

# Fairness-Aware Regression Robust to Adversarial Attacks

Yulu Jin,  *Member, IEEE* and Lifeng Lai,  *Senior Member, IEEE*

**Abstract**—In this paper, 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.

**Index Terms**—Fairness, minimax problem, adversarial robustness.

## I. INTRODUCTION

MACHINE learning 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 been shown that machine learning algorithms can mirror or even amplify biases against population subgroups [2], [3], for example, based on race or sex. With direct social and economic impact on individuals, it is imperative to build ML models ethically and responsibly to avoid these biases. To this end various algorithms have been developed to find fair machine models (FML) that satisfy different fairness measures [4], [5], [6], [7].

In the meantime, a large body of work has shown that machine learning models are vulnerable to various types of attacks [8], [9], [10]. 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 [11], [12], [13], [14].

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. 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 [15]. Although there exist many iterative methods for finding stationary points or local optima of nonconvex-concave or nonconvex-nonconcave minimax problems [16], [17], [18], [19], 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  $\beta$  of the defense model. We then analyze the minimization problem by transforming it to four sub-problems where each sub-problem

Manuscript received 23 May 2023; revised 21 September 2023; accepted 17 October 2023. Date of publication 2 November 2023; date of current version 10 November 2023. This work was supported by the National Science Foundation under Grants ECCS-2000415 and CCF-2232907. An earlier version of this paper was presented in part at the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [DOI: 10.1109/ICASSP49357.2023.10094637]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ketan Rajawat. (*Corresponding author: Lifeng Lai.*)

The authors are with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA (e-mail: yuljin@ucdavis.edu; llfai@ucdavis.edu).

Digital Object Identifier 10.1109/TSP.2023.3328111

is a non-convex quadratic minimization problem with multiple quadratic constraints, which is usually NP hard [20], [21], 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,  $\beta_{rob}^*$ , corresponds to the optimal robust fairness-aware model, and the best adversarial data sample is obtained by fitting  $\beta_{rob}^*$  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 [22]. 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.

The main contributions of this paper are summarized as:

- 1) We formulate a minimax framework to design fairness-aware learning algorithms that are robust to adversarial attacks;
- 2) We explore the attack scheme where the adversary is able to add one carefully designed adversarial data point to the dataset, which provides important insights on adversary behavior and robust model design when a set of adversarial data points can be injected into the dataset;
- 3) We explore an attack scheme where the adversary can directly modify the feature matrix of existing data points, which complements the previous attack scheme;
- 4) We propose novel algorithms to solve the challenging nonsmooth nonconvex-nonconcave minimax problems that arise in our proposed framework. Our approach involves carefully designed transformations that maintain the essence of the original problems while making them computationally tractable.

This journal paper is an extension of our conference paper [1], where we exclusively focused on the rank-one attack scheme, in which the adversary modifies the feature matrix without altering the group index. In this journal paper, we also explore another attack scheme where additional data samples can be added to the existing dataset. We carefully design the feature vector, outcome variable, and group membership index

TABLE I  
MAIN NOTATIONS

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

of the poisoned sample to explore the impact of such attack and derive the robust fairness-aware model. An important observation is that the formula of the objective function in the proposed minimax problem differs depending on the group index of the injected data point. As a result, when multiple adversarial data points can be added, all possibilities of their group index need to be considered and analyzed separately, providing crucial insight into designing practical attacks and robust fairness-aware models. Additionally, in the fairness problem, determining the group index is crucial as we aim to guarantee that the model performs equally well among different groups. Therefore, the added attack scenario complements the previous attack scheme presented in [1] and makes our work more comprehensive. Furthermore, we conduct more comprehensive numerical simulations and provide detailed theoretical analysis and proofs.

The main notations used in the paper are listed in Table I.

## II. RELATED WORK

**Adversarial attacks on FML.** There are many research works exploring the design of adversarial examples to reduce the testing accuracy and fairness of FML models. For example, [11] develops a gradient-based poisoning attack, [12] presents anchoring attack and influence attack, [13] provides three online attacks based on different group-based fairness measures, and [14] 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 [23], [24], [25], [26]. 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 [27].

**Intersection of fairness and robustness.** Fairness and robustness are critical elements of trustworthy AI that need to be addressed together [28]. 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 [28], [29], [30]. For example, [29] 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 [31]. Moreover, [32] and [33] 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.

### III. 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.

#### A. Problem Formulation

Using a set of training samples  $\{\mathbf{x}_i, y_i, G_i\}_{i=1}^n := \{\mathbf{X}, \mathbf{y}, \mathbf{G}\}$ , where  $\mathbf{x}_i \in \mathbb{R}^p$  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 paper, 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  $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2]$ ,  $\mathbf{y} = [\mathbf{y}_1; \mathbf{y}_2]$ .

