (SD)**: it contains 200 rows for two groups with 5 features. We suppose that the numbers of samples in two groups are the same, i.e. $m = n - m = 100$. For two different groups, the samples are generated by

$$\boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_{0,1} + \boldsymbol{c}_1 + \boldsymbol{\epsilon},\ \boldsymbol{y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_{0,2} + \boldsymbol{\epsilon}, \tag{4.23}$$

where elements in $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are uniformly distributed on $(0, 10)$, $\boldsymbol{\beta}_{0,1} = [1, 1, 1, 1, 1]^T$, $\boldsymbol{c}_1 = [1, \cdots, 1]^T$, $\boldsymbol{\beta}_{0,2} = [1.1, 1.1, 1.1, 1.1, 1.1]^T$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$. Under this setup, we verify the assumption in Propositions 7 and have that $\eta^2 \geq \eta^2_{min} = \max\left\{\frac{(n+1)v_{X_1,p}}{m(m+1)}, \frac{(n+1)v_{X_2,p}}{(n-m+1)(n-m)}\right\} = 15.98^2$ while the mean energy of a sample is $\eta_D = 29.08$, which indicates that the assumption on η is reasonable.

2. **Law School Dataset (LSD)** [189]: it contains 1,823 records for law students who took the bar passage study for law school admission, with gender as the sensitive attribute and undergraduate GPA as the target variable. The dimension of features is 8. There are 999 samples and 824 samples for two genders respectively. For the assumption on η, we have $\eta \geq \eta_{min} = 2.44$ and $\eta_D = 2.86$.

3. **Medical Insurance Cost Dataset (MICD)** [190]: it contains 1,338 medical expense examples for patients in the United States. In our experiment, we use gender as the sensitive attribute, charged medical expenses as the target variable, and consider 5 features. There are 662 samples and 676 samples for two genders respectively. Then we verify the assumption on η and have that $\eta \geq \eta_{min} = 1.58$ with $\eta_D = 2.34$.

For comparison purpose, we will introduce an unrobust fair regression model that does not consider the existence of the adversary and minimizes the objective function with respect to the

original dataset $\{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{G}\}$. In particular, the unrobust fair model is

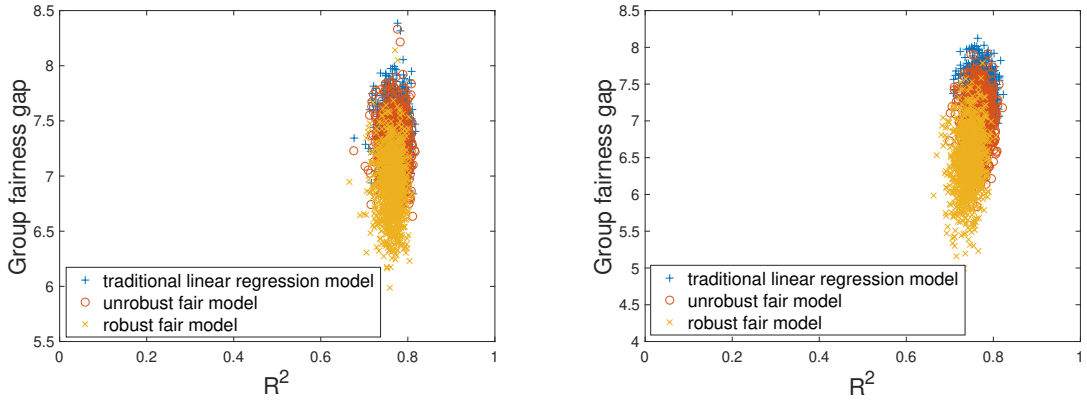
$$\boldsymbol{\beta}_{fair} = \arg\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{G}) + \lambda F(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{G}).$$

Moreover, for the rank-one attack scheme, we also compare our proposed adversarially robust model with other fair regression models, including the fair linear regression (FLR) model and fair kernel learning (FKL) model [191]. The optimal regression coefficient for each model is derived by fitting the model on the original dataset $\{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{G}\}$. To obtain the performance of each model on the poisoned dataset, we apply the derived optimal regression coefficient on the poisoned dataset, $\{\hat{\boldsymbol{X}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{G}}\}$, and calculate the MSE as well as the group fairness gap.

4.4.1 Attack with one adversarial data point

Firstly, for SD, by choosing $\eta = \eta_D$, we explore the performance differences among the proposed robust fairness-aware model, unrobust fair model and traditional linear model (ordinary linear regression model). In Fig. 4.1(a) and Fig. 4.1(b), following (4.23), we construct 500 datasets relying on the randomness in $\boldsymbol{\epsilon}$. For $\lambda = 0.2 < \min\{\frac{m+1}{n+1}, \frac{n-m+1}{n+1}\}$ (which implies $C_{h_1} \geq 0, D_{g_2} \geq 0$), according to Theorem 5, the best adversarial point is $\tilde{\boldsymbol{x}}_0 = \eta \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_2}$. As shown in Fig. 4.1(a), the group fairness gap for the proposed robust fairness-aware model is smaller than that of the unrobust fair model, while the measure of goodness of fit R^2 remains similar. In the meantime, since $\boldsymbol{\beta}_{fair}$ has taken the fairness issue into consideration, its performance is better than the traditional linear regression model. Likewise, for $\lambda = 0.8 > \max\{\frac{m+1}{n+1}, \frac{n-m+1}{n+1}\}$ (which implies $C_{h_1} < 0, D_{g_2} < 0$), according to Theorem 5, the best adversarial point will be in the form $\tilde{\boldsymbol{x}}_0 \perp \boldsymbol{b}$ or $\tilde{\boldsymbol{x}}_0 = \eta \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_2}$ based on the value of $g_i(\boldsymbol{\beta}^*_{rob})$ and $h_i(\boldsymbol{\beta}^*_{rob}), i = 1, 2$. As shown in Fig. 4.1(b), the performance results are similar to the case $\lambda = 0.2$.

Secondly, we explore the effects of the energy constraint parameter η as well as the trade-off parameter λ on two real-world datasets, LSD and MICD. We have three energy levels, $\eta = \eta_{min}$, $\eta = \eta_D$ and $\eta = 1.5\eta_D$. As shown in Fig. 4.2, when η is small, under different choices of λ,



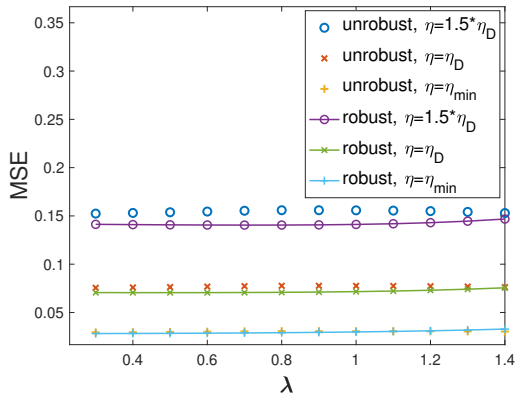
(a) $\lambda = 0.2$ (b) $\lambda = 0.8$

Figure 4.1: SD: comparison of robust fair model, unrobust fair model and traditional linear model (attack with one adversarial data point).

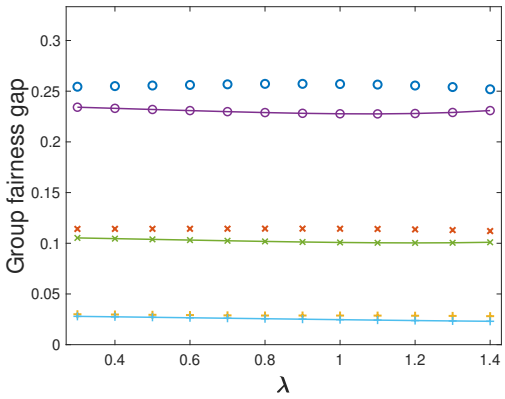
MSE and the group fairness gap for the robust fairness-aware model are both smaller than those for the unrobust fair model, which indicates that the proposed model has better robustness and achieves better performance in both accuracy and fairness. However, for MICD, when $\eta = 1.5\eta_D$, the MSE for the robust fair model becomes larger than that of the unrobust model as the power of the adversarial data point is large, which in turn affects the prediction performance considerably.

4.4.2 Rank-one attack

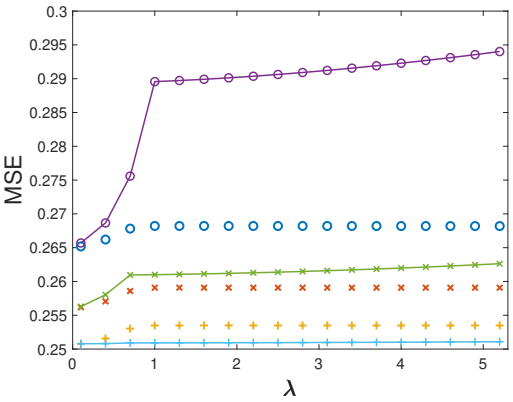
In the first experiment, we explore the effects of the energy constraint parameter η as well as the trade-off parameter λ. We carry out the attack with three different energy levels, $\eta = 0.2\sigma$, $\eta = 0.5\sigma$ and $\eta = 0.8\sigma$, where σ is the smallest singular value of the feature matrix of the training data. As shown in Fig. 4.3, we first observe that MSE and the group fairness gap for the adversarially robust model are almost always smaller than those for the unrobust fair model, which illustrates that the proposed robust model achieves better performance in both accuracy and fairness. We also notice that the performance of the adversarially robust model differs under different choices of λ. In particular, as λ increases, the value of MSE also increases because we care more about fairness and give more weight to the fairness-related term in the objective function. Especially, as shown in Fig. 4.3(c), when the energy constraint is comparable to the smallest singular value of the feature



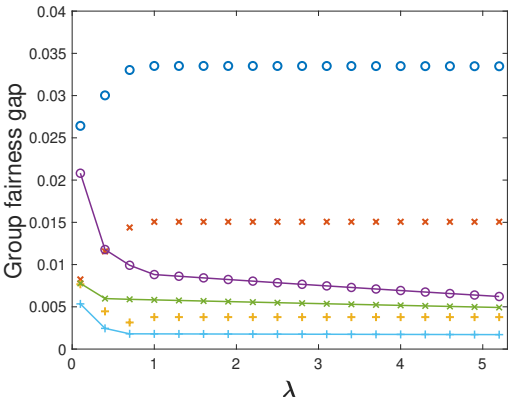
(a) LSD: MSE v.s. λ



(b) LSD: Group fairness gap v.s. λ

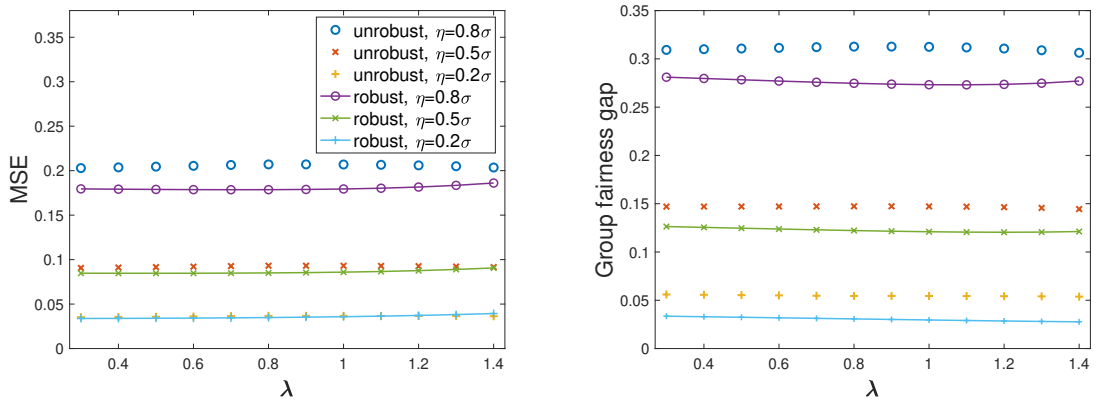


(c) MICD: MSE v.s. λ



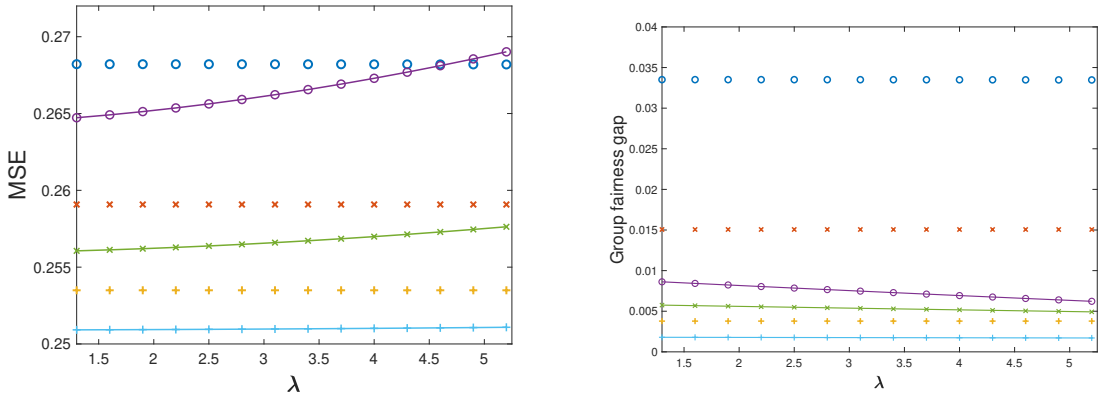
(d) MICD: Group fairness gap v.s. λ

Figure 4.2: Effects of λ and η on MSE and the group fairness gap (attack with one adversarial data point, samples with energy greater than $5\eta_D$ are removed).



(a) LSD: MSE v.s. λ

(b) LSD: Group fairness gap v.s. λ



(c) MICD: MSE v.s. λ

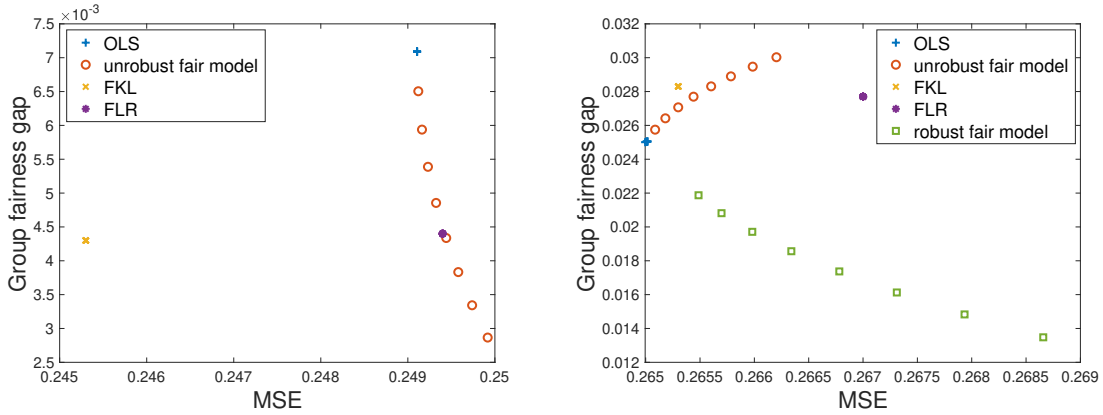
(d) MICD: Group fairness gap v.s. λ

Figure 4.3: Effects of λ and η on MSE and fairness gap (rank-one attack, samples with energy greater than 10σ are removed).

matrix ($\eta = 0.8\sigma$) and the trade-off parameter λ is large ($\lambda = 5.2$), the MSE of the robust model becomes larger than that of the unrobust model as the limitation on the adversary is small, which in turn affects the prediction performance considerably.

In the second experiment, we compare our proposed adversarially robust fair model with other fair regression models and ordinary least square (OLS). In Fig. 4.4, we provide the performance of different regression models on the original dataset as well as the poisoned dataset with $\eta = 0.5\sigma$. For the unrobust fair model and adversarially robust fair model, since the choice of the trade-off parameter λ will affect the model performance, we explore models with various choices of λ from the range $[0.5, 1.2]$. As shown in Fig. 4.4(a), on the original dataset, the overall performance of FKL is better than other models, since it is a nonlinear model based on kernels. FLR has similar perfor-

mance with the proposed unrobust fair regression model (with certain choice of λ). Moreover, for the unrobust fair model, it is observed that as λ increases, the group fairness gap decreases while the MSE increases. However, on the poisoned dataset, as shown in Fig. 4.4(b), the performance of FKL and FLR has been severely impacted. In particular, for FKL (which is the optimal model on the original dataset), the value of the group fairness gap has been increased from 4.3×10^{-3} to 2.8×10^{-2}, and the value of MSE also increases. Similar observations can be found for FLR. Besides, for the unrobust fair model, we observe a concave curve in the group fairness gap v.s. MSE plot, which is convex in the original dataset. Thus, we conclude that fair regression models are vulnerable to adversarial attacks and may not preserve their performance in adversarial environment. On the contrary, for the adversarially robust model, the curve between the group fairness gap and MSE locates in the lower left corner and is convex. Thus, by appropriately choosing λ, a model that performs well in terms of both fairness and prediction accuracy can be obtained.



(a) Original dataset

(b) Poisoned dataset

Figure 4.4: MICD: Group fairness gap v.s. MSE (rank-one attack).

4.5 Conclusion

In this chapter, we have proposed a minimax framework to characterize the best attacker that generates the optimal poisoned point or rank-one attack for the original dataset, as well as the adversarially robust fair defender that can achieve the best performance in terms of both predic-

tion accuracy and fairness guarantee, in the presence of the best attacker. We have discussed two types of attack schemes and provided the corresponding methods to solve the proposed nonsmooth nonconvex-nonconcave minimax problems. Moreover, we have performed numerical experiments on synthetic data and two real-world datasets, and shown that the proposed adversarially robust fair models can achieve better performance in both prediction accuracy and fairness guarantee than other fair regression models with a proper choice of λ.

Chapter 5

Conclusion and Future Directions

In this chapter, we summarize contributions presented in this dissertation. In addition, we discuss potential directions for further exploration.

5.1 Summary and conclusions

This dissertation has provided an exploration into the multifaceted challenges and opportunities in ML algorithms. Three overarching concerns, namely security, privacy protection, and fairness, have been addressed to advance ML models in different scenarios.

In Chapter 2, we have investigated the adversarial robustness of hypothesis testing rules. We have formulated it as a minimax hypothesis testing problem, in which the adversary aims at designing attack strategies to maximize the error probability, while the goal of the decision maker is to design decision rules to minimize the error probability. We have shown that the formulated minimax problem has a saddle-point solution, which reveals the structures of the optimal attack and defense strategies. Under certain assumptions, we have derived an upper-bound on the prediction error, which only depends on the PMFs before the attack. Afterwards, we have designed a specific attack scheme and have shown that the designed attack scheme achieves the upper-bound. In this way, we have characterized the optimal attack and the corresponding optimal decision rules for both hypothesis-aware and hypothesis-unaware adversary models.

In Chapter 3, we have proposed a general framework to design privacy-preserving mapping to achieve privacy-accuracy trade-off in the IAS scenarios. We have formulated optimization problems to find the desirable mapping. However, the formulated problem is a complicated non-concave problem with multiple constraints. To deal with that, we have transformed the optimization problem into a form that has three dominating arguments with certain nice concavity properties, through various transformations and variable augmentations. Then we have designed an iterative method to solve the complicated optimization problem, and have proved the convergence of the proposed method under certain assumptions.