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,  $\{\mathbf{x}_0, y_0, G_0\}$ , and add it into the existing dataset. After inserting this poisoned data point, we have the poisoned dataset  $\{\hat{\mathbf{X}}, \hat{\mathbf{y}}, \hat{\mathbf{G}}\}$ , where  $\hat{\mathbf{X}} = [\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_n]^T$ ,  $\hat{\mathbf{y}} = [y_0, y_1, \dots, y_n]^T$ ,  $\hat{\mathbf{G}} = [G_0, G_1, \dots, G_n]^T$ . 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

$$L = f(\beta, \hat{\mathbf{X}}, \hat{\mathbf{y}}, \hat{\mathbf{G}}) + \lambda F(\beta, \hat{\mathbf{X}}, \hat{\mathbf{y}}, \hat{\mathbf{G}}), \quad (1)$$

where  $\beta$  is the regression coefficient,  $f(\beta, \hat{\mathbf{X}}, \hat{\mathbf{y}}, \hat{\mathbf{G}})$  corresponds to the prediction accuracy loss,  $F(\beta, \hat{\mathbf{X}}, \hat{\mathbf{y}}, \hat{\mathbf{G}})$  corresponds to the group fairness gap and  $\lambda$  is the trade-off parameter. The goal of the adversary is to maximize (1) to make the model less fair and less accurate, while the robust fairness-aware regression model aims at minimizing (1). To make the problem meaningful, we introduce an energy constraint on the adversarial data point and use  $\ell_2$  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

$$\min_{\beta} \max_{\substack{(\mathbf{x}_0, y_0, G_0), \\ \text{s.t. } \|\mathbf{x}_0^T, y_0\|_2 \leq \eta}} L = f(\beta, \hat{\mathbf{X}}, \hat{\mathbf{y}}, \hat{\mathbf{G}}) + \lambda F(\beta, \hat{\mathbf{X}}, \hat{\mathbf{y}}, \hat{\mathbf{G}}). \quad (2)$$

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  $f(\beta, \hat{\mathbf{X}}, \hat{\mathbf{y}}, \hat{\mathbf{G}}) = \mathbb{E}[(Y - \hat{Y})^2]$ , where  $\hat{Y}$  is the prediction result. For the group fairness gap, various metrics have been proposed, including demographic parity [34], equality of opportunity [35], equalized odds [35], and accuracy parity [7]. 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 [7],  $\mathbb{E}[(Y - \hat{Y})^2 | G = 1] = \mathbb{E}[(Y - \hat{Y})^2 | G = 2]$ . Then the absolute difference between two groups can be used to measure the severity of violations [36] and we have  $F(\beta, \hat{\mathbf{X}}, \hat{\mathbf{y}}, \hat{\mathbf{G}}) = |\mathbb{E}[(Y - \hat{Y})^2 | G = 1] - \mathbb{E}[(Y - \hat{Y})^2 | G = 2]|$ .

#### B. Proposed Method

To solve the minimax problem in (2), we will first solve the inner maximization problem with respect to the adversary to design the optimal adversarial data point  $\{\mathbf{x}_0, y_0, G_0\}$  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

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

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

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

It is worth noting that for either  $L_1$  or  $L_2$ , the objective function of the minimax problem (2) is non-smooth nonconvex-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 1:* For any given  $\beta$ , we have

$$L \stackrel{(a)}{=} \max_{\substack{(\mathbf{x}_0, y_0, G_0), \\ \text{s.t. } \|\mathbf{x}_0^T, y_0\|_2 \leq \eta}} \{g_1(\beta), h_1(\beta), g_2(\beta), h_2(\beta)\},$$

where

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

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  $\tilde{\mathbf{x}}_0 = [\mathbf{x}_0^T, y_0]^T$ ,  $\mathbf{b} = [\boldsymbol{\beta}^T, -1]^T$ . 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  $\tilde{\mathbf{x}}_0^*(\boldsymbol{\beta}) = \eta \frac{\mathbf{b}}{\|\mathbf{b}\|_2}$  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  $\tilde{\mathbf{x}}_0^*(\boldsymbol{\beta}) \perp \mathbf{b}$  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  $\tilde{\mathbf{x}}_0^*(\boldsymbol{\beta}) = \eta \frac{\mathbf{b}}{\|\mathbf{b}\|_2}$  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  $\tilde{\mathbf{x}}_0^*(\boldsymbol{\beta}) \perp \mathbf{b}$  and  $G_0 = 2$ .

*Proof:* Please refer to Appendix A.  $\square$

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

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

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 1, the original min-max problem is converted to the following problem

$$\min_{\boldsymbol{\beta}} \max_{(\mathbf{x}_0, y_0, G_0)} L = \min_{\boldsymbol{\beta}} \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta}), g_2(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}. \quad (3)$$

As we seek to minimize the largest of four functions, (3) can be separated into four sub-problems. One of them is

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & g_1(\boldsymbol{\beta}), \\ \text{s.t.} \quad & g_1(\boldsymbol{\beta}) \geq g_2(\boldsymbol{\beta}), g_1(\boldsymbol{\beta}) \geq h_1(\boldsymbol{\beta}), g_1(\boldsymbol{\beta}) \geq h_2(\boldsymbol{\beta}), \end{aligned} \quad (4)$$

and other sub-problems can be written in a similar manner. Once these sub-problems are solved, the solution to (3) can be obtained.

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

In the following, we focus on solving (4). The analysis of other sub-problems can be done similarly. Specifically, (4) can be further written as

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & g_1(\boldsymbol{\beta}) = C_{g_1} \eta^2 (1 + \|\boldsymbol{\beta}\|_2^2) + C_{g_1} \|\mathbf{y}_1 - \mathbf{X}_1 \boldsymbol{\beta}\|_2^2 \\ & \quad + D_{g_1} \|\mathbf{y}_2 - \mathbf{X}_2 \boldsymbol{\beta}\|_2^2, \\ \text{s.t.} \quad & C_1(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - h_1(\boldsymbol{\beta}) \geq 0, \\ & C_2(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - g_2(\boldsymbol{\beta}) \geq 0, \\ & C_3(\boldsymbol{\beta}) = g_1(\boldsymbol{\beta}) - h_2(\boldsymbol{\beta}) \geq 0. \end{aligned} \quad (5)$$

For the objective function in (5), since  $D_{g_1}$  can be negative,  $\mathbf{M}_{g_1}$  is not necessarily positive-semidefinite. Hence, (5) is a non-convex quadratic minimization problem with several quadratic constraints (QCQP), which is NP hard in general [20]. 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 [37].

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

$$\begin{aligned} \mathbf{M}_{g_1} - \sum_{i=1}^3 \alpha_i \frac{\partial^2 C_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \succeq 0, \\ \frac{\partial g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}^*} - \sum_{i=1}^3 \alpha_i \frac{\partial C_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}^*} = 0, \\ \sum_{i=1}^3 \alpha_i C_i(\boldsymbol{\beta}^*) = 0, \\ C_i(\boldsymbol{\beta}^*) \geq 0, i = 1, 2, 3, \end{aligned} \quad (6)$$

then  $\boldsymbol{\beta}^*$  is a global minimizer of QCQP (5).

*Remark 2:* From (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  $\alpha_i$ s. By examining these 8 different cases, we can obtain the optimal regression coefficient  $\boldsymbol{\beta}^*$  of the sub-problem (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  $\alpha_i$ , i.e.  $\exists! \alpha_i > 0$  and  $\alpha_k = 0, \forall k \neq i$ ; 3) the case with two non-zero  $\alpha_i$ s, 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 1, if there exists  $\tilde{\boldsymbol{\beta}}$ , such that

$$\mathbf{M}_{g_1} \succeq 0, \quad (7)$$

$$\frac{\partial g_1(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\tilde{\boldsymbol{\beta}}} = 0, \quad (8)$$

$$C_i(\tilde{\beta}) \geq 0, i = 1, 2, 3, \quad (9)$$

then  $\tilde{\beta}$  is a global minimizer of (5). From (7), we require that  $M_{g_1}$  is positive-semidefinite, which can be true when  $\lambda$  is small, e.g. when  $D_{g_1} \geq 0$ . From (8), when  $M_{g_1}$  is invertible, we have

$$\tilde{\beta} = M_{g_1}^{-1} \left[ C_{g_1} X_1^T Y_1 + D_{g_1} X_2^T Y_2 \right]. \quad (10)$$

If (9) is satisfied at (10), then  $\tilde{\beta}$  is a global minimizer of (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 1, if there exists  $\tilde{\beta}$  and  $\alpha_1 > 0$ , such that

$$M_{g_1} - \alpha_1(M_{g_1} - M_{h_1}) \succeq 0, \quad (11)$$

$$\frac{\partial g_1(\beta)}{\partial \beta} \Big|_{\tilde{\beta}} - \alpha_1 \frac{\partial C_1(\beta)}{\partial \beta} \Big|_{\tilde{\beta}} = 0, \quad (12)$$

$$C_1(\tilde{\beta}) = 0, C_2(\tilde{\beta}) \geq 0, C_3(\tilde{\beta}) \geq 0, \quad (13)$$

then  $\tilde{\beta}$  is a global minimizer of (5).

*Proposition 2:* Denote the largest eigenvalue of  $X_1^T X_1$  as  $v_{X_1,p}$  and the largest eigenvalue of  $X_2^T X_2$  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  $A_{g_1 h_1} = \{\alpha : M_{g_1} - \alpha(M_{g_1} - M_{h_1}) \succ 0\} \neq \emptyset$ . By randomly selecting an  $\alpha_1^* \in A_{g_1 h_1}$ , for

$$\tilde{\beta} = [(1 - \alpha_1^* - \gamma^*)M_{g_1} + (\alpha_1^* + \gamma^*)M_{h_1}]^{-1} \cdot [(1 - \alpha_1^* - \gamma^*)E_{g_1} - (\alpha_1^* + \gamma^*)E_{h_1}],$$

where  $\gamma^*$  is a certain Lagrangian multiplier, and  $E_{g_1} = C_{g_1} X_1^T y_1 + D_{g_1} X_2^T y_2$ ,  $E_{h_1} = C_{h_1} X_1^T y_1 + D_{h_1} X_2^T y_2$ , if we have  $C_2(\tilde{\beta}) \geq 0, C_3(\tilde{\beta}) \geq 0$ , then  $\tilde{\beta}$  satisfies (11), (12), (13) and is a global minimizer of (5).