In Chapter 4, we have proposed a minimax framework to characterize the best adversarial attack as well as the adversarially robust fair model that can achieve the best performance in terms of both prediction accuracy and fairness guarantee. We have discussed two types of attack schemes and provided the corresponding minimax problems. However, the proposed minimax problems are nonsmooth nonconvex-nonconcave, which may not have a local saddle point in general. We have carefully examined the underlying structure of the inner maximization problem and the outer minimization problem, and then exploited the identified structure to design efficient algorithms. In particular, for the attack with poisoned data point, when solving the inner maximization problem, we have dealt with the non-smooth nature of the objective function and obtained a structure that characterizes the best adversary. We have then analyzed the minimization problem by transforming it to four sub-problems where each sub-problem is a non-convex quadratic minimization problem with quadratic constraints. For the rank-one attack scheme, we have transformed the maximization problem into a form with five arguments, four of which can be solved exactly. With this transformation, the original problem has been converted into several weakly-convex-weakly-concave minimax problems, which are approximately solvable using existing algorithms. Through numerical examples, we have shown that by properly choosing the trade-off parameter, the robust model can achieve desirable performance in both prediction accuracy and group-based fairness.

5.2 Extensions

The research in this dissertation can be extended in the following directions.

- **Adversarial robustness:** While our investigation into the adversarial robustness of hypothesis testing rules in Chapter 2 has provided valuable insights, the assumption of known underlying distributions may not hold in practical scenarios. Future work should extend this analysis to situations where the true underlying distributions are approximated or entirely unknown to both the attacker and decision-maker. In practical applications, there are cases where we have some knowledge of underlying distributions, but it's only an approximation due to data collection or modeling limitations. In such instances, the developed methods should account for distributional uncertainties and their effects on adversarial attacks. One promising approach to address these challenges is to utilize robust optimization methods. These techniques enable us to address a wider range of scenarios, even when distribution information is uncertain or hidden. By embracing robust optimization, we may design optimal attack and defense strategies even in the presence of distributional ambiguities.

- **Fairness in Adversarial Environments:** Chapter 4 of this dissertation has illuminated the intricate relationship between adversarial robustness, standard accuracy, and accuracy-based fairness measures. However, it is unclear whether such a trade-off is inherent, even in the linear setting. Consequently, there is a pressing need to understand the fundamental limits of adversarial attacks to fair machine learning models and to design new fairness-aware models that can withstand adversarial attacks and maintain robustness in adversarial environments. While this dissertation has primarily focused on the regression problem, classifiers are more commonly employed in decision-making tasks. However, the fairness measure in such tasks is often non-differentiable with respect to the model parameters, posing challenges in analyzing the impact of adversarial attacks. Furthermore, aside from devising specific adversarially robust fair models, it is vital to comprehend the impact of adversarial attacks on the model's performance and to analyze how enforcing robustness influences the fairness measure in con-

trast to standard training. To explore the impact of adversarial robustness to fair classifiers, we may initiate our investigation with a Gaussian mixture model and linear classifiers. In the absence of an adversary, unfairness in classification problems often arises from imbalances between different classes. However, when an adversarial attack is introduced, the impact of enforcing adversarial robustness can be divided into two parts: the first part stems from the inherent constraints of adversarial robustness itself, leading to the degradation of standard accuracy due to changes in the decision boundary; the second part may be attributed to the class imbalance ratio between the two classes under consideration. Additionally, it is also important to quantify the robustness of fair models. Specifically, we may design a framework that measures the model's robustness against adversarial attacks performed on the training data. By assessing the maximum change in any fairness measure, we can gain insights into the model's robustness against adversarial attacks. If the presence of poisoned training samples does not significantly alter the disparity in unfairness, it indicates a higher level of robustness. Through an in-depth analysis of the impact of adversarial attacks on fairness, we may enhance our understanding of the vulnerabilities and sensitivities inherent in fair machine learning models. These insights will be instrumental in the development of fair machine learning models that are robust to adversarial attacks.

Appendix A

Appendix of Chapter 2

A.1 Proof of Lemma 1

We first prove (2.14):

$$\begin{aligned} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}^*) &= \frac{1}{2}\left[P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}^*) + P_M(\boldsymbol{p}_1, \boldsymbol{B}, \boldsymbol{t}^*)\right] \\ &= \frac{1}{2}\left[\sum_{i=1}^{n} q_{0,i} t_i^* + \sum_{i=1}^{n} q_{1,i}(1 - t_i^*)\right] \\ &\overset{(a)}{=} \frac{1}{2} + \frac{1}{2}\sum_{i:q_{0,i}<q_{1,i}} (q_{0,i} - q_{1,i}) \\ &= \frac{1}{2} - \frac{1}{2}\sum_{i:q_{0,i}<q_{1,i}} |q_{0,i} - q_{1,i}| \\ &\overset{(b)}{=} \frac{1}{2} - \frac{1}{4}\sum_{i=1}^{n} |q_{0,i} - q_{1,i}| \,. \end{aligned} \tag{A.1}$$

Here, (a) is true to due to the form of $\boldsymbol{t}^*$ specified in (2.9). We now show (b) is true:

$$\begin{aligned} 0 = \sum_{i=1}^{n}(q_{0,i} - q_{1,i}) &= \sum_{i:q_{0,i}\geq q_{1,i}} (q_{0,i} - q_{1,i}) + \sum_{i:q_{0,i}<q_{1,i}} (q_{0,i} - q_{1,i}) \\ &= \sum_{i:q_{0,i}\geq q_{1,i}} |q_{0,i} - q_{1,i}| - \sum_{i:q_{0,i}<q_{1,i}} |q_{0,i} - q_{1,i}| \,, \end{aligned}$$

which implies

$$\sum_{i:q_{0,i}\geq q_{1,i}} |q_{0,i}-q_{1,i}| = \sum_{i:q_{0,i}<q_{1,i}} |q_{0,i}-q_{1,i}| = \frac{1}{2}\sum_{i=1}^{n} |q_{0,i}-q_{1,i}|\,.$$

We now prove (2.15). Using step (a) of (A.1), we have

$$\begin{aligned}
P_E(\boldsymbol{A},\boldsymbol{B},\boldsymbol{t}^*) &= \frac{1}{2} - \frac{1}{2}\sum_{i:q_{0,i}<q_{1,i}} (q_{1,i}-q_{0,i}) \\
&= \frac{1}{2}\left[\sum_{i:q_{0,i}\geq q_{1,i}} q_{1,i} + \sum_{i:q_{0,i}<q_{1,i}} q_{1,i} - \sum_{i:q_{0,i}<q_{1,i}} (q_{1,i}-q_{0,i})\right] \\
&= \frac{1}{2}\left[\sum_{i:q_{0,i}>q_{1,i}} q_{1,i} + \sum_{i:q_{0,i}=q_{1,i}} q_{1,i} + \sum_{i:q_{0,i}<q_{1,i}} q_{0,i}\right] \\
&= \frac{1}{2}\sum_{i=1}^{n} \min\{q_{0,i},q_{1,i}\}.
\end{aligned}$$

A.2 Proof of Theorem 2

For $\forall(\boldsymbol{A},\boldsymbol{B}) \in \mathcal{A}\times\mathcal{B}$, we have

$$\begin{aligned}
F_j(\boldsymbol{A},\boldsymbol{B}) &= 1 + \sum_{i=1}^{j} \min\{(q_{1,i}-q_{0,i}),0\} \\
&\overset{(a)}{\leq} 1 + \sum_{i=1}^{j} (q_{1,i}-q_{0,i}) \\
&= 1 + \sum_{i=1}^{j} q_{1,i} - \sum_{i=1}^{j} q_{0,j} \\
&= 1 - \sum_{i=1}^{j} (p_{0,i}-p_{1,i}) + K_{0,j} - K_{1,j} + I_{1,j} - I_{0,j}, \\
&\overset{(b)}{\leq} 1 - \sum_{i=1}^{j} (p_{0,i}-p_{1,i}) + \sum_{i=\max\{1,j-\delta+1\}}^{j} p_{0,i} + \sum_{i=j+1}^{\min\{j+\delta,n\}} p_{1,i} \\
&= 1 - \sum_{i=1}^{j-\delta} p_{0,i} + \sum_{i=1}^{j+\delta} p_{1,i} = G_j(\boldsymbol{p}_0,\boldsymbol{p}_1),
\end{aligned} \tag{A.2}$$

Here, the equality in (a) holds when $q_{1,i} - q_{0,i} \leq 0, 1 \leq i \leq j$, inequality (b) comes from the natural restrictions on I, K, in which the equality holds when $K_{0,j} - I_{0,j} = \sum_{i=\max\{1,j-\delta+1\}}^{j} p_{0,i}$, and $I_{1,j} - K_{1,j} = \sum_{i=j+1}^{\min\{j+\delta,n\}} p_{1,i}$.

Note that $2P_E(\boldsymbol{A}, \boldsymbol{B}) = F_n(\boldsymbol{A}, \boldsymbol{B}) \leq ... \leq F_m(\boldsymbol{A}, \boldsymbol{B}) \leq ... \leq F_0 = 1$. As (A.2) holds for $\forall \boldsymbol{A}, \boldsymbol{B} \in \Omega$, we have $F_n(\boldsymbol{A}, \boldsymbol{B}) \leq G_j(\boldsymbol{p}_0, \boldsymbol{p}_1), \forall 1 \leq j \leq n$. Therefore,

$$\begin{aligned} F_m(\boldsymbol{A}, \boldsymbol{B}) &\leq \min_{1\leq j\leq m} \{1, G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}, \\ F_n(\boldsymbol{A}, \boldsymbol{B}) &\leq \min_{1\leq j\leq n} \{1, G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}. \end{aligned}$$

Let $j^* = \arg\min_{1\leq j\leq n} \{G_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}$. If $G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) \leq 1$, we have $F_n(\boldsymbol{A}, \boldsymbol{B}) \leq G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1)$ and the equality is achieved when

- $F_{j^*}(\boldsymbol{A}, \boldsymbol{B}) = G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1)$, which is equivalent to

$$\begin{aligned} &q_{1,i} - q_{0,i} \leq 0, 1 \leq i \leq j^*, \\ &K_{0,j^*} - I_{0,j^*} = \sum_{i=\max\{1,j^*-\delta+1\}}^{j^*} p_{0,i}, \\ &I_{1,j^*} - K_{1,j^*} = \sum_{i=j^*+1}^{\min\{j^*+\delta,n\}} p_{1,i}. \end{aligned}$$

- $F_k(\boldsymbol{A}, \boldsymbol{B}) = F_{j^*}(\boldsymbol{A}, \boldsymbol{B}), j^* < k \leq n$.

If $G_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) > 1$, we have $F_n(\boldsymbol{A}, \boldsymbol{B}) \leq 1$ and the equality is achieved if

- $F_i(\boldsymbol{A}, \boldsymbol{B}) = 1, 1 \leq i \leq n$.

A.3 Proof of the designed attack matrices achieving $\hat{\boldsymbol{q}}_0, \hat{\boldsymbol{q}}_1$

We will calculate the PMF $\hat{\boldsymbol{q}}_0, \hat{\boldsymbol{q}}_1$ achieved by the attack matrices $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$ designed according to (2.22) and (2.23) for columns $2, \cdots, m$, and according to (2.22) and (2.32) for columns $m+1, \cdots, n$. We will show that these satisfy the desired conditions specified in Section 2.2.3.

For $j = 1$,

$$\begin{aligned} q_{0,1} &= \sum_{i=1}^{1+\delta} p_{0,i}\hat{A}_{i,1} = \min\{p_{0,1}, \hat{q}_{0,1}\} + \sum_{i=2}^{1+\delta} \min\{p_{0,i}, \max\{0, \hat{q}_{0,1} - \sum_{k=1}^{i-1} p_{0,k}\}\} \quad \text{(A.3)} \\ &\overset{(a)}{=} \hat{q}_{0,1}, \\ q_{0,1} &= \sum_{i=1}^{1+\delta} p_{1,i}\hat{B}_{i,1} = \sum_{i=1}^{1+\delta} p_{1,i}. \end{aligned}$$

Here, (a) is true because 1) if $p_{0,1} \geq \hat{q}_{0,1}$, then $\min\{p_{0,1}, \hat{q}_{0,1}\} = \hat{q}_{0,1}$ and $\max\{0, \hat{q}_{0,1} - \sum_{k=1}^{i-1} p_{0,k}\} = 0$, which indicates $q_{0,1} = \hat{q}_{0,1}$; 2) if $p_{0,1} < \hat{q}_{0,1}$, then

$$\begin{aligned} &\min\{p_{0,1}, \hat{q}_{0,1}\} + \sum_{i=2}^{1+\delta} \min\left\{p_{0,i}, \max\left\{0, \hat{q}_{0,1} - \sum_{k=1}^{i-1} p_{0,k}\right\}\right\} \\ &= p_{0,1} + \sum_{i=2}^{1+\delta} \min\left\{p_{0,i}, \max\left\{0, \hat{q}_{0,1} - \sum_{k=1}^{i-1} p_{0,k}\right\}\right\} \\ &= \min\{p_{0,2}, \max\{0, \hat{q}_{0,1}\}\} + \sum_{i=3}^{1+\delta} \min\left\{p_{0,i}, \max\left\{0, \hat{q}_{0,1} - \sum_{k=1}^{i-1} p_{0,k}\right\}\right\} \\ &= \min\{p_{0,2}, \hat{q}_{0,1}\} + \sum_{i=3}^{1+\delta} \min\left\{p_{0,i}, \max\left\{0, \hat{q}_{0,1} - \sum_{k=1}^{i-1} p_{0,k}\right\}\right\}. \quad \text{(A.4)} \end{aligned}$$

Note that (A.3) and (A.4) are in the same form. Then by continuing this process, we will have $q_{0,1} = \hat{q}_{0,1}$.

For $2 \leq j \leq m$, under $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}})$, we have

$$\begin{aligned} q_{1,j} &= \sum_{i=1}^{j+\delta} p_{1,i}\hat{B}_{i,j} = p_{1,j+\delta}, \\ q_{0,j} &= \sum_{i=1}^{j+\delta} p_{0,i}\hat{A}_{i,j} \\ &= \sum_{i=1}^{j+\delta} \min\left\{p_{0,i}(1 - \sum_{k=1}^{j-1} \hat{A}_{i,k}), \max\left\{\hat{q}_{0,j} - \sum_{k=1}^{i-1} p_{0,k}\left(1 - \sum_{t=1}^{j-1} \hat{A}_{k,t}\right), 0\right\}\right\} \quad \text{(A.5)} \\ &\overset{(b)}{=} \hat{q}_{0,j}, \end{aligned}$$

in which (b) can be derived using the similar steps discussed in $j = 1$. Specifically, for $i = 1$, the index term in (A.5) is $\min\left\{p_{0,1}(1 - \sum_{k=1}^{j-1} \hat{A}_{1,k}), \max\{\hat{q}_{0,j}, 0\}\right\}$. If $p_{0,1}(1-\sum_{k=1}^{j-1} \hat{A}_{1,k}) \geq \hat{q}_{0,j}$, we have $q_{0,j} = \hat{q}_{0,j}$ directly. On the other hand, if $p_{0,1}(1 - \sum_{k=1}^{j-1} \hat{A}_{1,k}) < \hat{q}_{0,j}$, the index term for $i = 1$ is $p_{0,1}(1-\sum_{k=1}^{j-1} \hat{A}_{1,k})$, which will cancel out with a term in $\hat{q}_{0,j} - \sum_{k=1}^{i-1} p_{0,k}\left(1 - \sum_{t=1}^{j-1} \hat{A}_{k,t}\right)$ and we will have $q_{0,j} = \hat{q}_{0,j}$ after a series of cancellations.

For $j \in R_1$, since the formula of $\hat{A}_{i,j}$ stays the same, $q_{0,j} = \hat{q}_{0,j}$ still holds. Under $\hat{B}_{i,j}$, suppose l is the smallest component with $\hat{B}_{l,j} > 0$, then we have

$$
\begin{aligned}
q_{1,j} &= \sum_{i=1}^{j+\delta} p_{1,i}\hat{B}_{i,j} \\
&= \sum_{i=1}^{j+\delta} \min\left\{p_{1,i}(1 - \sum_{k=1}^{j-1} \hat{B}_{i,k}), \hat{q}_{1,j} - \sum_{k=1}^{i-1} p_{1,k}\hat{B}_{k,j}\right\} \\
&= \sum_{i=l}^{j+\delta} \min\left\{p_{1,i}(1 - \sum_{k=1}^{j-1} \hat{B}_{i,k}), \hat{q}_{1,j} - \sum_{k=1}^{i-1} p_{1,k}\hat{B}_{k,j}\right\} \\
&= \min\left\{p_{1,l}(1 - \sum_{k=1}^{j-1} \hat{B}_{l,k}), \hat{q}_{1,j} - \sum_{k=1}^{l-1} p_{1,k}\hat{B}_{k,j}\right\} \\
&\quad + \sum_{i=l+1}^{j+\delta} \min\left\{p_{1,i}(1 - \sum_{k=1}^{j-1} \hat{B}_{i,k}), \hat{q}_{1,j} - \sum_{k=l}^{i-1} p_{1,k}\hat{B}_{k,j}\right\} \\
&= \min\left\{p_{1,l}(1 - \sum_{k=1}^{j-1} \hat{B}_{l,k}), \hat{q}_{1,j}\right\} + \sum_{i=l+1}^{j+\delta} \min\left\{p_{1,i}(1 - \sum_{k=1}^{j-1} \hat{B}_{i,k}), \hat{q}_{1,j} - \sum_{k=l}^{i-1} p_{1,k}\hat{B}_{k,j}\right\} \\
&\overset{(c)}{=} \hat{q}_{1,j},
\end{aligned}
\tag{A.6}
$$

in which (c) can be proved by considering two different cases. First, if $p_{1,l}(1 - \sum_{k=1}^{j-1} \hat{B}_{l,k}) \geq \hat{q}_{1,j}$, in (A.6), we have $\min\left\{p_{1,l}(1 - \sum_{k=1}^{j-1} \hat{B}_{l,k}), \hat{q}_{1,j}\right\} = \hat{q}_{1,j} = p_{1,l}\hat{B}_{l,j}$ and $\hat{q}_{1,j} - p_{1,l}\hat{B}_{l,j} = 0$, which implies $\hat{B}_{l+1,j} = 0$ and thus $\hat{B}_{i,j} = 0, n \geq i \geq l + 1$. Therefore, (c) holds. Second, if $p_{1,l}(1 - \sum_{k=1}^{j-1} \hat{B}_{l,k}) \leq \hat{q}_{1,j}$, in (A.6), we have $\min\left\{p_{1,l}(1 - \sum_{k=1}^{j-1} \hat{B}_{l,k}), \hat{q}_{1,j}\right\} = p_{1,l}(1 - \sum_{k=1}^{j-1} \hat{B}_{l,k}) = p_{1,l}\hat{B}_{l,j}$, which will cancel out with a term in $\hat{q}_{1,j} - \sum_{k=l}^{i-1} p_{1,k}\hat{B}_{k,j}$ and thus after a series of cancellation, we have $q_{0,j} = \hat{q}_{0,j}$.

A.4 Proof of Theorem 3

For $j = m - \delta$, we have

$$
\begin{aligned}
& F_{m-\delta}(\boldsymbol{A}) \\
= & \sum_{i=1}^{m-\delta} q_{1,i} + \sum_{i=m-\delta+1}^{n} q_{0,i} \\
= & 1 - \sum_{i=1}^{m-\delta}(p_{0,i} - p_{1,i}) + K_{0,m-\delta} - K_{1,m-\delta} + I_{1,m-\delta} - I_{0,m-\delta} \\
\overset{(a)}{\leq} & 1 - \sum_{i=1}^{m-\delta}(p_{0,i} - p_{1,i}) + \sum_{i=m-2\delta+1}^{m-\delta}(p_{0,i} - p_{1,i}) + 0 \\
= & 1 - \sum_{i=1}^{m-2\delta}(p_{0,i} - p_{1,i}) \\
= & E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1).
\end{aligned}
$$

For $\forall m - \delta + 1 \leq j \leq m + \delta, \forall \boldsymbol{A} \in \mathcal{A}$, we have

$$
\begin{aligned}
F_j(\boldsymbol{A}) = & F_{m-\delta}(\boldsymbol{A}) + \sum_{i=m-\delta+1}^{j} \min\{(q_{1,i} - q_{0,i}), 0\} \\
\overset{(b)}{\leq} & F_{m-\delta}(\boldsymbol{A}) + \sum_{i=m-\delta+1}^{j}(q_{1,i} - q_{0,i}) \\
= & \sum_{i=1}^{j} q_{1,i} + \sum_{i=j+1}^{n} q_{0,i} \\
= & 1 + \sum_{i=1}^{j}(q_{1,i} - q_{0,i}) \\
= & 1 - \sum_{i=1}^{j}(p_{0,i} - p_{1,i}) + K_{0,j} - K_{1,j} + I_{1,j} - I_{0,j} \\
\overset{(c)}{\leq} & 1 - \sum_{i=1}^{j}(p_{0,i} - p_{1,i}) + \sum_{i=j-\delta+1}^{\min\{j,m\}}(p_{0,i} - p_{1,i}) + \sum_{i=\max\{m+1,j+1\}}^{\min\{n,j+\delta\}}(p_{1,i} - p_{0,i}) \\
= & 1 - \sum_{i=1}^{j-\delta}(p_{0,i} - p_{1,i}) + \sum_{i=m+1}^{\min\{n,j+\delta\}}(p_{1,i} - p_{0,i}) \\
= & E_j(\boldsymbol{p}_0, \boldsymbol{p}_1),
\end{aligned}
$$

in which the inequalities in $(a), (c)$ follow from the observation about I, K and the equality in (b) holds when $q_{1,i} \leq q_{0,i}, m-\delta+1 \leq i \leq j$.

Since the above inequality holds for $\forall \boldsymbol{A} \in \mathcal{A}$ and we have shown that $F_{m+\delta}(\boldsymbol{A}) \leq F_{m+\delta-1}(\boldsymbol{A}) \leq ... \leq F_{m-\delta}(\boldsymbol{A})$, then

$$\begin{aligned} F_m(\boldsymbol{A}) &\leq \min_{m-\delta \leq j \leq m} \{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}, \\ F_{m+\delta}(\boldsymbol{A}) &\leq \min_{m-\delta \leq j \leq m+\delta} \{E_j(\boldsymbol{p}_0, \boldsymbol{p}_1)\}. \end{aligned}$$

Furthermore, if $j^* > m-\delta$, $F_{m+\delta}(\boldsymbol{A}) \leq E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1)$ and the equality is achieved when

(i)

$$K_{0,m-\delta} - K_{1,m-\delta} = \sum_{i=m-2\delta+1}^{m-\delta} (p_{0,i} - p_{1,i});$$

(ii) $q_{1,i} \leq q_{0,i}, m-\delta+1 \leq i \leq j^*$;

(iii)

$$\begin{aligned} K_{0,j^*} - K_{1,j^*} &= \sum_{i=j^*-\delta+1}^{\min\{j^*,m\}} (p_{0,i} - p_{1,i}), \\ I_{1,j^*} - I_{0,j^*} &= \sum_{i=\max\{m+1,j^*+1\}}^{\min\{n,j^*+\delta\}} (p_{1,i} - p_{0,i}); \end{aligned}$$

(iv) $F_k(\boldsymbol{A}) = F_{j^*}(\boldsymbol{A}), j^* < k \leq m+\delta$.

If $E_{j^*}(\boldsymbol{p}_0, \boldsymbol{p}_1) > E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1)$, the equality is achieved when

$$F_i(\boldsymbol{A}) = E_{m-\delta}(\boldsymbol{p}_0, \boldsymbol{p}_1), m-\delta \leq i \leq m+\delta.$$

Appendix B

Appendix of Chapter 3

B.1 Proof of Lemma 2

$$
\begin{aligned}
& I(S;U) + \sum_{u,y} p(y)p(u|y)D_{KL}[P_{S|y} \parallel P_{S|u}] \\
= & \sum_{s,u,y} p(s,u,y) \log \frac{p(s|u)}{p(s)} + \sum_{s,u,y} p(y)p(u|y)p(s|y) \log \frac{p(s|y)}{p(s|u)} \\
\overset{(a)}{=} & \sum_{s,u,y} p(s,u,y) \left[\log \frac{p(s|u)}{p(s)} + \log \frac{p(s|y)}{p(s|u)} \right] \\
= & \sum_{s,y} p(s,y) \log \frac{p(s|y)}{p(s)} = I(S;Y),
\end{aligned}
$$

where (a) uses the fact that $S \to Y \to U$ is a Markov chain since given Y, S and U are independent.

B.2 Proof of Lemma 3

First, prove that $\mathcal{F}[P_{U|Y}]$ is concave with respect to $P_{S|U}$. By applying Lemma 2, (3.1) can be written in the following form,

$$\mathcal{F}[P_{U|Y}] \;\; = \;\; I(S;Y) - \beta \mathbb{E}_{Y,U}[d(y,u)] - \sum_{u,y} p(y)p(u|y) D_{KL}[P_{S|y} \parallel P_{S|u}]. \tag{B.1}$$

Note that $I(S;Y)$ is a constant under our setup. Given $P_{U|Y}$ and P_U, $\mathbb{E}_{Y,U}[d(y,u)]$ is independent of $P_{S|U}$. Moreover, $P_{S|u}$ and $P_{S|u'}$ are two independent vectors. For given u and y, we have

$$D_{KL}[P_{S|y} \parallel P_{S|u}] = \sum_{s} p(s|y) \log \frac{p(s|y)}{p(s|u)}. \tag{B.2}$$

Since $a \log(x)$ is concave in x, (B.2) is convex in $P_{S|u}$ and $\mathcal{F}[P_{U|Y}]$ is concave with respect to $P_{S|U}$.

Second, we prove that $\mathcal{F}[P_{U|Y}]$ is concave w.r.t P_U when f is strictly convex. Note that P_U only shows up in $\mathbb{E}_{Y,U}[d(y,u)]$ and since f is strictly convex, taking the sum doesn't change the concavity and $\mathcal{F}[P_{U|Y}]$ is also concave in P_U.

Third, we consider $P_{U|Y}$. There are $|\mathcal{Y}|$ conditional distributions in the mapping $P_{U|Y}$, where $P_{U|y}$ and $P_{U|y'}$ are independent when $y \neq y'$. Then we consider a particular row $P_{U|y}$ and prove the concavity. The Hessian matrix of $\mathcal{F}$ with respect to $P_{U|y}$ is

$$\boldsymbol{H}_{\mathcal{F}} = \begin{bmatrix} \frac{\partial^2 \mathcal{F}[p(u|y)]}{\partial p(u_1|y)^2} & \cdots & \frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_1|y) \partial p(u_{|\mathcal{U}|}|y)} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 \mathcal{F}[p(u|y)]}{\partial p(u_{|\mathcal{U}|}|y) \partial p(u_1|y)} & \cdots & \frac{\partial^2 \mathcal{F}[p(u|y)]}{\partial p(u_{|\mathcal{U}|}|y)^2} \end{bmatrix}.$$

Then we calculate each element in $\boldsymbol{H}_{\mathcal{F}}$. Assume that $i \neq j$. Taking derivative based on the

form in (B.1), we have

$$\begin{aligned}\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)^2} &= -\beta \frac{\partial^2 \mathbb{E}_{Y,U}[d(y,u)]}{\partial p(u_i|y)^2}, \\ \frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)\partial p(u_j|y)} &= -\beta \frac{\partial^2 \mathbb{E}_{Y,U}[d(y,u)]}{\partial p(u_i|y)\partial p(u_j|y)},\end{aligned}$$

in which

$$\begin{aligned}\frac{\partial^2 \mathbb{E}_{Y,U}[d(y,u)]}{\partial p(u_i|y)^2} &= p(y)\left[f'(t)\frac{-p(u_i)}{p(u_i|y)^2} - f'(t)\frac{-p(u_i)}{p(u_i|y)^2} - tf''(t)\frac{-p(u_i)}{p(u_i|y)^2}\right] \\ &= p(y)f''(t)\frac{t^2}{p(u_i|y)} > 0, \\ \frac{\partial^2 \mathbb{E}_{Y,U}[d(y,u)]}{\partial p(u_i|y)\partial p(u_j|y)} &\overset{(a)}{=} 0,\end{aligned}$$

where $t = \frac{p(u_i)}{p(u_i|y)}$ and (a) is due to the fact that t is independent of $p(u_j|y)$ when $i \neq j$ and P_U is given. Then we have $\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)^2} < 0$ and $\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)\partial p(u_j|y)} = 0$. Thus, the Hessian matrix $\boldsymbol{H}_{\mathcal{F}}$ is a diagonal matrix with negative entries, which indicates that the objective function $\mathcal{F}$ is concave in $P_{U|y_i}$ and the lemma is proved.

B.3 Proof of Lemma 4

We first ignore (3.7), (3.9) and solve the optimization problem subject to (3.8) only. We will then check that the obtained solution satisfy constraints (3.7), (3.9).

For a $u \in \mathcal{U}$, the Lagrangian is

$$\mathcal{L}_{S|u} = \mathcal{F}[P_{S|U}|P_{U|Y}, P_U] + \alpha\left(\sum_s p(s|u) - 1\right),$$

where α is the Lagrangian multiplier with respect to constraint (3.8) . Since P_U and $P_{U|Y}$ are given,

$\mathcal{L}_{S|u}$ is a convex function with respect to $P_{S|u}$. By taking the derivative, we have

$$\frac{\partial \mathcal{L}_{S|u}}{\partial p(s|u)} = \frac{\sum_y p(y)p(u|y)p(s|y)}{p(s|u)} + \alpha = 0,$$

which indicates

$$p(s|u) = \frac{\sum_y p(y)p(u|y)p(s|y)}{-\alpha}. \tag{B.3}$$

Since $\sum_s p(s|u) = 1$, we have

$$\begin{aligned}
\sum_s p(s|u) = \sum_s \frac{\sum_y p(y)p(u|y)p(s|y)}{-\alpha} &= 1 \\
\implies \alpha &= -\sum_s \sum_y p(y)p(u|y)p(s|y) \\
&= -\sum_y p(y)p(u|y) \sum_s p(s|y) \\
&= -\sum_y p(y)p(u|y) = -p(u).
\end{aligned}$$

Plugging the value of α into (B.3), we have

$$p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)} \geq 0,$$

which guarantees the non-negativity condition in (3.7). It is also easy to check that this satisfies the constraint in (3.9) exactly, preserves the consistency of different arguments and thus is the solution to the $P_{S|U}$ subproblem.

B.4 Proof of Lemma 5 and 6

Note that for $i \neq j$,

$$\frac{\partial^2 \sum_{i=1}^m \lambda(u_i)\delta(u_i) - \frac{\rho}{2}\sum_{i=1}^m \delta(u_i)^2}{\partial p(u_i|y)^2} = -\rho p^2(y) \leq 0,$$
$$\frac{\partial^2 \sum_{i=1}^m \lambda(u_i)\delta(u_i) - \frac{\rho}{2}\sum_{i=1}^m \delta(u_i)^2}{\partial p(u_i|y)\partial p(u_j|y)} = 0.$$

In Lemma 3, we have shown that $\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)^2} < 0$ and $\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)\partial p(u_j|y)} = 0$. Hence, we have

$$\frac{\partial^2 \mathcal{L}[P_{U|Y}]}{\partial p(u_i|y)^2} = \frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)^2} - \rho p^2(y) < 0,$$
$$\frac{\partial^2 \mathcal{L}[P_{U|Y}]}{\partial p(u_i|y)\partial p(u_j|y)} = 0,$$

and that the Hessian matrix $\boldsymbol{H}_{\mathcal{L}}$ is negative-definite. Moreover, the constraint $\sum_{i=1}^m p(u_i|y) = 1, \forall y \in \mathcal{Y}$ defines a convex set and thus the sub-problem on $P_{U|y_i}$ is a convex problem. Similarly, we also have $\frac{\partial^2 \mathcal{L}[P_U]}{\partial p(u_i)^2} < 0$ and $\frac{\partial^2 \mathcal{L}[P_U]}{\partial p(u_i)\partial p(u_j)} = 0$, which indicates that the Hessian matrix of $\mathcal{L}$ with respect to P_U is negative-definite. Combined with the fact that the constraint set is convex, the sub-problem on P_U is a convex optimization problem.

B.5 Proof of Lemma 7

First, note that $I(S;U) \leq H(S)$, which is bounded. Thus, $\mathcal{F}[P_{U|Y}]$ is upper bounded if $\mathbb{E}_{Y,U}[d(y,u)]$ is bounded from above. Let $t(y,u) = \frac{p(u)}{p(u|y)}$. We have that

$$\mathbb{E}_{Y,U}[d(y,u)] = \sum_{y,u} p(y)p(u|y)f(t(y,u)) = \sum_{y,u} p(y)p(u)\frac{f(t(y,u))}{t(y,u)},$$

where $p(y)p(u) \leq 1$. Since $\epsilon \leq p(u|y) \leq 1$, we have that $t(y,u) \in \left[\epsilon, \frac{1}{\epsilon}\right], \forall y, u$. Since f is continuous, it's natural to have $\frac{f(t(y,u))}{t(y,u)} < +\infty$. Then $\mathbb{E}_{Y,U}[d(y,u)]$ is bounded from above.

B.6 Proof of Lemma 8

By the optimality of P_U, we have

$$\begin{cases} 0 = \nabla_{P_U} g\left(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1}\right) - \rho\left(-p(y_1)P_{U|y_1}^{t+1} - p(y_2)P_{U|y_2}^{t+1} + P_U^{t+1}\right) + \Lambda^t, \\ \Lambda^{t+1} = \Lambda^t - \rho\left(-p(y_1)P_{U|y_1}^{t+1} - p(y_2)P_{U|y_2}^{t+1} + P_U^{t+1}\right), \end{cases}$$

which implies

$$0 = \nabla_{P_U} g\left(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1}\right) + \Lambda^{t+1}. \tag{B.4}$$

Then we have

$$\begin{aligned} & \left\|\Lambda^{t+1} - \Lambda^t\right\|_2^2 \\ = & \left\|\nabla_{P_U} g\left(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1}\right) - \nabla_{P_U} g\left(P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^{t}\right)\right\|_2^2 \\ = & \sum_{u\in\mathcal{U}} \left(\frac{\partial g(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1})}{\partial p^{t+1}(u)} - \frac{\partial g(P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^{t})}{\partial p^{t}(u)}\right)^2. \end{aligned} \tag{B.5}$$

Given that $f^{'}(t)$ is l_f-Lipschitz continuous of t, we further have that for any given u,

$$\begin{aligned} & \left(\frac{\partial g(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1})}{\partial p^{t+1}(u)} - \frac{\partial g(P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^{t})}{\partial p^{t}(u)}\right)^2 \\ = & \left(\beta p(y_1)\left[f'\left(\frac{p^t(u)}{p^t(u|y_1)}\right) - f'\left(\frac{p^{t+1}(u)}{p^{t+1}(u|y_1)}\right)\right] + \beta p(y_2)\left[f'\left(\frac{p^t(u)}{p^t(u|y_2)}\right) - f'(\frac{p^{t+1}(u)}{p^{t+1}(u|y_2)})\right]\right)^2 \\ \leq & \beta^2 l_f^2\left[p(y_1)\left|\frac{p^t(u)}{p^t(u|y_1)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_1)}\right| + p(y_2)\left|\frac{p^t(u)}{p^t(u|y_2)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_2)}\right|\right]^2 \\ \leq & 2\beta^2 l_f^2\left[p(y_1)^2\left(\frac{p^t(u)}{p^t(u|y_1)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_1)}\right)^2 + p(y_2)^2\left(\frac{p^t(u)}{p^t(u|y_2)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_2)}\right)^2\right], \end{aligned} \tag{B.6}$$

where $\frac{p^t(u)}{p^t(u|y_1)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_1)} = \frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^t(u|y_1)p^{t+1}(u|y_1)}$. Using the assumption that $\frac{1}{p(u|y)} \leq \frac{1}{\epsilon} < \infty$, we have

$$\left| \frac{p^t(u)}{p^t(u|y_1)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_1)} \right| \leq \left(\frac{1}{\epsilon} \right)^2 \left| p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1) \right|.$$

To further bound $|p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)|$, we have

$$\begin{aligned} & \left| \frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^t(u) - p^{t+1}(u)} \right| \\ = \;& p^{t+1}(u|y_1) + p^{t+1}(u) \frac{|p^{t+1}(u|y_1) - p^t(u|y_1)|}{|p^t(u) - p^{t+1}(u)|} \\ \leq \;& 1 + \frac{|p^{t+1}(u|y_1) - p^t(u|y_1)|}{|p^t(u) - p^{t+1}(u)|}, \end{aligned}$$

and

$$\begin{aligned} & \left| \frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^{t+1}(u|y_1) - p^t(u|y_1)} \right| \\ = \;& p^t(u) + p^t(u|y_1) \frac{|p^t(u) - p^{t+1}(u)|}{|p^{t+1}(u|y_1) - p^t(u|y_1)|} \\ \leq \;& 1 + \frac{|p^t(u) - p^{t+1}(u)|}{|p^{t+1}(u|y_1) - p^t(u|y_1)|}. \end{aligned}$$

Moreover, $\min \left\{ \frac{|p^{t+1}(u|y_1) - p^t(u|y_1)|}{|p^t(u) - p^{t+1}(u)|}, \frac{|p^t(u) - p^{t+1}(u)|}{|p^{t+1}(u|y_1) - p^t(u|y_1)|} \right\} \leq 1$. Then we have

$$\min\{ \left| \frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^t(u) - p^{t+1}(u)} \right|, \left| \frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^{t+1}(u|y_1) - p^t(u|y_1)} \right| \} \leq 2,$$

and thus

$$\begin{aligned} & |p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)| \\ \leq \;& 2|p^t(u) - p^{t+1}(u)| + 2|p^{t+1}(u|y_1) - p^t(u|y_1)|, \end{aligned}$$

and thus

$$\begin{aligned}&\left|\frac{p^t(u)}{p^t(u|y_1)}-\frac{p^{t+1}(u)}{p^{t+1}(u|y_1)}\right|\\ \leq\ &\frac{2}{\epsilon^2}\left[|p^t(u)-p^{t+1}(u)|+|p^{t+1}(u|y_1)-p^t(u|y_1)|\right].\end{aligned}\tag{B.7}$$

Similarly, for $\left(\frac{p^t(u)}{p^t(u|y_2)}-\frac{p^{t+1}(u)}{p^{t+1}(u|y_2)}\right)$, we have

$$\begin{aligned}&\left|\frac{p^t(u)}{p^t(u|y_2)}-\frac{p^{t+1}(u)}{p^{t+1}(u|y_2)}\right|\\ \leq\ &\frac{2}{\epsilon^2}\left[|p^t(u)-p^{t+1}(u)|+|p^{t+1}(u|y_2)-p^t(u|y_2)|\right].\end{aligned}\tag{B.8}$$

Plugging (B.7) and (B.8) into (B.6) and (B.5), we have

$$\begin{aligned}&\left\|\Lambda^{t+1}-\Lambda^t\right\|_2^2\\ =\ &\sum_{u\in\mathcal{U}}\left(\frac{\partial g(P^{t+1}_{U|y_1},P^{t+1}_{U|y_2},P^{t+1}_U)}{\partial p^{t+1}(u)}-\frac{\partial g(P^{t}_{U|y_1},P^{t}_{U|y_2},P^{t}_U)}{\partial p^{t}(u)}\right)^2\\ \leq\ &2\beta^2l_f^2\left(\frac{2}{\epsilon^2}\right)^2\sum_{u\in\mathcal{U}}\big\{p(y_1)^2[|p^t(u)-p^{t+1}(u)|\\ &+|p^{t+1}(u|y_1)-p^t(u|y_1)|]^2+p(y_2)^2[|p^t(u)-p^{t+1}(u)|+|p^{t+1}(u|y_2)-p^t(u|y_2)|]^2\big\}\\ \leq\ &2\beta^2l_f^2\left(\frac{2}{\epsilon^2}\right)^2\sum_{u\in\mathcal{U}}\big\{2p(y_1)^2[(p^t(u)-p^{t+1}(u))^2\\ &+(p^{t+1}(u|y_1)-p^t(u|y_1))^2]+2p(y_2)^2[(p^t(u)-p^{t+1}(u))^2+(p^{t+1}(u|y_2)-p^t(u|y_2))^2]\big\}\\ =\ &4\beta^2l_f^2\left(\frac{2}{\epsilon^2}\right)^2\sum_{u\in\mathcal{U}}\big[p(y_1)^2(p^{t+1}(u|y_1)-p^t(u|y_1))^2\\ &+p(y_2)^2(p^{t+1}(u|y_2)-p^t(u|y_2))^2+(p(y_1)^2+p(y_2)^2)(p^t(u)-p^{t+1}(u))^2\big]\\ \leq\ &4\beta^2l_f^2\left(\frac{2}{\epsilon^2}\right)^2\sum_{u\in\mathcal{U}}\big[(p^{t+1}(u|y_1)-p^t(u|y_1))^2+(p^{t+1}(u|y_2)-p^t(u|y_2))^2+(p^t(u)-p^{t+1}(u))^2\big]\\ =\ &l_\Lambda\left(\left\|P^{t+1}_{U|y_1}-P^t_{U|y_1}\right\|_2^2+\left\|P^{t+1}_{U|y_2}-P^t_{U|y_2}\right\|_2^2+\left\|P^{t+1}_U-P^t_U\right\|_2^2\right),\end{aligned}$$

where $l_\Lambda=4\beta^2l_f^2\left(\frac{2}{\epsilon^2}\right)^2=\frac{16\beta^2l_f^2}{\epsilon^4}$.