*Proof:* Please refer to Appendix B.  $\square$

*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 1, if there exists  $\hat{\beta}$  and  $\alpha_1 > 0, \alpha_2 > 0$ , such that

$$M_{g_1} - \alpha_1(M_{g_1} - M_{h_1}) - \alpha_2(M_{g_1} - M_{g_2}) \succeq 0, \quad (14)$$

$$\frac{\partial g_1(\beta)}{\partial \beta} \Big|_{\hat{\beta}} - \alpha_1 \frac{\partial C_1(\beta)}{\partial \beta} \Big|_{\hat{\beta}} - \alpha_2 \frac{\partial C_2(\beta)}{\partial \beta} \Big|_{\hat{\beta}} = 0, \quad (15)$$

$$C_1(\hat{\beta}) = 0, C_2(\hat{\beta}) = 0, \quad (16)$$

$$C_3(\hat{\beta}) \geq 0, \quad (17)$$

then  $\hat{\beta}$  is a global minimizer of (5).

*Proposition 3:* For

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

where  $\gamma_1^*, \gamma_2^*$  are certain Lagrangian multipliers, and  $E_{g_2} = C_{g_2} X_1^T y_1 + D_{g_2} X_2^T y_2$ , if  $C_3(\check{\beta}) \geq 0$ , then  $\check{\beta}$  satisfies (14), (15), (16), (17) and is a global minimizer of (5).

*Proof:* Please refer to Appendix C.  $\square$

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

---

### Algorithm 1 Attack with one adversarial point

---

**Input:**  $\{X, y, G\}, \lambda, \eta$ .

**Output:** Optimal adversarial point  $(x_0^*, y_0^*, G_0^*)$ . Robust fairness-aware regression coefficient  $\beta_{rob}^*$ .

**Procedure:**

- 1: Separate (2) into 4 sub-problems.
  - 2: **for** each sub-problem **do**
  - 3:   step through *Case 1* to *Case 4* until a  $\beta$  that satisfies the sufficient conditions in Proposition 1 is found.
  - 4: Select  $\beta$  that minimizes  $\max\{g_1(\beta), h_1(\beta), g_2(\beta), h_2(\beta)\}$  from the solution set as  $\beta_{rob}^*$ .
  - 5: Plug  $\beta_{rob}^*$  into the optimal attack strategy to obtain  $(x_0^*, y_0^*, G_0^*)$ .
- 

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

$$M_{g_1} - \alpha_1(M_{g_1} - M_{h_1}) - \alpha_2(M_{g_1} - M_{g_2}) - \alpha_3(M_{g_1} - M_{h_2}) \succeq 0, \quad (18)$$

$$\frac{\partial g_1(\beta)}{\partial \beta} \Big|_{\hat{\beta}} - \alpha_1 \frac{\partial C_1(\beta)}{\partial \beta} \Big|_{\hat{\beta}} - \alpha_2 \frac{\partial C_2(\beta)}{\partial \beta} \Big|_{\hat{\beta}} - \alpha_3 \frac{\partial C_3(\beta)}{\partial \beta} \Big|_{\hat{\beta}} = 0, \quad (19)$$

$$C_1(\hat{\beta}) = 0, C_2(\hat{\beta}) = 0, C_3(\hat{\beta}) = 0, \quad (20)$$

then  $\hat{\beta}$  is a global minimizer of (5). From Remark 1, we note that with (20), there are three equations on  $\|\beta\|_2^2, \|\mathbf{y}_1 - X_1\beta\|_2^2$  and  $\|\mathbf{y}_2 - X_2\beta\|_2^2$ , which indicates that there will be deterministic solutions for them or the feasible set is empty.

When the feasible set of (20) is nonempty (for example, when  $\lambda > \max\{\frac{m+1}{n+1}, \frac{n-m+1}{n+1}\}$ ), the value of  $g_1(\beta), C_1(\beta), C_2(\beta), C_3(\beta)$  is determined as there have been deterministic solutions for  $\|\beta\|_2^2, \|\mathbf{y}_1 - X_1\beta\|_2^2$  and  $\|\mathbf{y}_2 - X_2\beta\|_2^2$ . Then the process of finding  $\hat{\beta}$  is

1. Solve (20) and derive the solution for  $\|\beta\|_2^2, \|\mathbf{y}_1 - X_1\beta\|_2^2$  and  $\|\mathbf{y}_2 - X_2\beta\|_2^2$ .
2. Calculate the value of  $g_1(\beta), C_1(\beta), C_2(\beta), C_3(\beta)$ .
3. Select  $\alpha_1, \alpha_2, \alpha_3$  such that (18) is satisfied. Then (19) is satisfied naturally as  $g_1(\beta), C_1(\beta), C_2(\beta), C_3(\beta)$  are constants.

Algorithm 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 [20], our proposed QCQP (5) is different from the setting considered in [20]. Notably, the objective function in (5) is nonconvex, and the constraints are not concave, which distinguishes our problem from the one in [20]. Moreover, while [20] provides approximation bounds for a norm minimization problem with multiple concave quadratic constraints, our objective is to find a global minimizer of (5) based on the sufficient conditions of global minimizers to QCQP, as stated in Proposition 1. Consequently, the methods proposed in [20] are not applicable to our problem.