B.7 Proof of Lemma 9

Since $f'(t)$ is l_f-Lipschitz continuous and $t = \frac{p(u)}{p(u|y_i)} \in (0, \frac{1}{\epsilon})$, we have that g is differentiable and $\nabla_{P_{U|y_1}} g, \nabla_{P_{U|y_2}} g, \nabla_{P_U} g$ are Lipschitz continuous with constants l_{y_1}, l_{y_2}, l_u for $P_{U|y_1}, P_{U|y_2}, P_U$ respectively. In particular, we have $l_{y_1} = l_{y_2} = \frac{\beta l_f}{\epsilon^3}$ and $l_u = \frac{\beta l_f}{\epsilon}$. Then we have

$$
\begin{aligned}
& \mathcal{L}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t}, P_U^t; \Lambda^t] - \mathcal{L}[P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^t; \Lambda^t] \\
= \; & \mathcal{F}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t}, P_U^t] - \mathcal{F}[P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^t] + \langle \Lambda^t, p(y_1)(P_{U|y_1}^t - P_{U|y_1}^{t+1}) \rangle \\
& + \frac{\rho}{2} \| P_U^t - p(y_1) P_{U|y_1}^t - p(y_2) P_{U|y_2}^t \|_2^2 - \frac{\rho}{2} \| P_U^t - p(y_1) P_{U|y_1}^{t+1} - p(y_2) P_{U|y_2}^t \|_2^2 \\
\overset{(a)}{=} \; & \mathcal{F}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t}, P_U^t] - \mathcal{F}[P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^t] + \langle \Lambda^t, p(y_1)(P_{U|y_1}^t - P_{U|y_1}^{t+1}) \rangle \\
& + \langle \rho(P_U^t - p(y_1) P_{U|y_1}^{t+1} - p(y_2) P_{U|y_2}^t), p(y_1)(P_{U|y_1}^{t+1} - P_{U|y_1}^t) \rangle + \frac{\rho}{2} \| p(y_1)(P_{U|y_1}^t - P_{U|y_1}^{t+1}) \|_2^2 \\
= \; & \mathcal{F}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t}, P_U^t] - \mathcal{F}[P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^t] + \frac{\rho}{2} \| p(y_1)(P_{U|y_1}^t - P_{U|y_1}^{t+1}) \|_2^2 \\
& + \langle P_{U|y_1}^{t+1} - P_{U|y_1}^t, -p(y_1)\Lambda^t + p(y_1)\rho(P_U^t - p(y_1) P_{U|y_1}^{t+1} - p(y_2) P_{U|y_2}^t) \rangle \\
= \; & \mathcal{F}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t}, P_U^t] - \mathcal{F}[P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^t] + \frac{\rho}{2} \| p(y_1)(P_{U|y_1}^t - P_{U|y_1}^{t+1}) \|_2^2 \\
& - \langle P_{U|y_1}^{t+1} - P_{U|y_1}^t, \nabla_{P_{U|y_1}} \mathcal{F}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t}, P_U^t] \rangle \\
\overset{(b)}{\geq} \; & \left[\frac{\rho}{2} p(y_1)^2 - \frac{l_{y_1}}{2} \right] \| P_{U|y_1}^{t+1} - P_{U|y_1}^t \|_2^2,
\end{aligned} \tag{B.9}
$$

where (a) follows from the cosine rule and (b) follows from the fact that $\mathcal{F} = h_1 + h_2 + g$, $h_i(P_{U|y_i})$ is linear in $P_{U|y_i}$, and $\nabla_{P_{U|y_1}} g$ is l_{y_1}-Lipschitz continuous of $P_{U|y_1}$.

Similarly, for the update of $P_{U|y_2}$, we have

$$
\begin{aligned}
& \mathcal{L}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^t; \Lambda^t] - \mathcal{L}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t}, P_U^t; \Lambda^t] \\
\geq \; & \left[\frac{\rho}{2} p(y_2)^2 - \frac{l_{y_1}}{2} \right] \| P_{U|y_1}^{t+1} - P_{U|y_1}^t \|_2^2.
\end{aligned} \tag{B.10}
$$

For the update of P_U and Λ, we have

$$\begin{aligned}
& \mathcal{L}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1}; \Lambda^{t+1}] - \mathcal{L}[P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^t; \Lambda^t] \\
= \; & g(P_{U|Y}^{t+1}, P_U^{t+1}) - g(P_{U|Y}^{t+1}, P_U^t) + \langle \Lambda^{t+1}, P_U^{t+1} - P_U^t \rangle + \frac{\rho}{2}\|P_U^{t+1} - P_U^t\|_2^2 - \frac{1}{\rho}\|\Lambda^{t+1} - \Lambda^t\|_2^2 \\
\geq \; & \frac{\rho - l_u}{2}\|P_U^{t+1} - P_U^t\|_2^2 - \frac{1}{\rho}\|\Lambda^{t+1} - \Lambda^t\|_2^2. \qquad \text{(B.11)}
\end{aligned}$$

Combining (B.9), (B.10) and (B.11), we have

$$\begin{aligned}
& \mathcal{L}[P_{U|Y}^{t+1}, P_U^{t+1}; \Lambda^{t+1}] - \mathcal{L}[P_{U|Y}^t, P_U^t; \Lambda^t] \\
\geq \; & \left[\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2}\right]\|P_{U|y_1}^{t+1} - P_{U|y_1}^t\|_2^2 + \left[\frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2}\right]\left\|P_{U|y_2}^{t+1} - P_{U|y_2}^t\right\|_2^2 \\
& + \frac{\rho - l_u}{2}\left\|P_U^{t+1} - P_U^t\right\|_2^2 - \frac{1}{\rho}\left\|\Lambda^{t+1} - \Lambda^t\right\|_2^2 \\
\overset{(c)}{\geq} \; & \left[\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}\right]\|P_{U|y_1}^{t+1} - P_{U|y_1}^t\|_2^2 + \left[\frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}\right]\left\|P_{U|y_2}^{t+1} - P_{U|y_2}^t\right\|_2^2 \\
& + \left(\frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\right)\left\|P_U^{t+1} - P_U^t\right\|_2^2,
\end{aligned}$$

where (c) follows from Lemma 8.

B.8 Proof of Proposition 2

1) If $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$, according to Lemma 9, we have

$$\mathcal{L}[P_{U|Y}^{t+1}, P_U^{t+1}; \Lambda^{t+1}] - \mathcal{L}[P_{U|Y}^t, P_U^t, \Lambda^t] \geq 0.$$

2) $\forall t \in \mathbb{N}$, $\mathcal{L}[P^t_{U|Y}, P^t_U, P^t_{S|U}; \Lambda^t]$ is upper-bounded.

Assume that there exists P'_U, such that $P'_U - (P^t_{U|Y})^T P_Y = \mathbf{0}$. Then we have

$$
\begin{aligned}
& \mathcal{L}[P^t_{U|Y}, P^t_U; \Lambda^t] \\
= \;& h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U) + \sum_u \lambda^t(u)\delta^t(u) - \frac{\rho}{2}\sum_u \delta^t(u)^2 \\
= \;& h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U) + (\Lambda^t)^T[P^t_U - (P^t_{U|Y})^T P_Y] \\
& -\frac{\rho}{2}[P^t_U - (P^t_{U|Y})^T P_Y]^T[P^t_U - (P^t_{U|Y})^T P_Y] \\
\leq \;& h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U) + (\Lambda^t)^T[P^t_U - (P^t_{U|Y})^T P_Y] \\
= \;& h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U) + \langle \Lambda^t, P^t_U - P'_U \rangle \\
\overset{(a)}{=} \;& h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U) - \langle \nabla_{P_U} g\left(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U\right), P^t_U - P'_U \rangle \\
\overset{(b)}{\leq} \;& h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P'_U) + \frac{l_u}{2}\|P^t_U - P'_U\|_2^2 \\
< \;& \infty,
\end{aligned}
$$

where (a) follows from (B.4) and (b) is true as $\nabla_{P_U} g$ is l_u-Lipschitz continuous.

3) $\{P^t_{U|Y}, P^t_U, \Lambda^t\}$ is bounded.

Since $\forall t \in \mathbb{N}$, $P^t_{U|y_1}, P^t_{U|y_2}$ are PMFs, $\{P_{U|Y}\}^t$ is bounded. Similarly, $\{P_U\}^t$ is also bounded. For Λ^t, Lemma 8 can be generalized to the case where the iteration difference is k and we have

$$
\left\|\Lambda^{t+k} - \Lambda^t\right\|_2^2 \;\leq\; l_\Lambda \left(\left\| P^{t+k}_{U|y_1} - P^t_{U|y_1} \right\|_2^2 + \left\| P^{t+k}_{U|y_2} - P^t_{U|y_2} \right\|_2^2 + \left\| P^{t+k}_U - P^t_U \right\|_2^2 \right), \forall k \in \mathbb{N}^+.
$$

Thus, since $\{P_{U|Y}\}^t$ and $\{P_U\}^t$ are bounded, $\{\Lambda\}^t$ is also bounded.

B.9 Proof of Proposition 3

When ρ is sufficiently large, e.g. $\rho = \frac{7\beta l_f}{\epsilon^3 \min\{p(y_1)^2, p(y_2)^2\}}$, we will have $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$. In this case, since $\mathcal{L}[P_{U|Y}, P_U; \Lambda]$ is non-decreasing between

iterations and upper-bounded, there exists t_0, such that

$$\begin{aligned}
\infty &> \sum_{t=t_0}^{\infty} \left| \mathcal{L}\left[P_{U|y_1}^t, P_{U|y_2}^t, P_U^t; \Lambda^t\right] - \mathcal{L}\left[P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1}; \Lambda^{t+1}\right] \right| \\
&\overset{(b)}{\geq} \left[\frac{\rho}{2} p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}\right] \sum_{t=t_0}^{\infty} \|P_{U|y_1}^{t+1} - P_{U|y_1}^t\|_2^2 \\
&\quad + \left[\frac{\rho}{2} p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}\right] \sum_{t=t_0}^{\infty} \left\|P_{U|y_2}^{t+1} - P_{U|y_2}^t\right\|_2^2 + \left(\frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\right) \sum_{t=t_0}^{\infty} \left\|P_U^{t+1} - P_U^t\right\|_2^2,
\end{aligned}$$

where (b) is from Lemma 9. Then as $t \to \infty$, we have $\left\|P_{U|y_1}^{t+1} - P_{U|y_1}^t\right\|_2 \to 0$, $\left\|P_{U|y_2}^{t+1} - P_{U|y_2}^t\right\|_2 \to 0$, and $\left\|P_U^{t+1} - P_U^t\right\|_2 \to 0$. By Lemma 8, we have $\left\|\Lambda^{t+1} - \Lambda^t\right\|_2 \to 0$, which implies

$$P_U^{t+1} - p\left(y_1\right) P_{U|y_1}^{t+1} - p\left(y_2\right) P_{U|y_2}^{t+1} \to 0.$$

B.10 Proof of Proposition 4

Since $\{P_{U|Y}^t, P_U^t, \Lambda^t\}$ is bounded, there exists a subsequence $\{P_{U|Y}^{t_s}, P_U^{t_s}, \Lambda^{t_s}\}$ that converges to the limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$, i.e. $\lim_{s\to\infty}(P_{U|Y}^{t_s}, P_U^{t_s}, \Lambda^{t_s}) = (\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$. For the limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$, we will show that it is the stationary point of (3.11).

By the optimality of $P_{U|y_1}$, $P_{U|y_2}$ and P_U, we have

$$\begin{aligned}
\mathbf{0} &\in \partial_{P_{U|y_1}} \mathcal{F}[P_{U|y_1}^{t_{s+1}}, P_{U|y_2}^{t_s}] - p(y_1)\Lambda^{t_s} + \rho p(y_1)[P_U^{t_s} - p(y_1)P_{U|y_1}^{t_{s+1}} - p(y_1)P_{U|y_2}^t], \\
\mathbf{0} &\in \partial_{P_{U|y_2}} \mathcal{F}[P_{U|Y}^{t_{s+1}}] - p(y_2)\Lambda^{t_s} + \rho p(y_2)[P_U^{t_s} - (P_{U|Y}^{t_{s+1}})^T P_Y], \\
\mathbf{0} &\in \partial_{P_U} g\left(P_{U|y_1}^{t_{s+1}}, P_{U|y_2}^{t_{s+1}}, P_U^{t_{s+1}}\right) + \Lambda^{t_s} - \rho\left(P_U^{t_{s+1}} - p(y_1)P_{U|y_1}^{t_{s+1}} - p(y_2)P_{U|y_2}^{t_{s+1}}\right).
\end{aligned}$$

Taking the limit along the subsequence and using Proposition 3, we have

$$\begin{aligned} 0 &\in \partial_{P_{U|y_1}} \mathcal{F}[\hat{P}_{U|y_1}] - p(y_1)\hat{\Lambda}, \\ 0 &\in \partial_{P_{U|y_2}} \mathcal{F}[\hat{P}_{U|y_2}] - p(y_2)\hat{\Lambda}, \\ 0 &\in \partial_{P_U} \mathcal{F}[\hat{P}_U] + \hat{\Lambda}, \end{aligned}$$

which indicates that the stationary condition is satisfied at the limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$.

Now we check all constraints in (3.11) are also satisfied at the limit point.

- Since $P_{U|Y}^{t_s} \in \mathcal{P}_{U|Y}, \forall s$, and $\mathcal{P}_{U|Y}$ is a closed set, we have $\hat{P}_{U|Y} \in \mathcal{P}_{U|Y}$;
- By taking limit along the subsequence on both sides of the equation in Proposition 3 5), we have

$$\hat{P}_U = p(y_1)\hat{P}_{U|y_1} + p(y_2)\hat{P}_{U|y_1}; \tag{B.12}$$

- Based on (B.12), we have $\hat{p}(u) > 0, \forall u$, and

$$\sum_u \hat{p}(u) = \sum_u \sum_y \hat{p}(u|y)p(y) = \sum_y p(y) \sum_u \hat{p}(u|y) = \sum_y p(y) = 1,$$

  which indicate that $\hat{P}_U \in \mathcal{P}_U$.

B.11 Proof of Theorem 4

Since $\mathcal{L}[P_{U|Y}^t, P_U^t, \Lambda^t]$ is non-decreasing between iterations and bounded from above, we have that $\mathcal{L}[P_{U|Y}^{t_s}, P_U^{t_s}, \Lambda^{t_s}]$ is also monotonic non-decreasing and upper-bounded. Then we have $\lim_{s\to\infty} \mathcal{L}[P_{U|Y}^{t_s}, P_U^{t_s}, \Lambda^{t_s}] = \mathcal{L}[\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda}]$ as $\mathcal{L}$ is continuous for $P_{U|Y} \in \mathcal{P}_{U|Y}, P_U \in \mathcal{P}_U$, and Theorem 4 is proved following from Proposition 4.

B.12 Proof of Lemma 10

The optimality condition of v-subproblem yields

$$0 = \nabla_v g\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}\right) - \Lambda^k + \rho\left(p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^{k+1}\right) - \nabla\phi\left(v^{k+1}\right) + \nabla\phi\left(v^k\right).$$

As $\Lambda^{k+1} = \Lambda^k - \rho\left(p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^{k+1}\right)$, we have $\Lambda^{k+1} = \nabla_v g\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}\right) - \nabla\phi\left(v^{k+1}\right) + \nabla\phi\left(v^k\right)$. Thus,

$$\begin{aligned}
&\left\|\Lambda^{k+1} - \Lambda^k\right\|_2^2 \\
&= \left\|\nabla_v g\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}\right) - \nabla_v g\left(x_1^k, x_2^k, v^k\right) - \nabla\phi\left(v^{k+1}\right) + \nabla\phi\left(v^k\right) + \nabla\phi\left(v^k\right) - \nabla\phi\left(v^{k-1}\right)\right\|_2^2 \\
&\leq 3\left(\left\|\nabla_v g\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}\right) - \nabla_v g\left(x_1^k, x_2^k, v^k\right)\right\|_2^2\right. \\
&\left. + \left\|\nabla\phi\left(v^{k-1}\right) - \nabla\phi\left(v^k\right)\right\|_2^2 + \left\|\nabla\phi\left(v^k\right) - \nabla\phi\left(v^{k+1}\right)\right\|_2^2\right) \\
&\leq 3l_g^2\left(\left\|x_1^{k+1} - x_1^k\right\|_2^2 + \left\|x_2^{k+1} - x_2^k\right\|_2^2\right) + 3\left(l_g^2 + l_\phi^2\right)\left\|v^{k+1} - v^k\right\|_2^2 + 3l_\phi^2\left\|v^k - v^{k-1}\right\|_2^2.
\end{aligned}$$

B.13 Proof of Lemma 11

From the update of x_1, we have

$$\begin{aligned}
&h_1\left(x_1^{k+1}\right) + \left\langle x_1^{k+1} - x_1^k, \nabla_{x_1} g\left(u^k\right)\right\rangle + \left\langle \Lambda^k, p(y_1)x_1^{k+1} + p(y_2)x_2^k - v^k\right\rangle \\
&\quad - \frac{\rho}{2}\left\|p(y_1)x_1^{k+1} + p(y_2)x_2^k - v^k\right\|_2^2 - \Delta_{\varphi_1}\left(x_1^{k+1}, x_1^k\right) \\
&\geq h_1\left(x_1^k\right) + \left\langle \Lambda^k, r_k\right\rangle - \frac{\rho}{2}\left\|r_k\right\|_2^2,
\end{aligned}$$

where $u^k = (x_1^k, x_2^k, y^k)^T$ and $r_k = p(y_1)x_1^k + p(y_2)x_2^k - v^k$.

From the update of x_2, we have

$$\begin{aligned}
&h_2\left(x_2^{k+1}\right) + \left\langle x_2^{k+1} - x_2^k, \nabla_{x_2} g\left(u^k\right)\right\rangle + \left\langle \Lambda^k, p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^k\right\rangle \\
&- \frac{\rho}{2}\left\|p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^k\right\|_2^2 - \Delta_{\varphi_2}\left(x_2^{k+1}, x_2^k\right) \\
&\geq h_2\left(x_2^k\right) + \left\langle \Lambda^k, p(y_1)x_1^{k+1} + p(y_2)x_2^k - v^k\right\rangle - \frac{\rho}{2}\left\|p(y_1)x_1^{k+1} + p(y_2)x_2^k - v^k\right\|_2^2.
\end{aligned}$$

From the update of v, we have

$$\begin{aligned}
&g\left(u^{k+1}\right)+\left\langle \Lambda^k, r_{k+1}\right\rangle-\frac{\rho}{2}\left\|r_{k+1}\right\|_2^2-\Delta_\phi\left(v^{k+1}, v^k\right) \\
\geq & g\left(x_1^{k+1}, x_2^{k+1}, v^k\right)-\frac{\rho}{2}\left\|p(y_1)x_1^{k+1}+p(y_2)x_2^{k+1}-v^k\right\|_2^2 \\
&+\left\langle \Lambda^k, p(y_1)x_1^{k+1}+p(y_2)x_2^{k+1}-v^k\right\rangle,
\end{aligned}$$

where $u^{k+1}=(x_1^{k+1}, x_2^{k+1}, y^{k+1})^T$ and $r_{k+1}=p(y_1)x_1^{k+1}+p(y_2)x_2^{k+1}-v^{k+1}$.

Adding up the above three inequalities, we have

$$\begin{aligned}
&\mathcal{L}\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^k\right) \\
=& h_1\left(x_1^{k+1}\right)+h_2\left(x_2^{k+1}\right)+g\left(u^{k+1}\right)+\left\langle \Lambda^k, r_{k+1}\right\rangle-\frac{\rho}{2}\left\|r_{k+1}\right\|_2^2 \\
\geq & h_1\left(x_1^k\right)+h_2\left(x_2^k\right)+g\left(x_1^{k+1}, x_2^{k+1}, v^k\right)+\left\langle \Lambda^k, r_k\right\rangle \\
&-\left[\left\langle x_1^{k+1}-x_1^k, \nabla_{x_1} g\left(u^k\right)\right\rangle+\left\langle x_2^{k+1}-x_2^k, \nabla_{x_2} g\left(u^k\right)\right\rangle\right] \\
&+\Delta_{\varphi_1}\left(x_1^{k+1}, x_1^k\right)+\Delta_{\varphi_2}\left(x_2^{k+1}, x_2^k\right)+\Delta_\phi\left(v^{k+1}, v^k\right)-\frac{\rho}{2}\left\|r_k\right\|_2^2 \\
=& h_1\left(x_1^k\right)+h_2\left(x_2^k\right)+g\left(u^k\right)+\left\langle \Lambda^k, r_k\right\rangle-\frac{\rho}{2}\left\|r_k\right\|_2^2 \\
&-g\left(u^k\right)+g\left(x_1^{k+1}, x_2^{k+1}, v^k\right)-\left\langle\left(x_1^{k+1}-x_1^k, x_2^{k+1}-x_2^k, 0\right), \nabla g\left(u^k\right)\right\rangle \\
&+\Delta_{\varphi_1}\left(x_1^{k+1}, x_1^k\right)+\Delta_{\varphi_2}\left(x_2^{k+1}, x_2^k\right)+\Delta_\phi\left(v^{k+1}, v^k\right) \\
=& \mathcal{L}\left(w^k\right)+g\left(x_1^{k+1}, x_2^{k+1}, v^k\right)-g\left(u^k\right)-\left\langle\left(x_1^{k+1}-x_1^k, x_2^{k+1}-x_2^k, 0\right), \nabla g\left(u^k\right)\right\rangle \\
&+\Delta_{\varphi_1}\left(x_1^{k+1}, x_1^k\right)+\Delta_{\varphi_2}\left(x_2^{k+1}, x_2^k\right)+\Delta_\phi\left(v^{k+1}, v^k\right) \\
\overset{(a)}{\geq} & \mathcal{L}\left(w^k\right)-\frac{l_g}{2}\left[\left\|x_1^{k+1}-x_1^k\right\|_2^2+\left\|x_2^{k+1}-x_2^k\right\|_2^2\right] \\
&+\frac{\delta_{\varphi_1}}{2}\left\|x_1^{k+1}-x_1^k\right\|_2^2+\frac{\delta_{\varphi_2}}{2}\left\|x_2^{k+1}-x_2^k\right\|_2^2+\frac{\delta_\phi}{2}\left\|v^{k+1}-v^k\right\|_2^2,
\end{aligned}$$

where (a) follows from the assumption 3) and the fact from [192] that if $h:\mathbb{R}^n\longrightarrow\mathbb{R}$ is a continuous differentiable function where gradient ∇h is Lipschitz continuous with the modulus $l_h>0$, then for any $x, y\in\mathbb{R}^n$, we have $|h(y)-h(x)-\langle\nabla h(x), y-x\rangle|\leq\frac{l_h}{2}\|y-x\|_2^2$, and apply this result on g here.

By using the fact that $\langle \Lambda^{k+1} - \Lambda^k, r_{k+1} \rangle = -\frac{1}{\rho}\|\Lambda^{k+1} - \Lambda^k\|_2^2$, we have

$$
\begin{aligned}
& \mathcal{L}\left(w^{k+1}\right) - \mathcal{L}\left(w^k\right) \\
= \; & \mathcal{L}\left(w^{k+1}\right) - \mathcal{L}\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^k\right) + \mathcal{L}\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^k\right) - \mathcal{L}\left(w^k\right) \\
= \; & -\frac{1}{\rho}\left\|\Lambda^{k+1} - \Lambda^k\right\|_2^2 + \mathcal{L}\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^k\right) - \mathcal{L}\left(w^k\right) \\
\geq \; & \left(\frac{\delta_{\varphi_1} - l_g}{2} - \frac{3l_g^2}{\rho}\right)\left\|x_1^{k+1} - x_1^k\right\|_2^2 + \left(\frac{\delta_{\varphi_2} - l_g}{2} - \frac{3l_g^2}{\rho}\right)\left\|x_2^{k+1} - x_2^k\right\|_2^2 \\
& + \left(\frac{\delta_\phi}{2} - \frac{3l_g^2 + 3l_\phi^2}{\rho}\right)\left\|v^{k+1} - v^k\right\|_2^2 - \frac{3l_\phi^2}{\rho}\left\|v^k - v^{k-1}\right\|_2^2,
\end{aligned}
$$

which implies

$$
\begin{aligned}
& \left(\mathcal{L}\left(w^{k+1}\right) - \frac{3l_\phi^2}{\rho}\left\|v^{k+1} - v^k\right\|_2^2\right) - \left(\mathcal{L}\left(w^k\right) - \frac{3l_\phi^2}{\rho}\left\|v^k - v^{k-1}\right\|_2^2\right) \\
\geq \; & \left(\frac{\delta_{\varphi_1} - l_g}{2} - \frac{3l_g^2}{\rho}\right)\left\|x_1^{k+1} - x_1^k\right\|_2^2 + \left(\frac{\delta_{\varphi_2} - l_g}{2} - \frac{3l_g^2}{\rho}\right)\left\|x_2^{k+1} - x_2^k\right\|_2^2 \\
& + \left(\frac{\delta_\phi}{2} - \frac{3l_g^2 + 6l_\phi^2}{\rho}\right)\left\|v^{k+1} - v^k\right\|_2^2.
\end{aligned}
$$

Appendix C

Appendix of Chapter 4

C.1 Proof of Theorem 5

We will prove the maximum value of L under two cases: $G_0 = 1$ and $G_0 = 2$ separately. For $G_0 = 1$, we will show $\max_{\substack{(\boldsymbol{x}_0, y_0, 1), \\ \text{s.t. } \|[\boldsymbol{x}_0^T, y_0]\|_2 \leq \eta}} L_1 \overset{(b)}{=} \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta})\}$. Similarly, for $G_0 = 2$, we will have that $\max_{\substack{(\boldsymbol{x}_0, y_0, 2), \\ \text{s.t. } \|[\boldsymbol{x}_0^T, y_0]\|_2 \leq \eta}} L_2 \overset{(c)}{=} \max\{g_2(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}$. Then (a) follows directly from (b) and (c). Since the case $G_0 = 2$ is similar to the case $G_0 = 1$, we will only verify the equality (b).

Firstly, for the adversarial point, under the constraint that $\|\tilde{\boldsymbol{x}}_0^T\|_2 = \|[\boldsymbol{x}_0^T, y_0]\|_2 \leq \eta$, we have

$$0 \leq \|\tilde{\boldsymbol{x}}_0^T \boldsymbol{b}\|_2^2 \leq \eta^2 \|\boldsymbol{b}\|_2^2 = \eta^2 (1 + \|\boldsymbol{\beta}\|_2^2). \tag{C.1}$$

Then we notice that

$$\begin{aligned} L_1 \overset{(d)}{\leq} & \max \Bigg\{ \left(\frac{\lambda}{m+1} + \frac{1}{n+1} \right) \eta^2 (1 + \|\boldsymbol{\beta}\|_2^2) + \left(\frac{\lambda}{m+1} + \frac{1}{n+1} \right) \|\boldsymbol{y_1} - \boldsymbol{X}_1 \boldsymbol{\beta}\|_2^2 \\ & + \left(-\frac{\lambda}{n-m} + \frac{1}{n+1} \right) \|\boldsymbol{y_2} - \boldsymbol{X}_2 \boldsymbol{\beta}\|_2^2, \max\{0, -\frac{\lambda}{m+1} + \frac{1}{n+1}\} \eta^2 (1 + \|\boldsymbol{\beta}\|_2^2) \\ & + \left(-\frac{\lambda}{m+1} + \frac{1}{n+1} \right) \|\boldsymbol{y_1} - \boldsymbol{X}_1 \boldsymbol{\beta}\|_2^2 + \left(\frac{\lambda}{n-m} + \frac{1}{n+1} \right) \|\boldsymbol{y_2} - \boldsymbol{X}_2 \boldsymbol{\beta}\|_2^2 \Bigg\} \\ = & \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta})\}, \end{aligned}$$

where (d) is from (C.1). Then we verify the achievability of the equality in (d). Define a set $B_1 \coloneqq \{\boldsymbol{\beta} : g_1(\boldsymbol{\beta}) \geq h_1(\boldsymbol{\beta}) = \{\boldsymbol{\beta} : \frac{1}{m+1}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 - \frac{1}{n-m}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2 \geq \max\left\{-\frac{1}{2(m+1)} - \frac{1}{2\lambda(n+1)}, -\frac{1}{m+1}\right\} \cdot \eta^2(1 + \|\boldsymbol{\beta}\|_2^2)$. In the sequel, we will verify the achievability of the equality in (d) with two cases: $\boldsymbol{\beta} \in B_1$ and $\boldsymbol{\beta} \in B_1^c$.

Case 1: $\boldsymbol{\beta} \in B_1$: By taking $\tilde{\boldsymbol{x}}_0 = \eta\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_2}$, we have

$$\begin{aligned} & \frac{1}{m+1}(\|y_0 - \boldsymbol{x}_0^T\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2) - \frac{1}{n-m}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2 \\ = & \frac{1}{m+1}\left[\eta^2(1 + \|\boldsymbol{\beta}\|_2^2) + \|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2\right] - \frac{1}{n-m}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2 \\ \overset{(e)}{\geq} & \max\left\{\frac{1}{2(m+1)} - \frac{1}{2\lambda(n+1)}, 0\right\}\eta^2(1 + \|\boldsymbol{\beta}\|_2^2)\}, \end{aligned} \tag{C.2}$$

where (e) is from the definition of set B_1. Then we have $L_1 \overset{(f)}{=} g_1(\boldsymbol{\beta})$, where (f) follows from (C.2). Therefore, for $\boldsymbol{\beta} \in B_1$, we have $h_1(\boldsymbol{\beta}) \leq g_1(\boldsymbol{\beta})$ and $L_1 \leq g_1(\boldsymbol{\beta})$, in which the equality can be achieved for $\tilde{\boldsymbol{x}}_0 = \eta\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_2}$.

Case 2: $\boldsymbol{\beta} \in B_1^c$: On the one hand, if $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} \leq 0$, by taking $\tilde{\boldsymbol{x}}_0 = \eta\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_2}$, we have

$$\begin{aligned} & \frac{1}{m+1}(\|y_0 - \boldsymbol{x}_0^T\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2) - \frac{1}{n-m}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2 \\ \overset{(g)}{<} & \max\left\{\frac{1}{2(m+1)} - \frac{1}{2\lambda(n+1)}, 0\right\}\eta^2(1 + \|\boldsymbol{\beta}\|_2^2) \\ \overset{(h)}{=} & 0, \end{aligned} \tag{C.3}$$

where (g) is from the definition of set B_1 and (h) is because $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} \leq 0$. Then we have

$$\begin{aligned} L_1 \overset{(j)}{=} & \frac{1}{n+1}\left[\eta^2(1 + \|\boldsymbol{\beta}\|_2^2) + \|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2\right. \\ & \left. + \|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2\right] - \lambda\left[\frac{1}{m+1}\eta^2(1 + \|\boldsymbol{\beta}\|_2^2) + \frac{1}{m+1}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 - \frac{1}{n-m}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2\right] \overset{(k)}{=} h_1(\boldsymbol{\beta}), \end{aligned}$$

where (j) is from (C.3) and (k) is true because $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} \leq 0$. Therefore, for $\boldsymbol{\beta} \in B_1^c$ and $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} \leq 0$, we have $h_1(\boldsymbol{\beta}) \geq g_1(\boldsymbol{\beta})$ and $L_1 \leq h_1(\boldsymbol{\beta})$, in which the equality can be achieved for $\tilde{\boldsymbol{x}}_0 = \eta\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_2}$.

On the other hand, if $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} > 0$, by taking $\tilde{\boldsymbol{x}}_0$ to be a vector such that $\tilde{\boldsymbol{x}}_0 \perp \boldsymbol{b}$, we have

$$\begin{aligned}
& \frac{1}{m+1}(\|y_0 - \boldsymbol{x}_0^T\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2) - \frac{1}{n-m}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2 \\
= & \frac{1}{m+1}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 - \frac{1}{n-m}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2 \\
\overset{(l)}{<} & \max\{-\frac{1}{2(m+1)} - \frac{1}{2\lambda(n+1)}, -\frac{1}{m+1}\}\eta^2(1 + \|\boldsymbol{\beta}\|_2^2) \\
< & 0,
\end{aligned} \tag{C.4}$$

where (l) is from the definition of set B_1. Then we have

$$\begin{aligned}
L_1 \overset{(s)}{=} & \frac{1}{n+1}\left(\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2\right) \\
& -\lambda\left[\frac{1}{m+1}\|\boldsymbol{y_1} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 - \frac{1}{n-m}\|\boldsymbol{y_2} - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2\right] \\
\overset{(t)}{=} & h_1(\boldsymbol{\beta}),
\end{aligned}$$

where (s) is from (C.4) and (t) is because $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} > 0$. Therefore, for $\boldsymbol{\beta} \in B_1^c$ and $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} > 0$, we have $h_1(\boldsymbol{\beta}) \geq g_1(\boldsymbol{\beta})$ and $L_1 \leq h_1(\boldsymbol{\beta})$, in which the equality can be achieved when $\tilde{\boldsymbol{x}}_0 \perp \boldsymbol{b}$.

C.2 Proof of Proposition 7

First, we summarize the process of finding $\bar{\boldsymbol{\beta}}$ as follows.

1. Check whether $A = \{\alpha_1 : \boldsymbol{M_{g_1}} - \alpha_1(\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_1}}) \succeq 0\} \neq \emptyset$. If $A = \emptyset$, there does not exist a global minimizer in this case.

2. By randomly selecting an $\alpha_1^* \in A_{g_1h_1} \coloneqq \{\alpha : \boldsymbol{M_{g_1}} - \alpha(\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_1}}) \succ 0\}$, we solve the optimization problem

$$\begin{aligned}
\min_{\boldsymbol{\beta}} \quad & k(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - \alpha_1^*[g_1(\boldsymbol{\beta}) - h_1(\boldsymbol{\beta})], \\
\text{s.t.} \quad & C_1(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - h_1(\boldsymbol{\beta}) = 0,
\end{aligned} \tag{C.5}$$

where $k(\boldsymbol{\beta})$ is positive-definite and the choice of α_1^* does not affect the solution to the problem.

3. For the solution to (C.5), check whether $\alpha_1 > 0$, (4.11), (4.12) and (4.13) are satisfied.

Now we explore the details of steps 1, 2 and 3.

In step 1, the assumption $\eta^2 \geq \eta^2_{\min} = \max\left\{\frac{(n+1)v_{X_1,p}}{m(m+1)}, \frac{(n+1)v_{X_2,p}}{(n-m+1)(n-m)}\right\}$ will guarantee that A is nonempty. To be exact, we denote $A_{g_1g_2} = \{\alpha : \boldsymbol{M_{g_1}} - \alpha(\boldsymbol{M_{g_1}} - \boldsymbol{M_{g_2}}) \succ 0\}, A_{g_1h_2} = \{\alpha : \boldsymbol{M_{g_1}} - \alpha(\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_2}}) \succ 0\}, A_{h_1g_2} = \{\alpha : \boldsymbol{M_{h_1}} - \alpha(\boldsymbol{M_{h_1}} - \boldsymbol{M_{g_2}}) \succ 0\}, A_{h_1h_2} = \{\alpha : \boldsymbol{M_{h_1}} - \alpha(\boldsymbol{M_{h_1}} - \boldsymbol{M_{h_2}}) \succ 0\}, A_{g_2h_2} = \{\alpha : \boldsymbol{M_{g_2}} - \alpha(\boldsymbol{M_{g_2}} - \boldsymbol{M_{h_2}}) \succ 0\}$. Then under the assumption that $\eta^2 \geq \eta^2_{\min} = \max\left\{\frac{(n+1)v_{X_1,p}}{m(m+1)}, \frac{(n+1)v_{X_2,p}}{(n-m+1)(n-m)}\right\}$, we are able to derive that $A_{g_1h_1} \neq \emptyset, A_{g_1g_2} \neq \emptyset, A_{g_1h_2} \neq \emptyset, A_{h_1g_2} \neq \emptyset, A_{h_1h_2} \neq \emptyset, A_{g_2h_2} \neq \emptyset$. The detailed proof is omitted here. Particularly, in this case study, we have $A_{g_1h_1} \subset A$ and $A \neq \emptyset$.

In step 2, (C.5) is a strictly convex quadratic optimization problem with one quadratic equality constraint, which has been discussed in [193]. Define the Lagrangian function of (C.5) as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \gamma) &= k(\boldsymbol{\beta}) - \gamma C_1(\boldsymbol{\beta}) \\
&= g_1(\boldsymbol{\beta}) - (\alpha_1^* + \gamma)(g_1(\boldsymbol{\beta}) - h_1(\boldsymbol{\beta})) \\
&= (1 - \alpha_1^* - \gamma)g_1(\boldsymbol{\beta}) + (\alpha_1^* + \gamma)h_1(\boldsymbol{\beta}),
\end{aligned}$$

where γ is the Lagrangian multiplier. According to [193], the global minimizer $\check{\boldsymbol{\beta}}$ and the corresponding multiplier γ^* of (C.5) satisfy first-order, second-order and the constraint conditions

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}}|_{\check{\boldsymbol{\beta}}} = (1 - \alpha_1^* - \gamma)\frac{\partial g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\check{\boldsymbol{\beta}}} + (\alpha_1^* + \gamma)\frac{\partial h_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\check{\boldsymbol{\beta}}} = \mathbf{0}, \tag{C.6}$$

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta}^2} = 2\left[(1 - \alpha_1^* - \gamma)\boldsymbol{M}_{g_1} + (\alpha_1^* + \gamma^*)\boldsymbol{M}_{h_1}\right] \succeq \mathbf{0},$$

$$C_1(\check{\boldsymbol{\beta}}) = 0. \tag{C.7}$$

From (C.6), we have

$$\check{\boldsymbol{\beta}} = \left[(1 - \alpha_1^* - \gamma^*)\boldsymbol{M}_{g_1} + (\alpha_1^* + \gamma^*)\boldsymbol{M}_{h_1}\right]^{-1} \cdot \left[(1 - \alpha_1^* - \gamma^*)\boldsymbol{E}_{g_1} - (\alpha_1^* + \gamma^*)\boldsymbol{E}_{h_1}\right]. \tag{C.8}$$

Substituting (C.8) into (C.7), we derive an equation for γ, $K(\gamma) = C_1(\check{\boldsymbol{\beta}}) = 0$, whose root is γ^*. By plugging $\gamma = \gamma^*$ back into (C.8), the exact solution for $\check{\boldsymbol{\beta}}$ is obtained.