## IV. RANK-ONE ATTACK

In Section III, 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 [26]. In particular, the attacker will carefully design a rank-one feature modification matrix  $\Delta$  and add it to the original feature matrix  $X$ , so as to obtain the modified feature matrix  $\hat{X} = X + \Delta$ . Since  $\Delta$  is of rank 1, we can write  $\Delta = cd^T$ , where  $c \in \mathbb{R}^n$  and  $d \in \mathbb{R}^p$ . Moreover, recall that there are samples from two groups, we denote the modification matrix of the first group as  $\Delta_1$ , i.e., the first  $m$  rows of  $\Delta$ , and assume that  $\Delta_1 = c_1 d^T$ , where  $c_1$  consists of the first  $m$  components of  $c$ . Similarly, for the second group, the modification matrix is  $\Delta_2 = c_2 d^T$ . Then the modified feature matrices are  $\hat{X}_1 = X_1 + \Delta_1$  and  $\hat{X}_2 = X_2 + \Delta_2$ .

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

$$\min_{\beta} \max_{\|\Delta\|_F \leq \eta} f(\beta, \hat{X}) + \lambda F(\beta, \hat{X}). \quad (21)$$

To solve (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 (21), we have

$$\begin{aligned} f(\beta, \hat{X}) + \lambda F(\beta, \hat{X}) &= \frac{1}{n} \|y - \hat{X}\beta\|_2^2 \\ &+ \lambda \left| \frac{1}{m} \|y_1 - \hat{X}_1\beta\|_2^2 - \frac{1}{n-m} \|y_2 - \hat{X}_2\beta\|_2^2 \right| \\ &= \max\{g(\beta, \hat{X}), h(\beta, \hat{X})\}, \end{aligned}$$

in which  $g(\beta, \hat{X}) = C_g \|y_1 - \hat{X}_1\beta\|_2^2 + D_g \|y_2 - \hat{X}_2\beta\|_2^2$ ,  $h(\beta, \hat{X}) = C_h \|y_1 - \hat{X}_1\beta\|_2^2 + D_h \|y_2 - \hat{X}_2\beta\|_2^2$ , 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 1:* For  $g(\beta, \hat{X})$  and  $h(\beta, \hat{X})$ , we have that

- 1) if  $D_g \geq 0$ ,  $g(\beta, \hat{X})$  is convex in  $c_1$  for any given  $c_2, d$ , and also convex in  $c_2$  for any given  $c_1, d$ ; otherwise,  $g(\beta, \hat{X})$  is convex in  $c_1$  for any given  $c_2, d$ , and concave in  $c_2$  for any given  $c_1, d$ ;
- 2) if  $C_h \geq 0$ ,  $h(\beta, \hat{X})$  is convex in  $c_1$  for any given  $c_2, d$ , and also convex in  $c_2$  for any given  $c_1, d$ ; otherwise,  $h(\beta, \hat{X})$  is concave in  $c_1$  for any given  $c_2, d$ , and convex in  $c_2$  for any given  $c_1, d$ .

Based on Lemma 1, we now solve the maximization problem in (21). First, note that

$$\begin{aligned} &\max_{\|cd^T\|_F \leq \eta} \max\{g(\beta, \hat{X}), h(\beta, \hat{X})\} \\ &= \max \left\{ \max_{\|cd^T\|_F \leq \eta} g(\beta, \hat{X}), \max_{\|cd^T\|_F \leq \eta} h(\beta, \hat{X}) \right\}, \end{aligned}$$

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  $g(\beta, \hat{X})$  in detail and the sub-problem of  $h(\beta, \hat{X})$  can be analyzed similarly.

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

According to Lemma 1, the value of  $D_g$  will affect the property of  $g(\beta, \hat{X})$ . In the following, we will first explore the case  $D_g \geq 0$  and obtain Lemma 2 as well as Proposition 4, and then explore the case  $D_g < 0$  and obtain Lemma 3 as well as Proposition 5.

*Lemma 2:* For  $D_g \geq 0$ , we have

$$\begin{aligned} &\max_{\|cd^T\|_F \leq \eta} g(\beta, \hat{X}) \\ &= \max_{0 < \eta_c \leq \eta} \max_{0 \leq \eta_{c_1} \leq \eta_c} \max_{\|d\|_2 \leq 1} \max_{\|c_2\|_2 = \sqrt{\eta_c^2 - \eta_{c_1}^2}} \max_{\|c_1\|_2 = \eta_{c_1}} g(\beta, \hat{X}) \\ &= \max_{0 < \eta_c \leq \eta} \max_{0 \leq \eta_{c_1} \leq \eta_c} \max_{\|d\|_2 \leq 1} g_{m_1}(\eta_{c_1}, \beta, d), \end{aligned}$$

where  $g_{m_1}(\eta_{c_1}, \beta, d) = C_g (\|y_1 - X_1\beta\|_2 + \eta_{c_1} d^T \beta)^2 + D_g (\|y_2 - X_2\beta\|_2 + \sqrt{\eta_c^2 - \eta_{c_1}^2} d^T \beta)^2$ .