For step 3, if $C_2(\check{\boldsymbol{\beta}}) \geq 0, C_3(\check{\boldsymbol{\beta}}) \geq 0$, then we have that:

(1) if $\alpha_1^* + \gamma^* > 0$, $\boldsymbol{\beta} = \check{\boldsymbol{\beta}}$ is a global minimizer satisfying (4.11), (4.12), (4.13) with $\alpha_1 = \alpha_1^* + \gamma^*$;

(2) if $\alpha_1^* + \gamma^* = 0$, $\boldsymbol{\beta} = \check{\boldsymbol{\beta}}$ is a global minimizer in **Case 1** that satisfies (4.7), (4.8), (4.9);

(3) if $\alpha_1^* + \gamma^* < 0$, $\boldsymbol{\beta} = \check{\boldsymbol{\beta}}$ satisfies global optimality conditions for the minimization of $h_1(\boldsymbol{\beta})$ with multipliers $\alpha_1' = 1 - \alpha_1^* - \gamma^*, \alpha_2' = \alpha_3' = 0$.

C.3 Proof of Proposition 8

First, we summarize the process of finding $\hat{\boldsymbol{\beta}}$ as follows.

1. Check $AA = \{(\alpha_1, \alpha_2) : \boldsymbol{M_{g_1}} - \alpha_1(\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_1}}) - \alpha_2(\boldsymbol{M_{g_1}} - \boldsymbol{M_{g_2}}) \succeq 0\} \neq \emptyset$. Under the assumption made in Proposition 7 that $\eta^2 \geq \eta^2_{\min} = \max\left\{\frac{(n+1)v_{X_1,p}}{m(m+1)}, \frac{(n+1)v_{X_2,p}}{(n-m+1)(n-m)}\right\}$, we have $A_{g_1h_1} = \{\alpha : \boldsymbol{M_{g_1}} - \alpha(\boldsymbol{M_{g_1}} - \boldsymbol{M_{h_1}}) \succ 0\} \neq \emptyset$ and $A_{g_1g_2} = \{\alpha : \boldsymbol{M_{g_1}} - \alpha(\boldsymbol{M_{g_1}} - \boldsymbol{M_{g_2}}) \succ 0\} \neq \emptyset$, which implies $AA \neq \emptyset$.

2. Solve the optimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & k(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - \alpha_1^*[g_1(\boldsymbol{\beta}) - h_1(\boldsymbol{\beta})], \\ \text{s.t.} \quad & C_1(\boldsymbol{\beta}) = C_2(\boldsymbol{\beta}) = 0. \end{aligned} \tag{C.9}$$

3. For the solution to (C.9), check whether $\alpha_1 > 0, \alpha_2 > 0$, and (4.17) are satisfied.

We now provide more details of steps 2 and 3. In step 2, define the Lagrangian function of (C.9) as

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\beta}, \gamma_i) = k(\boldsymbol{\beta}) - \gamma_1 C_1(\boldsymbol{\beta}) - \gamma_2 C_1(\boldsymbol{\beta}) \\ &= (1 - \alpha_1^* - \gamma_1 - \gamma_2) g_1(\boldsymbol{\beta}) + (\alpha_1^* + \gamma_1) h_1(\boldsymbol{\beta}) + \gamma_2 h_1(\boldsymbol{\beta}). \end{aligned}$$

Then the derived optimal solution $\check{\boldsymbol{\beta}}$ and the corresponding Lagrangian multipliers γ_1^*, γ_2^* satisfy first-order, second-order and the constraint conditions

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}}|_{\check{\boldsymbol{\beta}}} &= (1-\alpha_1^*-\gamma_1^*-\gamma_2^*)\frac{\partial g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\check{\boldsymbol{\beta}}} + (\alpha_1^*+\gamma_1^*)\frac{\partial h_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\check{\boldsymbol{\beta}}} + \gamma_2^*\frac{\partial g_2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}|_{\check{\boldsymbol{\beta}}} = \mathbf{0}, \quad \text{(C.10)} \\
\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta}^2} &= 2[(1-\alpha_1^*-\gamma_1^*-\gamma_2^*)\boldsymbol{M}_{g_1} + (\alpha_1^*+\gamma_1^*)\boldsymbol{M}_{h_1} + \gamma_2^*\boldsymbol{M}_{g_2}] \succeq \mathbf{0}, \quad \text{(C.11)} \\
& C_1(\check{\boldsymbol{\beta}}) = 0, C_2(\check{\boldsymbol{\beta}}) = 0. \quad \text{(C.12)}
\end{aligned}$$

From (C.10), we have

$$\begin{aligned}
\mathbf{0} &= [(1-\alpha_1^*-\gamma_1^*-\gamma_2^*)\boldsymbol{M}_{g_1} + (\alpha_1^*+\gamma_1^*)\boldsymbol{M}_{h_1} + \gamma_2^*\boldsymbol{M}_{g_2}]\,\check{\boldsymbol{\beta}} - (1-\alpha_1^*-\gamma_1^*-\gamma_2^*)\boldsymbol{E}_{g_1} \\
&\quad -(\alpha_1^*+\gamma_1^*)\boldsymbol{E}_{h_1} - \gamma_2^*\boldsymbol{E}_{g_2},
\end{aligned}$$

where $\boldsymbol{E}_{g_2} = C_{g_2}\boldsymbol{X}_1^T\boldsymbol{y_1} + D_{g_2}\boldsymbol{X}_2^T\boldsymbol{y_2}$. Then we have

$$\begin{aligned}
\check{\boldsymbol{\beta}} = &[(1-\alpha_1^*-\gamma_1^*-\gamma_2^*)\boldsymbol{M}_{g_1} + (\alpha_1^*+\gamma_1^*)\boldsymbol{M}_{h_1} + \gamma_2^*\boldsymbol{M}_{g_2}]^{-1} \\
&\cdot [(1-\alpha_1^*-\gamma_1^*-\gamma_2^*)\boldsymbol{E}_{g_1} + (\alpha_1^*+\gamma_1^*)\boldsymbol{E}_{h_1} + \gamma_2^*\boldsymbol{E}_{g_2}]\,. \quad \text{(C.13)}
\end{aligned}$$

Plugging (C.13) into (C.12), we have

$$K_1(\gamma_1,\gamma_2) = C_1(\check{\boldsymbol{\beta}}) = 0, K_2(\gamma_1,\gamma_2) = C_2(\check{\boldsymbol{\beta}}) = 0,$$

with solution (γ_1^*, γ_2^*). By substituting $\gamma_1 = \gamma_1^*, \gamma_2 = \gamma_2^*$ into (C.13), we obtain the solution for $\check{\boldsymbol{\beta}}$.

For step 3, the verification process is given as follows.

(1) If $\alpha_1^* + \gamma_1^* > 0$ and $\gamma_2^* > 0$, (4.14), (4.15), (4.16) are satisfied for $\alpha_1 = \alpha_1^* + \gamma_1^*, \alpha_2 = \gamma_2^*$ and $\boldsymbol{\beta} = \check{\boldsymbol{\beta}}$ based on (C.10), (C.11), (C.12). If we further have $C_3(\check{\boldsymbol{\beta}}) \geq 0$, then $\check{\boldsymbol{\beta}}$ is a global minimizer of (4.5).

(2) If $\alpha_1^* + \gamma_1^* < 0$, we could consider the minimization of $h_1(\boldsymbol{\beta})$.

(3) If $\gamma_2^* < 0$, we consider the minimization of $g_2(\boldsymbol{\beta})$.

C.4 Proof of Lemma 13

Note that $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ are independent without considering the optimization on η_{c_1}. In particular, the first term in $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ only involves $\boldsymbol{c}_1$ and the second term in $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ only involves $\boldsymbol{c}_2$. Thus, we firstly focus on the first term in $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ and solve the maximization with respect to $\boldsymbol{c}_1$.

$$\begin{aligned} & \max_{\eta_{c_1}} \max_{\|\boldsymbol{d}\|_2 \le 1} \max_{\|\boldsymbol{c}_1\|_2 = \eta_{c_1}} \|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta} - \boldsymbol{c}_1\boldsymbol{d}^T\boldsymbol{\beta}\|_2^2 \\ = & \max_{\eta_{c_1}} \max_{\|\boldsymbol{d}\|_2 \le 1} \max_{\|\boldsymbol{c}_1\|_2 = \eta_{c_1}} (\boldsymbol{d}^T\boldsymbol{\beta})^2 \|\boldsymbol{e}_1\|_2^2 \\ = & \max_{\eta_{c_1}} \max_{\|\boldsymbol{d}\|_2 \le 1} (\boldsymbol{d}^T\boldsymbol{\beta})^2 \max_{\|\boldsymbol{c}_1\|_2 = \eta_{c_1}} \|\boldsymbol{e}_1\|_2^2, \end{aligned}$$

in which $\boldsymbol{e}_1 = \boldsymbol{f}_1 - \boldsymbol{c}_1$ with $\boldsymbol{f}_1 = \frac{1}{\boldsymbol{d}^T\boldsymbol{\beta}}(\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta})$. For the maximization problem on $\boldsymbol{c}_1$, we have

$$\begin{aligned} & \max_{\boldsymbol{c}_1} \|\boldsymbol{e}_1\|_2^2, \ \text{ s.t. } \|\boldsymbol{c}_1\|_2 = \eta_{c_1}, \\ \Longleftrightarrow \ & \min_{\boldsymbol{e}_1} -\|\boldsymbol{e}_1\|_2^2, \ \text{ s.t. } \|\boldsymbol{f}_1 - \boldsymbol{e}_1\|_2^2 = \eta_{c_1}^2. \end{aligned} \tag{C.14}$$

Although (C.14) is not a convex optimization problem, we can first investigate its KKT necessary conditions. The Lagrangian function of (C.14) is $\mathcal{L}(\boldsymbol{e}_1, \gamma_{\boldsymbol{e}_1}) = -\|\boldsymbol{e}_1\|_2^2 + \gamma_{\boldsymbol{e}_1}(\|\boldsymbol{f}_1 - \boldsymbol{e}_1\|_2^2 - \eta_{c_1}^2)$, where $\gamma_{\boldsymbol{e}_1}$ is the Lagrangian multiplier. According to the KKT conditions, we have

$$\begin{aligned} & \partial\mathcal{L}(\boldsymbol{e}_1, \gamma_{\boldsymbol{e}_1}) = -2\boldsymbol{e}_1^T - 2\gamma_{\boldsymbol{e}_1}(\boldsymbol{f}_1 - \boldsymbol{e}_1)^T = \boldsymbol{0}, \\ & \|\boldsymbol{f}_1 - \boldsymbol{e}_1\|_2^2 = \eta_{c_1}^2, \end{aligned}$$

from which we can derive that the solution to (C.14) is $\boldsymbol{e}_1^* = \boldsymbol{f}_1 + \frac{\eta_{c_1}}{\|\boldsymbol{f}_1\|_2}\boldsymbol{f}_1$, and the maximum value is

$$\max_{\|\boldsymbol{c}_1\|_2 = \eta_{c_1}} \|\boldsymbol{e}_1\|_2^2 = \|\boldsymbol{e}_1^*\|_2^2 = \left(1 + \frac{\eta_{c_1}}{\|\boldsymbol{f}_1\|_2}\right)^2 \|\boldsymbol{f}_1\|_2^2.$$

Then we focus on the second term in $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$, solve the maximization on η_{c_2}, and derive the formulation for $g_{m_1}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d})$.

C.5 Proof of Proposition 9

We observe that $g_{m_1}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d})$ is a quadratic function with respect to $\boldsymbol{d}^T\boldsymbol{\beta}$, i.e.

$$g_{m_1}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d}) = A(\boldsymbol{d}^T\boldsymbol{\beta})^2 + B(\boldsymbol{d}^T\boldsymbol{\beta}) + C, \tag{C.15}$$

in which A, B, C are three coefficients. In particular, we have

$$\begin{aligned} A &= (C_g - D_g)\,\eta_{c_1}^2 + D_g\eta_c^2, && \text{(C.16)} \\ B &= 2\left[C_g\eta_{c_1}\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + D_g\eta_{c_2}\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2\right] \geq 0, \\ C &= C_g\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2 + D_g\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2^2. && \text{(C.17)} \end{aligned}$$

Since $A > 0$, $-\frac{B}{2A} \leq 0$ and $\boldsymbol{d}^T\boldsymbol{\beta} \in [-\|\boldsymbol{\beta}\|_2, \|\boldsymbol{\beta}\|_2]$, we can conclude that the maxima of $g_{m_1}(\boldsymbol{d}|\eta_{c_1}, \boldsymbol{\beta})$ is attained when $\boldsymbol{d}^T\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_2$ and the maximum value is

$$\begin{aligned} &\max_{\|\boldsymbol{d}\|_2 \leq 1} g_{m_1}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d}) \\ = \;& C_g(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 \\ &+ D_g(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta_c^2 - \eta_{c_1}^2}\|\boldsymbol{\beta}\|_2)^2, \end{aligned}$$

which provides the form of g_a.

C.6 Proof of Lemma 14

In this case, the analysis for the first term in $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ remains the same. However, for the second term, we have

$$\begin{aligned} & \min_{\eta_{c_2}} \min_{\|\boldsymbol{d}\|_2 \leq \frac{\eta}{\eta_{c_2}}} \min_{\|\boldsymbol{c}_2\|_2 \leq \eta_{c_2}} \|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta} - \boldsymbol{c}_2\boldsymbol{d}^T\boldsymbol{\beta}\|_2^2 \\ = & \min_{\eta_{c_2}} \min_{\|\boldsymbol{d}\|_2 \leq \frac{\eta}{\eta_{c_2}}} (\boldsymbol{d}^T\boldsymbol{\beta})^2 \min_{\|\boldsymbol{f}_2 - \boldsymbol{e}_2\|_2 \leq \eta_{c_2}} \|\boldsymbol{e}_2\|_2^2, \end{aligned}$$

where $\eta_{c_2} = \sqrt{\eta_c^2 - \eta_{c_1}^2}$, $\boldsymbol{f}_2 = \frac{1}{\boldsymbol{d}^T\boldsymbol{\beta}}(\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta})$ and $\boldsymbol{e}_2 = \boldsymbol{f}_2 - \boldsymbol{c}_2$. Thus, the minimization on $\boldsymbol{e}_2$ is a convex problem. By exploring the KKT conditions of the minimization problem, we are able to find the optimal solution. Particularly, the Lagrangian function of the minimization problem on $\boldsymbol{e}_2$ is

$$\mathcal{L}(\boldsymbol{e}_2, \gamma_{\boldsymbol{e}_2}) = \|\boldsymbol{e}_2\|_2^2 + \gamma_{\boldsymbol{e}_2}(\|\boldsymbol{f}_2 - \boldsymbol{e}_2\|_2^2 - \eta_{c_2}^2),$$

in which $\gamma_{\boldsymbol{e}_2}$ is the Lagrangian multiplier. By exploring the KKT conditions, we have

$$\nabla\mathcal{L}(\boldsymbol{e}_2, \gamma_{\boldsymbol{e}_2}) = 2\boldsymbol{e}_2^T - 2\gamma_{\boldsymbol{e}_2}(\boldsymbol{f}_2 - \boldsymbol{e}_2)^T = 0, \tag{C.18}$$

$$\|\boldsymbol{f}_2 - \boldsymbol{e}_2\|_2^2 \leq \eta_{c_2}^2,$$

$$\gamma_{\boldsymbol{e}_2}(\|\boldsymbol{f}_2 - \boldsymbol{e}_2\|_2^2 - \eta_c^2) = 0, \tag{C.19}$$

$$\gamma_{\boldsymbol{e}_2} \geq 0.$$

By inspecting the complementary slackness condition (C.19), we consider two cases based on the value of $\gamma_{\boldsymbol{e}_2}$.

Case 1: $\gamma_{\boldsymbol{e}_2} = 0$. In this case, we have $\boldsymbol{e}_2 = \boldsymbol{0}$ according to (C.18), which can be true when $\|\boldsymbol{f}_2\|_2 \leq \eta_{c_2}$. Moreover, note that

$$\|\boldsymbol{f}_2\|_2 \leq \eta_{c_2} \iff \|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 \leq |\boldsymbol{d}^T\boldsymbol{\beta}|\eta_{c_2} \overset{(a)}{\leq} \frac{\eta}{\eta_{c_2}}\|\boldsymbol{\beta}\|_2\eta_{c_2} = \eta\|\boldsymbol{\beta}\|_2,$$

where the equality in (a) is achieved if $\boldsymbol{d} = \frac{\eta}{\eta_{c_2}\|\boldsymbol{\beta}\|_2}\boldsymbol{\beta}$. Thus, if $\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 \leq \eta\|\boldsymbol{\beta}\|_2$, the minimum value of $\|\boldsymbol{e}_2\|_2^2$ is 0.

Case 2: $\gamma_{\boldsymbol{e}_2} > 0$. If there is no feasible solution in **Case 1**, we can conclude that $\|\boldsymbol{f}_2\|_2 > \eta_{c_2}$. Moreover, by (C.18) and (C.19), we have $\boldsymbol{e}_2^* = \frac{\gamma_{\boldsymbol{e}_2}^*\boldsymbol{f}_2}{\gamma_{\boldsymbol{e}_2}^*+1}$, $\eta_{c_2} = \|\boldsymbol{f}_2 - \boldsymbol{e}_2^*\|_2 = \frac{1}{\gamma_{\boldsymbol{e}_2}^*+1}\|\boldsymbol{f}_2\|_2$, which implies $\gamma_{\boldsymbol{e}_2}^* = \frac{\|\boldsymbol{f}_2\|_2}{\eta_{c_2}} - 1$, $\boldsymbol{e}_2^* = \boldsymbol{f}_2 - \frac{\eta_{c_2}}{\|\boldsymbol{f}_2\|_2}\boldsymbol{f}_2$. Then we have $\min_{\|\boldsymbol{f}_2 - \boldsymbol{e}_2\|_2 \leq \eta_{c_2}} \|\boldsymbol{e}_2\|_2^2 = \|\boldsymbol{e}_2^*\|_2^2 = \left(1 - \frac{\eta_{c_2}}{\|\boldsymbol{f}_2\|_2}\right)^2 \|\boldsymbol{f}_2\|_2^2$. By combining these two cases, Lemma 14 is proved.

C.7 Proof of Proposition 10

Now we solve the maximization problem on $\boldsymbol{d}$. Firstly, consider the case when $\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 \leq \eta\|\boldsymbol{\beta}\|_2$. In this case, we notice that as long as $\eta_{c_1} \neq 0$, $g_{m_2}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d})$ is a quadratic function for $\boldsymbol{d}^T\boldsymbol{\beta}$ with $A = C_g\eta_{c_1}^2 > 0$, $B = 2C_g\eta_{c_1}\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 \geq 0$ and $-\frac{B}{2A} \leq 0$. Thus, the maxima is attained when $\boldsymbol{d}^T\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_2$ and the maximum value of $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ is

$$g_{b_1}(\eta_{c_1}, \boldsymbol{\beta}) = C_g(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2.$$

For $\eta_{c_1} = 0$, the attacker only changes the feature matrix of the second group and the maximum value of $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ can also be derived as $g_{b_1}(\eta_{c_1}, \boldsymbol{\beta})$.

Secondly, consider the case when $\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 > \eta\|\boldsymbol{\beta}\|_2$. In this case, $g_{m_2}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d})$ can also be written in the form of (C.15) with coefficients A, B, C. In particular, A and C are defined the same as (C.16) and (C.17), and B is defined as $B = 2C_g\eta_{c_1}\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 - 2D_g\eta_{c_2}\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 \geq 0$. Since the coefficient of the quadratic term A can be positive, negative or zero, the maxima of g_{m_2} varies. By investigating into these three different cases, we have that when $\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 > \eta\|\boldsymbol{\beta}\|_2$, the maximum value of $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ is $g_{b_2}(\eta_{c_1}, \boldsymbol{\beta})$.

If $A > 0$, we have $-\frac{B}{2A} \leq 0$ and the maxima is attained when $\boldsymbol{d}^T\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_2$ with the maximum value to be $\max_{\|\boldsymbol{d}\|_2 \leq 1} g_{m_2}(\eta_{c_1}, \boldsymbol{\beta}, \boldsymbol{d}) = C_g(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 + D_g(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 -$

$\eta_{c_2}\|\boldsymbol{\beta}\|_2)^2$, which implies that

$$\begin{aligned}
& \max_{\|\boldsymbol{c}\boldsymbol{d}^T\|_F\leq\eta} g(\boldsymbol{\beta},\hat{\boldsymbol{X}}) \\
= & \max_{0<\eta_c\leq\eta}\max_{0\leq\eta_{c_1}\leq\eta_c}\max_{\|\boldsymbol{d}\|_2\leq 1} g_{m_2}(\eta_{c_1},\boldsymbol{\beta},\boldsymbol{d}) \\
\overset{(a)}{=} & \max_{0\leq\eta_{c_1}\leq\eta}\left[C_g(\|\boldsymbol{y}_1-\boldsymbol{X}_1\boldsymbol{\beta}\|_2+\eta_{c_1}\|\boldsymbol{\beta}\|_2)^2+D_g(\|\boldsymbol{y}_2-\boldsymbol{X}_2\boldsymbol{\beta}\|_2-\max_{0<\eta_c\leq\eta}\eta_{c_2}\|\boldsymbol{\beta}\|_2)^2\right] \\
= & \max_{0\leq\eta_{c_1}\leq\eta} g_{b_2}(\eta_{c_1},\boldsymbol{\beta}),
\end{aligned}$$

where (a) follows from the fact that $D_g<0$ and $\|\boldsymbol{y}_2-\boldsymbol{X}_2\boldsymbol{\beta}\|_2>\eta\|\boldsymbol{\beta}\|_2\geq\eta_{c_2}\|\boldsymbol{\beta}\|_2$.

If $A=0$, from the expression of A, we have $\eta_{c_1}^2=\frac{-\frac{1}{n}+\frac{\lambda}{n-m}}{\frac{\lambda}{m}+\frac{\lambda}{n-m}}\eta_c^2$, which is feasible as $\frac{-\frac{1}{n}+\frac{\lambda}{n-m}}{\frac{\lambda}{m}+\frac{\lambda}{n-m}}\in(0,1)$. Then since $B\geq 0$, g_{m_2} is a linearly non-decreasing function in $\boldsymbol{d}^T\boldsymbol{\beta}$ and the maxima is attained when $\boldsymbol{d}^T\boldsymbol{\beta}=\|\boldsymbol{\beta}\|_2$ with the maximum value to be the same as $g_{b_2}(\eta_{c_1},\boldsymbol{\beta})$.

Otherwise, if $A<0$, g_{m_2} is a concave quadratic function in $\boldsymbol{d}^T\boldsymbol{\beta}$ with $-\frac{B}{2A}>\frac{\left(\frac{\lambda}{n-m}-\frac{1}{n}\right)\eta_{c_2}\|\boldsymbol{y}_2-\boldsymbol{X}_2\boldsymbol{\beta}\|_2}{-\left(\frac{\lambda}{n-m}-\frac{1}{n}\right)\eta_{c_1}^2+\left(\frac{\lambda}{n-m}-\frac{1}{n}\right)\eta_c^2}\overset{(g)}{>}\frac{\eta_{c_2}\eta\|\boldsymbol{\beta}\|_2}{\eta_{c_2}^2}\geq\|\boldsymbol{\beta}\|_2$, in which (g) is from the fact that $\|\boldsymbol{y}_2-\boldsymbol{X}_2\boldsymbol{\beta}\|_2>\eta\|\boldsymbol{\beta}\|_2$. Thus, the maxima is attained when $\boldsymbol{d}^T\boldsymbol{\beta}=\|\boldsymbol{\beta}\|_2$ and the maximum value is also $g_{b_2}(\eta_{c_1},\boldsymbol{\beta})$.

C.8 Proof of Lemma 15

Since the forms of $g_a, g_{b_1}, g_{b_2}, h_a, h_{b_1}, h_{b_2}$ are similar, we only show the weakly-convex-weakly-concave property of g_a. For η_{c_1}, we have $\frac{\partial^2 g_a(\eta_{c_1},\boldsymbol{\beta})}{\partial\eta_{c_1}^2}=2\left(\frac{\lambda}{m}+\frac{\lambda}{n-m}\right)\|\boldsymbol{\beta}\|_2^2-2D_g\frac{\eta^2}{\eta_{c_2}^3}\|\boldsymbol{\beta}\|_2\|\boldsymbol{y}_2-\boldsymbol{X}_2\boldsymbol{\beta}\|_2$. Since $D_g\geq 0$, as long as $\|\boldsymbol{\beta}\|_2$ is bounded, there always exist a constant $\rho_1<\infty$ such that $\frac{\partial^2 g_a(\eta_{c_1},\boldsymbol{\beta})}{\partial\eta_{c_1}^2}\leq\rho_1$, indicating that g_a is weakly-concave in η_{c_1}.

For $\boldsymbol{\beta}$, we have

$$\begin{aligned}
\frac{\partial^2 g_a(\eta_{c_1},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2}\geq & 2C_g\left[\eta_{c_1}\left(\eta_{c_1}-2\frac{\mathrm{Tr}(\boldsymbol{X}_1^T\boldsymbol{X}_1)}{\|\boldsymbol{X}_1\|_F}\right)\boldsymbol{I}+\boldsymbol{X}_1^T\boldsymbol{X}_1\right] \\
& +2D_g\left[\eta_{c_2}\left(\eta_{c_2}-2\frac{\mathrm{Tr}(\boldsymbol{X}_2^T\boldsymbol{X}_2)}{\|\boldsymbol{X}_2\|_F}\right)\boldsymbol{I}+\boldsymbol{X}_2^T\boldsymbol{X}_2\right].
\end{aligned}$$

Since $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are feature matrices with finite norm, there always exist $\rho_2<\infty$ such that

$\frac{\partial^2 g_a(\eta_{c_1},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \succeq -\rho_2 \boldsymbol{I}$, which indicates that g_a is weakly-convex in $\boldsymbol{\beta}$.

C.9 Proof of Lemma 16

For g_a, we have

$$\frac{\partial g_a(\eta_{c_1},\boldsymbol{\beta})}{\partial \eta_{c_1}} = 2C_g\|\boldsymbol{\beta}\|_2\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + 2\left(C_g - D_g\right)\eta_{c_1}\|\boldsymbol{\beta}\|_2^2 - 2D_g\frac{\eta_{c_1}}{\eta_{c_2}}\|\boldsymbol{\beta}\|_2\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 = 0,$$

which implies

$$\begin{aligned}
&\left(\eta_{c_1}\|\boldsymbol{\beta}\|_2 + \frac{C_g}{C_g - D_g}\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2\right)\cdot\left(\eta_{c_2}\|\boldsymbol{\beta}\|_2 - \frac{D_g}{C_g - D_g}\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2\right) \\
&= -\frac{C_g D_g}{\left(\frac{\lambda}{n-m} + \frac{\lambda}{m}\right)^2}\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2. \qquad \text{(C.20)}
\end{aligned}$$

From (C.20), we note that η_{c_1} and η_{c_2} are inversely proportional. Since we also have $\eta_{c_1}^2 + \eta_{c_2}^2 = \eta^2$, $\eta_{c_1} \geq 0$ and $\eta_{c_2} \geq 0$, there is a unique solution for (C.20) (which can be seen geometrically), denoted as $\eta_{c_1}^*$. Moreover, we have

- $\eta_{c_1} < \eta_{c_1}^*$, left hand side of (C.20) is positive;
- $\eta_{c_1} > \eta_{c_1}^*$, left hand side of (C.20) is negative.

Thus, g_a is a unimodal function that increases first and then decreases. The results can be easily generalized to other sub-functions.

Bibliography

[1] Sid Ahmed Fezza, Yassine Bakhti, Wassim Hamidouche, and Olivier Déforges. Perceptual evaluation of adversarial attacks for CNN-based image classification. In *Proc. Eleventh Int. Conference on Quality of Multimedia Experience*, pages 1–6, Berlin, Germany, Jun. 2019.

[2] Piergiuseppe Mallozzi, Patrizio Pelliccione, Alessia Knauss, Christian Berger, and Nassar Mohammadiha. Autonomous vehicles: state of the art, future trends, and challenges. *Automotive systems and software engineering: State of the art and future trends*, pages 347–367, Jul. 2019.

[3] Cumhur Erkan Tuncali, Georgios Fainekos, Hisahiro Ito, and James Kapinski. Simulation-based adversarial test generation for autonomous vehicles with machine learning components. In *proc. IEEE Intelligent Vehicles Symposium*, pages 1555–1562, Changshu, China, Jun. 2018.

[4] Harsha Jakkanahalli Vishnukumar, Björn Butting, Christian Müller, and Eric Sax. Machine learning and deep neural network—artificial intelligence core for lab and real-world test and validation for adas and autonomous vehicles: Ai for efficient and quality test and validation. In *proc. Intelligent Systems Conference*, pages 714–721, London, United Kingdom, Sep. 2017.

[5] Li Gao, Hongjie Jiang, Kaiming Fu, and Weikai He. On understanding degradation kinetics of pharmaceutic gelatin matrices for precision medicine: A deep learning approach. In

proc. IEEE International Conference on Big Data, pages 6060–6062, Los Angeles, CA, Dec. 2019.

[6] Hema Sekhar Reddy Rajula, Giuseppe Verlato, Mirko Manchia, Nadia Antonucci, and Vassilios Fanos. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, 56(9):455, Sep. 2020.

[7] Mohamed El-Shamouty, Kilian Kleeberger, Arik Lämmle, and Marco Huber. Simulation-driven machine learning for robotics and automation. *tm-Technisches Messen*, 86(11):673–684, Nov. 2019.

[8] Mohsen Soori, Behrooz Arezoo, and Roza Dastres. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, Apr. 2023.

[9] Tao Li, Lei Lin, Minsoo Choi, Kaiming Fu, Siyuan Gong, and Jian Wang. Youtube av 50k: an annotated corpus for comments in autonomous vehicles. In *proc. International Joint Symposium on Artificial Intelligence and Natural Language Processing*, pages 1–5, Pattaya, Thailand, Nov. 2018.

[10] Jayashree Padmanabhan and Melvin Jose Johnson Premkumar. Machine learning in automatic speech recognition: A survey. *IETE Technical Review*, 32(4):240–251, Jul. 2015.

[11] Sankar K Pal, Anima Pramanik, Jhareswar Maiti, and Pabitra Mitra. Deep learning in multi-object detection and tracking: state of the art. *Applied Intelligence*, 51:6400–6429, Sep. 2021.

[12] Pulkit Goel. Realtime object detection using tensorflow an application of ml. *International Journal of Sustainable Development in Computing Science*, 3(3):11–20, Oct. 2021.

[13] Kaiming Fu, Yulu Jin, and Zhousheng Chen. Test set optimization by machine learning algorithms. In *proc. IEEE International Conference on Big Data*, pages 5673–5675, Dec. 2020.

[14] Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165:113986, Mar. 2021.

[15] Candice Schumann, Jeffrey Foster, Nicholas Mattei, and John Dickerson. We need fairness and explainability in algorithmic hiring. In *proc. International Conference on Autonomous Agents and Multi-Agent Systems*, May 2020.

[16] Sarah Brayne and Angèle Christin. Technologies of crime prediction: The reception of algorithms in policing and criminal courts. *Social problems*, 68(3):608–624, Aug. 2021.

[17] Yanlin Qi, Jia Li, and Michael Zhang. Enabling cmf estimation in data-constrained scenarios: A semantic-encoding knowledge mining model. *arXiv preprint arXiv:2311.08690*, Nov. 2023.

[18] Yanlin Qi, Feng Cheng, and Muyang Wang. A new approach for personal safety forewarning based on gps positioning and spatiotemporal analysis. In *proc. International Conference on Information and Automation*, pages 3157–3162, Yunnan, China, Aug. 2015.

[19] Tao Li, Kaiming Fu, Minsoo Choi, Xudong Liu, and Ying Chen. Toward robust and efficient training of generative adversarial networks with bayesian approximation. In *proc. Approximation Theory and Machine Learning Conference*, volume 6, West Lafayette, IN, Sep. 2018.

[20] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, Mar. 2019.

[21] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, Jan. 2017.

[22] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, Dec. 2017.

[23] Yulu Jin and Lifeng Lai. On the adversarial robustness of hypothesis testing. *IEEE Transactions on Signal Processing*, 69:515–530, Dec. 2020.

[24] Fuwei Li, Lifeng Lai, and Shuguang Cui. Optimal feature manipulation attacks against linear regression. *IEEE Transactions on Signal Processing*, 69:5580–5594, Sep. 2021.

[25] Yanlin Qi, Gengchen Mai, Rui Zhu, and Michael Zhang. Evkg: An interlinked and interoperable electric vehicle knowledge graph for smart transportation system. *Transactions in GIS*, Apr. 2023.

[26] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

[27] Cynthia Dwork. Differential privacy: A survey of results. In *Proc. Int. Conference on Theory and Applications of Models of Computation*, pages 1–19, Xi'an, China, Apr. 2008.

[28] Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *Proc. Annual Allerton Conference on Communication, Control, and Computing*, pages 1401–1408, Monticello, IL, Oct. 2012.

[29] Cornelius Glackin, Gerard Chollet, Nazim Dugan, Nigel Cannings, Julie Wall, Shahzaib Tahir, Indranil Ghosh Ray, and Muttukrishnan Rajarajan. Privacy preserving encrypted phonetic search of speech data. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pages 6414–6418, New Orleans, LA, Mar. 2017.

[30] Xin Wang, Hideaki Ishii, Linkang Du, Peng Cheng, and Jiming Chen. Privacy-preserving distributed machine learning via local randomization and ADMM perturbation. *IEEE Transactions on Signal Processing*, 68:4226–4241, Jul. 2020.

[31] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[32] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proc. International Conference on Machine Learning*, pages 6755–6764, Vienna, Austria, Nov. 2020.

[33] Yulu Jin and Lifeng Lai. Privacy protection in learning fair representations. In *proc. International Conference on Acoustics, Speech and Signal Processing*, pages 2964–2968, Singapore, Jun. 2022.

[34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv pp.1312.6199*, Dec. 2013.

[35] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Proc. IEEE symposium on security and privacy*, pages 582–597, San Jose, CA, May 2016.

[36] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. Int. Conference on Learning Representations*, San Diego, CA, May. 2015.

[37] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proc. IEEE Symposium on Security and Privacy*, pages 39–57, San Jose, CA, May. 2017.

[38] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv pp.1710.11342*, Oct. 2017.

[39] Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya Nori, and Antonio Criminisi. Measuring neural net robustness with constraints. In *Proc. Advances in Neural Information Processing Systems*, pages 2613–2621, Quebec, Canada, Dec. 2016.

[40] Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, Sep. 2018.

[41] Nicholas Carlini and David Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

[42] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[43] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Fundamental limits on adversarial robustness. In *Proc. Int. Conference on Machine Learning, Workshop on Deep Learning*, Lille, France, Jul. 2015.

[44] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Proc. Advances in Neural Information Processing Systems*, pages 1178–1187, Quebec, Canada, Dec. 2018.

[45] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, Oxford, UK, Feb. 2013.

[46] Battista Biggio, Paolo Russu, Luca Didaci, and Fabio Roli. Adversarial biometric recognition : A review on biometric system security from the adversarial machine-learning perspective. *IEEE Signal Processing Magazine*, 32(5):31–41, Aug. 2015.

[47] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology*, 11(3):1–41, Apr. 2020.

[48] Saurabh Sihag and Ali Tajer. Secure estimation under causative attacks. *IEEE Transactions on Information Theory*, 66(8):5145–5166, Aug. 2020.

[49] Meghana Bande and Venugopal V. Veeravalli. Adversarial multi-user bandits for uncoordinated spectrum access. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pages 1–5, Brighton, UK, May 2019.

[50] Yuanxi Yang. Robust estimation for dependent observations. *Manuscripta geodaetica*, 19(1):10–17, Oct. 1994.

[51] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, San Francisco, CA, 2011.

[52] Peter J Huber. *Robust statistics*. Springer, New York, NY, 2011.

[53] Gökhan Gül and Abdelhak M Zoubir. Robust hypothesis testing with α -divergence. *IEEE Transactions on Signal Processing*, 64(18):4737–4750, May. 2016.

[54] Serkan Sarıtaş, Sinan Gezici, and Serdar Yüksel. Hypothesis testing under subjective priors and costs as a signaling game. *IEEE Transactions on Signal Processing*, 67(19):5169–5183, Aug. 2019.

[55] Nir Halay, Koby Todros, and Alfred O Hero. Binary hypothesis testing via measure transformed quasi-likelihood ratio test. *IEEE Transactions on Signal Processing*, 65(24):6381–6396, Sep. 2017.

[56] Gökhan Gül and Abdelhak M Zoubir. Minimax robust hypothesis testing. *IEEE Transactions on Information Theory*, 63(9):5572–5587, Apr. 2017.

[57] Bernard C Levy. Robust hypothesis testing with a relative entropy tolerance. *IEEE Transactions on Information Theory*, 55(1):413–421, Dec. 2008.

[58] Lifeng Lai and Erhan Bayraktar. On the adversarial robustness of robust estimators. *IEEE Transactions on Information Theory*, 66(8):5097–5109, Apr. 2020.

[59] Tyler Hunt, Congzheng Song, Reza Shokri, Vitaly Shmatikov, and Emmett Witchel. Chiron: Privacy-preserving machine learning as a service. *arXiv preprint arXiv:1803.05961*, 2018.

[60] Francesca Meneghello, Matteo Calore, Daniel Zucchetto, Michele Polese, and Andrea Zanella. IoT: Internet of threats? a survey of practical security vulnerabilities in real IoT devices. *IEEE Internet of Things Journal*, 6(5):8182–8201, May. 2019.

[61] Alireza Ahrabian, Sefki Kolozali, Shirin Enshaeifar, Clive Cheong-Took, and Payam Barnaghi. Data analysis as a web service: A case study using iot sensor data. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pages 6000–6004, New Orleans, LA, Mar. 2017.

[62] Meng Sun, Wee Peng Tay, and Xin He. Toward information privacy for the internet of things: A nonparametric learning approach. *IEEE Transactions on Signal Processing*, 66(7):1734–1747, Jan. 2018.

[63] Jacob Wurm, Khoa Hoang, Orlando Arias, Ahmad-Reza Sadeghi, and Yier Jin. Security analysis on consumer and industrial IoT devices. In *Proc. Asia and South Pacific Design Automation Conference*, pages 519–524, Macao, China, Jan. 2016.