*Proof:* Please refer to Appendix D.  $\square$

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

*Proposition 4:*

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

where  $g_a(\eta_{c_1}, \beta) = C_g (\|y_1 - X_1\beta\|_2 + \eta_{c_1} \|\beta\|_2)^2 + D_g (\|y_2 - X_2\beta\|_2 + \sqrt{\eta^2 - \eta_{c_1}^2} \|\beta\|_2)^2$ .

*Proof:* Please refer to Appendix E.  $\square$

*Lemma 3:* For  $D_g < 0$ , we have

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

where

$$g_{m_2}(\eta_{c_1}, \beta, d) = \begin{cases} C_g (\|y_1 - X_1\beta\|_2 + \eta_{c_1} d^T \beta)^2, & \text{if } \|y_2 - X_2\beta\|_2 \leq \eta \|\beta\|_2, \\ C_g (\|y_1 - X_1\beta\|_2 + \eta_{c_1} d^T \beta)^2 \\ + D_g (\|y_2 - X_2\beta\|_2 - \sqrt{\eta_c^2 - \eta_{c_1}^2} d^T \beta)^2, & \text{otherwise,} \end{cases}$$

*Proof:* Please refer to Appendix F.  $\square$

From the above lemma, we have the following proposition.

*Proposition 5:*

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

where

$$g_b(\eta_{c_1}, \beta) = \begin{cases} g_{b_1}(\eta_{c_1}, \beta), & \text{if } \|y_2 - X_2\beta\|_2 \leq \eta \|\beta\|_2, \\ g_{b_2}(\eta_{c_1}, \beta), & \text{otherwise.} \end{cases}$$

$$g_{b_1}(\eta_{c_1}, \beta) = C_g (\|y_1 - X_1\beta\|_2 + \eta_{c_1} \|\beta\|_2)^2,$$

$$g_{b_2}(\eta_{c_1}, \beta) = [C_g (\|y_1 - X_1\beta\|_2 + \eta_{c_1} \|\beta\|_2)^2 + D_g (\|y_2 - X_2\beta\|_2 - \sqrt{\eta^2 - \eta_{c_1}^2} \|\beta\|_2)^2].$$

*Proof:* Please refer to Appendix G.  $\square$

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

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

- if  $C_h \geq 0$ , we have

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

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

- if  $C_h < 0$ , we have

$$\max_{\|\mathbf{c}d^T\|_F \leq \eta} h(\beta, \hat{\mathbf{X}}) = \max_{0 \leq \eta_{c_1} \leq \eta} h_b(\eta_{c_1}, \beta),$$

where

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

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

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

*Transformation of the Minimax Problem:* After solving sub-problems above, the minimax problem (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  $C_h < 0$ , (21) can be represented as

$$\min_{\beta} \max_{0 \leq \eta_{c_1} \leq \eta} \max\{g_a(\eta_{c_1}, \beta), h_b(\eta_{c_1}, \beta)\}. \quad (22)$$

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

*Lemma 4:* If the norm of  $\beta$  is bounded, i.e.  $\|\beta\|_2 \leq B_\beta$ , then we have

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

*Proof:* Please refer to Appendix H.  $\square$

*Lemma 5:* For any given  $\beta$ ,  $g_a$ ,  $g_{b_2}$ ,  $h_a$  and  $h_{b_2}$  are all unimodal functions with respect to  $\eta_{c_1}$  that increase first and then decrease.

*Proof:* Please refer to Appendix I.  $\square$

Moreover, to deal with the piece-wise structure in the objective function, we further transform the minimax problem to several sub-problems. For example, (22) can be transformed to three sub-problems:

- 1)  $\min_{\beta} \max_{0 \leq \eta_{c_1} \leq \eta} h_{b_1}(\eta_{c_1}, \beta)$ , s.t.  $g_a(\eta_{c_1}, \beta) < h_{b_1}(\eta_{c_1}, \beta)$ ,  $\|\mathbf{y}_1 - \mathbf{X}_1\beta\|_2 \leq \eta\|\beta\|_2$ ;
- 2)  $\min_{\beta} \max_{0 \leq \eta_{c_1} \leq \eta} h_{b_2}(\eta_{c_1}, \beta)$ , s.t.  $g_a(\eta_{c_1}, \beta) < h_{b_2}(\eta_{c_1}, \beta)$ ,  $\|\mathbf{y}_1 - \mathbf{X}_1\beta\|_2 > \eta\|\beta\|_2$ .
- 3)  $\min_{\beta} \max_{0 \leq \eta_{c_1} \leq \eta} g_a(\eta_{c_1}, \beta)$ , s.t.  $g_a(\eta_{c_1}, \beta) \geq h_b(\eta_{c_1}, \beta)$ .