[64] Charith Perera, Arkady Zaslavsky, Peter Christen, and Dimitrios Georgakopoulos. Context aware computing for the internet of things: A survey. *IEEE Communications Surveys & Tutorials*, 16(1):414–454, Aug. 2013.

[65] Chengliang Zhang, Minchen Yu, Wei Wang, and Feng Yan. Mark: Exploiting cloud services for cost-effective, slo-aware machine learning inference serving. In *Proc. Annual Technical Conference*, pages 1049–1062, Renton, WA, Jul. 2019.

[66] Arpan Gujarati, Sameh Elnikety, Yuxiong He, Kathryn S McKinley, and Björn B Brandenburg. Swayam: Distributed autoscaling to meet slas of machine learning inference services with resource efficiency. In *Proc. ACM/IFIP/USENIX Middleware Conference*, pages 109–120, Las Vegas, NV, Dec. 2017.

[67] Ehsan Hesamifard, Hassan Takabi, Mehdi Ghasemi, and Catherine Jones. Privacy-preserving machine learning in cloud. In *Proc. ACM Conference on Computer and Communications Security*, pages 39–43, Dallas,TX, Nov. 2017.

[68] Tebaa Maha, E Saïd, and E Abdellatif. Homomorphic encryption applied to the cloud computing security. In *Proc. World Congress on Engineering*, volume 1, pages 4–6, London, U.K, Jul. 2012.

[69] Fabian Boemer, Anamaria Costache, Rosario Cammarota, and Casimir Wierzynski. ngraph-he2: A high-throughput framework for neural network inference on encrypted data. In *Proc. ACM Workshop on Encrypted Computing & Applied Homomorphic Cryptography*, pages 45–56, London, UK, Nov. 2019.

[70] Craig Gentry. *A fully homomorphic encryption scheme*, volume 20. Stanford university, 2009.

[71] Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Galen Reeves, and Guillermo Sapiro. Adversarially learned representations for information obfuscation and inference. In *Proc. Int. Conference on Machine Learning*, pages 614–623, Long Beach, CA, Jun. 2019.

[72] Jihun Hamm. Minimax filter: Learning to preserve privacy from inference attacks. *The Journal of Machine Learning Research*, 18(1):4704–4734, Apr. 2017.

[73] Ardhendu Tripathy, Ye Wang, and Prakash Ishwar. Privacy-preserving adversarial networks. In *Proc. 57th Annual Allerton Conference on Communication, Control, and Computing*, pages 495–505, Monticello, IL, Sep. 2019.

[74] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, Jan. 2018.

[75] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[76] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, Feb. 2019.

[77] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

[78] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proc. Int. conference on machine learning*, pages 325–333, Atlanta, Georgia, Jun. 2013.

[79] Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

[80] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*, 2018.

[81] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Proc. Conference on Fairness, Accountability and Transparency*, pages 119–133, New York, NY, Feb. 2018.

[82] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

[83] David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. *arXiv preprint arXiv:2004.07401*, Apr. 2020.

[84] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. *arXiv preprint arXiv:2012.08723*, Dec. 2020.

[85] Minh-Hao Van, Wei Du, Xintao Wu, and Aidong Lu. Poisoning attacks on fair machine learning. *arXiv preprint arXiv:2110.08932*, Oct. 2021.

[86] Hongyan Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi, and Reza Shokri. On adversarial bias and the robustness of fair machine learning. *arXiv preprint arXiv:2006.08669*, Jun. 2020.

[87] Ky Fan. Minimax theorems. *National Academy of Sciences of the United States of America*, 39(1):42, 1953.

[88] Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. Robust hypothesis testing using wasserstein uncertainty sets. In *Proc. Neural Information Processing Systems*, pages 7913–7923, Montreal, Canada, Dec. 2018.

[89] Peter J Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):73–79, Jan. 2011.

[90] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, Norwell, MA, 2011.

[91] Wei Deng and Wotao Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, May 2016.

[92] Bingsheng He and Xiaoming Yuan. On the o(1/n) convergence rate of the douglas–rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, Sep. 2012.

[93] Renato DC Monteiro and Benar F Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1):475–507, Oct. 2013.

[94] Caihua Chen, Bingsheng He, Yinyu Ye, and Xiaoming Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1-2):57–79, Apr. 2016.

[95] Bingsheng He, Min Tao, and Xiaoming Yuan. A splitting method for separable convex programming. *IMA Journal of Numerical Analysis*, 35(1):394–426, Sep. 2015.

[96] Tian-Yi Lin, Shi-Qian Ma, and Shu-Zhong Zhang. On the sublinear convergence rate of multi-block ADMM. *Journal of the Operations Research Society of China*, 3(3):251–274, Jun. 2015.

[97] Min Li, Defeng Sun, and Kim-Chuan Toh. A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block. *Asia-Pacific Journal of Operational Research*, 32(04):1550024, Jan. 2015.

[98] Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. In *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pages 3836–3840, Brisbane, Australia, Apr. 2015.

[99] Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. Iteration complexity analysis of multi-block ADMM for a family of convex minimization without strong convexity. *Journal of Scientific Computing*, 69(1):52–81, Mar. 2016.

[100] Ruoyu Sun, Zhi-Quan Luo, and Yinyu Ye. On the expected convergence of randomly permuted ADMM. *arXiv preprint arXiv:1503.06387*, 4(6), Dec. 2015.

[101] Mingyi Hong, Tsung-Hui Chang, Xiangfeng Wang, Meisam Razaviyayn, Shiqian Ma, and Zhi-Quan Luo. A block successive upper-bound minimization method of multipliers for linearly constrained convex optimization. *Mathematics of Operations Research*, 45(3):833–861, Feb. 2020.

[102] Ying Cui, Xudong Li, Defeng Sun, and Kim-Chuan Toh. On the convergence properties of a majorized alternating direction method of multipliers for linearly constrained convex optimization problems with coupled objective functions. *Journal of Optimization Theory and Applications*, 169(3):1013–1041, Mar. 2016.

[103] Xiang Gao and Shu-Zhong Zhang. First-order algorithms for convex optimization with nonseparable objective and coupled constraints. *Journal of the Operations Research Society of China*, 5(2):131–159, Jan. 2017.

[104] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, Mar. 1948.

[105] Andrei Serjantov and George Danezis. Towards an information theoretic metric for anonymity. In *Proc. Int. Workshop on Privacy Enhancing Technologies*, pages 41–53, San Francisco, CA, Apr. 2002. Springer.

[106] Sebastian Clauß and Stefan Schiffner. Structuring anonymity metrics. In *Proc. ACM Conference on Computer and Communications Security*, pages 55–62, Alexandria, VA, Nov. 2006.

[107] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, Oct. 2002.

[108] Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proc. ACM Symposium on Information, Computer and Communications Security*, pages 32–33, Seoul, Korea, May 2012.

[109] Charu C Aggarwal. On k-anonymity and the curse of dimensionality. In *Proc. Int. Conference on Very Large Data Bases*, volume 5, pages 901–909, Trondheim, Norway, Aug. 2005.

[110] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proc. Int. Conference on Management of Data*, pages 229–240, New York, NY, Jun. 2006.

[111] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. T-closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. Int. Conference on Data Engineering*, pages 106–115, Istanbul, Turkey, Apr. 2007.

[112] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[113] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proc. Int. Conference on Management of data*, pages 193–204, Athens, Greece, Jun. 2011.

[114] Márk Jelasity and Kenneth P Birman. Distributional differential privacy for large-scale smart metering. In *Proc. ACM workshop on Information hiding and multimedia security*, pages 141–146, Salzburg, Austria, Jun. 2014.

[115] Isabel Wagner. Genomic privacy metrics: A systematic comparison. In *Proc. IEEE Security and Privacy Workshops*, pages 50–59, San Jose, CA, May 2015.

[116] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proc. 21th Int. Conference on Knowledge Discovery and Data Mining*, pages 259–268, Sydney, Australia, Aug. 2015.

[117] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Proc. Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA, Dec. 2017.

[118] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the

country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

[119] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, Sep. 2016.

[120] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proc. International conference on world wide web*, pages 1171–1180, Perth, Australia, Apr. 2017.

[121] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Proc. Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA, Dec. 2017.

[122] Reuben Binns. On the apparent conflict between individual and group fairness. In *Proc. Conference on fairness, accountability, and transparency*, pages 514–524, Barcelona, Spain, Jan. 2020.

[123] Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. *Colum. Bus. L. Rev.*, page 494, 2019.

[124] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proc. Innovations in theoretical computer science conference*, pages 214–226, New York, NY, Jan. 2012.

[125] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Proc. Advances in neural information processing systems*, volume 29, Barcelona, Spain, Dec. 2016.

[126] Mauro Barni and Benedetta Tondi. Binary hypothesis testing game with training data. *IEEE Transactions on Information Theory*, 60(8):4848–4866, Jun. 2014.

[127] Mauro Barni and Benedetta Tondi. Multiple-observation hypothesis testing under adversarial conditions. In *Proc. IEEE Int. Workshop on Information Forensics and Security*, pages 91–96, Guangzhou, China, Nov. 2013.

[128] Mauro Barni and Benedetta Tondi. Adversarial source identification game with corrupted training. *IEEE Transactions on Information Theory*, 64(5):3894–3915, Feb. 2018.

[129] Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, May 2017.

[130] Taposh Banerjee, Hamed Firouzi, and Alfred O Hero. Quickest detection for changes in maximal KNN coherence of random matrices. *IEEE Transactions on Signal Processing*, 66(17):4490–4503, Jul. 2018.

[131] Yao Xie and David Siegmund. Sequential multi-sensor change-point detection. In *Proc. Information Theory and Applications Workshop*, pages 1–20. San Diego, CA, May 2013.

[132] Shaofeng Zou, Georgios Fellouris, and Venugopal V Veeravalli. Quickest change detection under transient dynamics: Theory and asymptotic analysis. *IEEE Transactions on Information Theory*, 65(3):1397–1412, Oct. 2018.

[133] H Vincent Poor and Olympia Hadjiliadis. *Quickest detection*. Cambridge University Press, 2008.

[134] Yuan Wang and Yajun Mei. Large-scale multi-stream quickest change detection via shrinkage post-change estimation. *IEEE Transactions on Information Theory*, 61(12):6926–6938, Dec. 2015.

[135] Olympia Hadjiliadis, Hongzhong Zhang, and H Vincent Poor. One shot schemes for decentralized quickest change detection. *IEEE Transactions on Information Theory*, 55(7):3346–3359, Jun. 2009.

[136] Rittwik Jana and Subhrakanti Dey. Change detection in teletraffic models. *IEEE Transactions on Signal Processing*, 48(3):846–853, Mar. 2000.

[137] Vasanthan Raghavan and Venugopal V Veeravalli. Quickest change detection of a Markov process across a sensor array. *IEEE Transactions on Information Theory*, 56(4):1961–1981, Mar. 2010.

[138] Georgios Fellouris, Erhan Bayraktar, and Lifeng Lai. Efficient Byzantine sequential change detection. *IEEE Transactions on Information Theory*, 64(5):3346–3360, Sep. 2017.

[139] Taposh Banerjee and Venugopal V Veeravalli. Data-efficient quickest change detection in sensor networks. *IEEE Transactions on Signal Processing*, 63(14):3727–3735, May 2015.

[140] Alexander G Tartakovsky and Venugopal V Veeravalli. Asymptotically optimal quickest change detection in distributed sensor systems. *Sequential Analysis*, 27(4):441–475, Oct. 2008.

[141] George V Moustakides. Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14(4):1379–1387, Nov. 1986.

[142] Shang Li, Yasin Yılmaz, and Xiaodong Wang. Quickest detection of false data injection attack in wide-area smart grids. *IEEE Transactions on Smart Grid*, 6(6):2725–2735, Dec. 2014.

[143] Di Li, Lifeng Lai, and Shuguang Cui. Quickest change detection and identification across a sensor array. In *Proc. IEEE Global Conference on Signal and Information Processing*, pages 145–148. Austin, TX, Dec. 2013.

[144] Yulu Jin and Lifeng Lai. Adversarially robust hypothesis testing. In *Asilomar Conference on Signals, Systems, and Computers*, pages 1806–1810, Pacific Grove, CA, Nov. 2019.

[145] Yulu Jin and Lifeng Lai. Privacy-accuracy trade-off of inference as service. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2645–2649, Jun. 2021.

[146] Yulu Jin and Lifeng Lai. Optimal accuracy-privacy trade-off of inference as service. *IEEE Transactions on Signal Processing*, 70:4031–4046, Jul. 2022.

[147] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. In *Proc. Advances in Neural Information Processing Systems*, volume 34, pages 815–827, Dec. 2021.

[148] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, Jun. 2018.

[149] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proc. International Conference on Machine Learning*, pages 3384–3393, Stockholm, Sweden, Jul. 2018.

[150] Philipp Benz, Chaoning Zhang, Soomin Ham, Adil Karjauv, Gyusang Cho, and In So Kweon. Trade-off between accuracy, robustness, and fairness of deep classifiers. 2021.

[151] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proc. ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, Mar. 2021.

[152] Qi Tian, Kun Kuang, Kelu Jiang, Fei Wu, and Yisen Wang. Analysis and applications of class-wise robustness in adversarial training. In *Proc. ACM Conference on Knowledge Discovery & Data Mining*, pages 1561–1570, Virtual Event, Singapore, Aug. 2021.

[153] Jie Jiang and Xiaojun Chen. Optimality conditions for nonsmooth nonconvex-nonconcave min-max problems and generative adversarial networks. *arXiv preprint arXiv:2203.10914*, Mar. 2022.

[154] Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proc. International Conference on Machine Learning*, pages 6083–6093, Jul. 2020.

[155] Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, Apr. 2020.

[156] Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *Proc. Advances in Neural Information Processing Systems*, volume 34, Dec. 2021.

[157] Dmitrii M Ostrovskii, Babak Barazandeh, and Meisam Razaviyayn. Nonconvex-nonconcave min-max optimization with a small maximization domain. *arXiv preprint arXiv:2110.03950*, Oct. 2021.

[158] Zhi-Quan Luo, Nicholas D Sidiropoulos, Paul Tseng, and Shuzhong Zhang. Approximation bounds for quadratic optimization with homogeneous quadratic constraints. *SIAM Journal on Optimization*, 18(1):1–28, Jan. 2007.

[159] Kejun Huang and Nicholas D Sidiropoulos. Consensus-admm for general quadratically constrained quadratic programming. *IEEE Transactions on Signal Processing*, 64(20):5297–5310, Jul. 2016.

[160] Mingrui Liu, Hassan Rafique, Qihang Lin, and Tianbao Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22:169–1, Jan. 2021.

[161] Yulu Jin and Lifeng Lai. Adversarially robust fairness-aware regression. In *proc. International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, Jun. 2023.

[162] Yulu Jin and Lifeng Lai. Fairness-aware regression robust to adversarial attacks. *IEEE Transactions on Signal Processing*, Nov. 2023.

[163] Yuan Chen, Soummya Kar, and Jose MF Moura. Resilient distributed estimation through adversary detection. *IEEE Transactions on Signal Processing*, 66(9):2455–2469, Mar. 2018.

[164] Fuwei Li, Lifeng Lai, and Shuguang Cui. On the adversarial robustness of subspace learning. *IEEE Transactions on Signal Processing*, 68:1470–1483, Mar. 2020.

[165] Jean-Pierre Aubin and Ivar Ekeland. *Applied nonlinear analysis*. Courier Corporation, North Chelmsford, MA, 2006.

[166] Rémi Cogranne, Tomas Denemark, and Jessica Fridrich. Theoretical model of the fld ensemble classifier based on hypothesis testing theory. In *Proc. IEEE Int. Workshop on Information Forensics and Security*, pages 167–172, Atlanta, GA, Dec. 2014.

[167] Domenico Ciuonzo, Pierluigi Salvo Rossi, and Peter Willett. Generalized Rao test for decentralized detection of an uncooperative target. *IEEE Signal Processing Letters*, 24(5):678–682, May 2017.

[168] Erfan Soltanmohammadi, Mahdi Orooji, and Mort Naraghi-Pour. Decentralized hypothesis testing in wireless sensor networks in the presence of misbehaving nodes. *IEEE Transactions on Information Forensics and Security*, 8(1):205–215, Nov. 2012.

[169] Domenico Ciuonzo, Antonio De Maio, and P Salvo Rossi. A systematic framework for composite hypothesis testing of independent Bernoulli trials. *IEEE Signal Processing Letters*, 22(9):1249–1253, Jan. 2015.

[170] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2233–2246, Jan. 2012.

[171] Wei Deng, Mingjun Lai, Zhimin Peng, and Wotao Yin. Parallel multi-block ADMM with o (1/k) convergence. *Journal of Scientific Computing*, 71(2):712–736, May 2017.

[172] Xiangfeng Wang, Mingyi Hong, Shiqian Ma, and Zhiquan Luo. Solving multiple-block separable convex minimization problems using two-block alternating direction method of multipliers. *arXiv preprint arXiv:1308.5294*, Aug. 2013.

[173] Albrecht Böttcher and David Wenzel. The frobenius norm and the commutator. *Linear algebra and its applications*, 429(8-9):1864–1885, Oct. 2008.

[174] Xingju Cai, Deren Han, and Xiaoming Yuan. On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. *Computational Optimization and Applications*, 66(1):39–73, Jan. 2017.

[175] Ying Cui, Xudong Li, Defeng Sun, and Chuan Toh. On the convergence properties of a majorized ADMM for linearly constrained convex optimization problems with coupled objective functions. *arXiv preprint arXiv:1502.00098*, Jan. 2015.

[176] Caihua Chen, Min Li, Xin Liu, and Yinyu Ye. Extended ADMM and BCD for nonseparable convex minimization models with quadratic coupling terms: convergence analysis and insights. *Mathematical Programming*, 173(1-2):37–77, Mar. 2019.

[177] Fenghui Wang, Wenfei Cao, and Zongben Xu. Convergence of multi-block bregman ADMM for nonconvex composite problems. *Science China Information Sciences*, 61(12):1–12, Dec. 2018.

[178] Bent Fuglede and Flemming Topsoe. Jensen-Shannon divergence and Hilbert space embedding. In *Proc. IEEE Int. Symposium on Information Theory*, page 31, Parma, Italy, Oct. 2004.

[179] Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, NY, 1997.

[180] Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, Berlin, Germany, 2012.

[181] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, Nov. 2016.

[182] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, Jun. 2017.

[183] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proc. Int. Conference on Machine Learning*, pages 7472–7482, Long Beach, CA, Jun. 2019.

[184] Fuwei Li, Lifeng Lai, and Shuguang Cui. On the adversarial robustness of subspace learning. *IEEE Transactions on Signal Processing*, 68:1470–1483, Mar. 2020.

[185] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *Proc. International Conference on Machine Learning*, pages 11492–11501, Jul. 2021.

[186] Jianfeng Chi, Yuan Tian, Geoffrey J Gordon, and Han Zhao. Understanding and mitigating accuracy disparity in regression. In *proc. International conference on machine learning*, pages 1866–1876, Jul. 2021.

[187] Abhin Shah, Yuheng Bu, Joshua Ka-Wing Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W Wornell. Selective regression under fairness criteria. *arXiv preprint arXiv:2110.15403*, Oct. 2021.

[188] Vaithilingam Jeyakumar, Alex M Rubinov, and Zhi-You Wu. Non-convex quadratic minimization problems with quadratic constraints: global optimality conditions. *Mathematical Programming*, 110(3):521–541, Sep. 2007.

[189] Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

[190] Brett Lantz. *Machine learning with R: expert techniques for predictive modeling*. Packt Publishing ltd, 2019.

[191] Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 339–355, Skopje, Macedonia, Sep. 2017. Springer.

[192] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, Berlin, Germany, 2003.

[193] Hatem Hmam. Quadratic optimisation with one quadratic equality constraint. Technical report, Defence Science and Technology Organisation Edinburgh (Australia) Electronic Warfare and Radar Division, Jun. 2010.