For the sub-problem 1), the maximization on  $\eta_{c_1}$  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  $\|\beta\|_2 \leq B_\beta$ , which is reasonable in reality, and have that: the feasible set  $\{\beta : \|\beta\|_2 \leq B_\beta\} \times [0, \eta]$  is convex and compact; the objective function is weakly-convex-weakly-concave by Lemma 4; the saddle-point exists by Lemma 5. Based on those properties,

we are able to apply a first-order algorithms proposed by [22] to solve the non-convex non-concave minimax problem as in sub-problem 2) and derive the nearly  $\epsilon$ -stationary solution. In particular, define  $\mathcal{Z} = \{\beta : \|\beta\|_2 \leq B_\beta\} \times [0, \eta]$  and the mapping  $H(z) := (\partial_\beta h_{b_2}(\eta_{c_1}, \beta), \partial_{\eta_{c_1}}[-h_{b_2}(\eta_{c_1}, \beta)])^T$ , where  $z = (\beta, \eta_{c_1})$ . The minty variational inequality (MVI) problem corresponding to the saddle-point problem in sub-problem 2) is to find  $z^* \in \mathcal{Z}$  such that  $\langle \xi, z - z^* \rangle \geq 0, \forall z \in \mathcal{Z}, \forall \xi \in H(z)$ . Then the saddle-point problem can be solved through the lens of MVI. In [22], the proposed inexact proximal point method consists of approximately solving a sequence of strongly monotone MVIs constructed by adding a strongly monotone mapping to  $H(z)$  with a sequentially updated proximal center. Thus, the complex non-convex non-concave minimax problem can be decomposed into a sequence of easier strongly-convex strongly-concave problems.

In comparison to [22], 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 in our paper differs from that in [22]. In our transformed minimax problem (22), the objective function is a max-type function, which is not an exact weakly-convex-weakly-concave minimax problem.

## V. NUMERICAL RESULTS

In this section, we provide numerical examples to illustrate the results in this paper. 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

$$\mathbf{y}_1 = \mathbf{X}_1\beta_{0,1} + \mathbf{c}_1 + \epsilon, \quad \mathbf{y}_2 = \mathbf{X}_2\beta_{0,2} + \epsilon, \quad (23)$$

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

**2. Law School Dataset (LSD) [38]:** 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  $\eta$ , we have  $\eta \geq \eta_{min} = 2.44$  and  $\eta_D = 2.86$ .

**3. Medical Insurance Cost Dataset (MICD) [39]:** 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

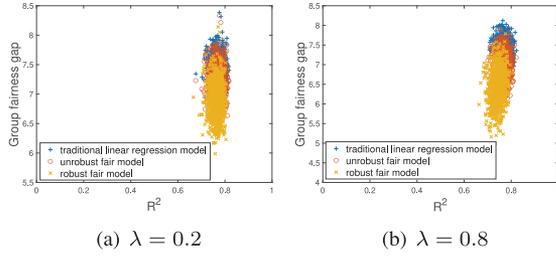


Fig. 1. SD: comparison of robust fair model, unrobust fair model and traditional linear model (attack with one adversarial data point).

two genders respectively. Then we verify the assumption on  $\eta$  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  $\{\mathbf{X}, \mathbf{y}, \mathbf{G}\}$ . In particular, the unrobust fair model is

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

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 [40]. The optimal regression coefficient for each model is derived by fitting the model on the original dataset  $\{\mathbf{X}, \mathbf{y}, \mathbf{G}\}$ . To obtain the performance of each model on the poisoned dataset, we apply the derived optimal regression coefficient on the poisoned dataset,  $\{\tilde{\mathbf{X}}, \tilde{\mathbf{y}}, \tilde{\mathbf{G}}\}$ , and calculate the MSE as well as the group fairness gap.

#### A. 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. 1(a) and Fig. 1(b), following (23), we construct 500 datasets relying on the randomness in  $\epsilon$ . 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 1, the best adversarial point is  $\tilde{\mathbf{x}}_0 = \eta \frac{\mathbf{b}}{\|\mathbf{b}\|_2}$ . As shown in Fig. 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  $\beta_{fair}$  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 1, the best adversarial point will be in the form  $\tilde{\mathbf{x}}_0 \perp \mathbf{b}$  or  $\tilde{\mathbf{x}}_0 = \eta \frac{\mathbf{b}}{\|\mathbf{b}\|_2}$  based on the value of  $g_i(\beta_{rob}^*)$  and  $h_i(\beta_{rob}^*), i = 1, 2$ . As shown in Fig. 1(b), the performance results are similar to the case  $\lambda = 0.2$ .

Secondly, we explore the effects of the energy constraint parameter  $\eta$  as well as the trade-off parameter  $\lambda$  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. 2, when  $\eta$  is small, under different choices of  $\lambda$ , MSE and the group fairness gap for the robust fairness-aware model are both

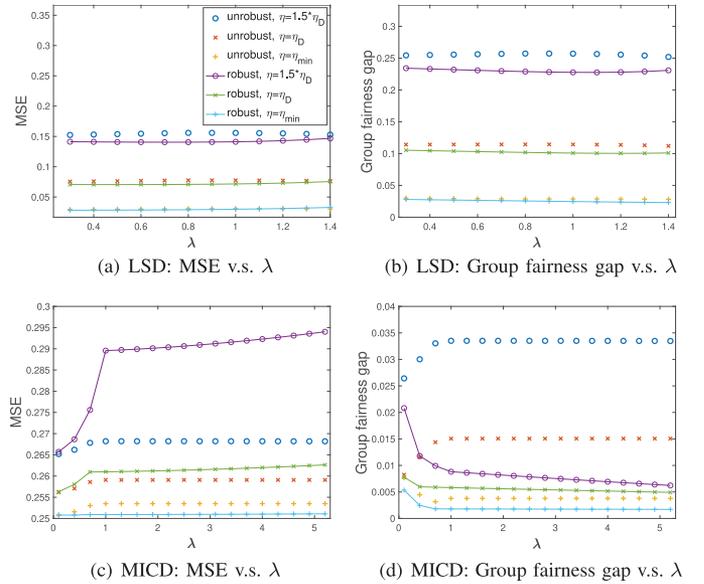


Fig. 2. Effects of  $\lambda$  and  $\eta$  on MSE and the group fairness gap (attack with one adversarial data point, samples with energy greater than  $5\eta_D$  are removed).

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.

#### B. Rank-One Attack

In the first experiment, we explore the effects of the energy constraint parameter  $\eta$  as well as the trade-off parameter  $\lambda$ . We carry out the attack with three different energy levels,  $\eta = 0.2\sigma$ ,  $\eta = 0.5\sigma$  and  $\eta = 0.8\sigma$ , where  $\sigma$  is the smallest singular value of the feature matrix of the training data. As shown in Fig. 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  $\lambda$ . In particular, as  $\lambda$  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. 3(c), when the energy constraint is comparable to the smallest singular value of the feature matrix ( $\eta = 0.8\sigma$ ) and the trade-off parameter  $\lambda$  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, 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

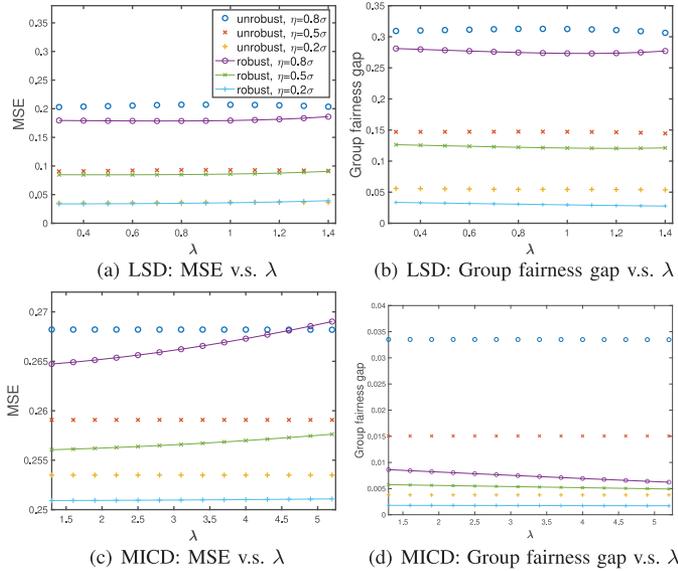


Fig. 3. Effects of  $\lambda$  and  $\eta$  on MSE and fairness gap (rank-one attack, samples with energy greater than  $10\sigma$  are removed).

fair model and adversarially robust fair model, since the choice of the trade-off parameter  $\lambda$  will affect the model performance, we explore models with various choices of  $\lambda$  from the range  $[0.5, 1.2]$ . As shown in Fig. 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 performance with the proposed unrobust fair regression model (with certain choice of  $\lambda$ ). Moreover, for the unrobust fair model, it is observed that as  $\lambda$  increases, the group fairness gap decreases while the MSE increases. However, on the poisoned dataset, as shown in Fig. 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  $4.3 \times 10^{-3}$  to  $2.8 \times 10^{-2}$ , 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  $\lambda$ , a model that performs well in terms of both fairness and prediction accuracy can be obtained.

## VI. CONCLUSION

In this paper, 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 prediction 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

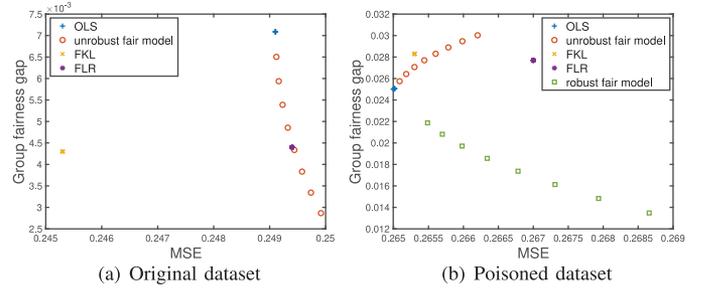


Fig. 4. MICD: Group fairness gap v.s. MSE (rank-one attack).

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  $\lambda$ .

## APPENDIX A PROOF OF THEOREM 1

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  $\max_{(\mathbf{x}_0, y_0, 1), \text{ s.t. } \|\mathbf{x}_0^T, y_0\|_2 \leq \eta} L_1 \stackrel{(b)}{=} \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta})\}$ . Similarly, for  $G_0 = 2$ , we will have that  $\max_{(\mathbf{x}_0, y_0, 2), \text{ s.t. } \|\mathbf{x}_0^T, y_0\|_2 \leq \eta} L_2 \stackrel{(c)}{=} \max\{g_2(\boldsymbol{\beta}), h_2(\boldsymbol{\beta})\}$ . 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  $\|\tilde{\mathbf{x}}_0^T\|_2 = \|\mathbf{x}_0^T, y_0\|_2 \leq \eta$ , we have

$$0 \leq \|\tilde{\mathbf{x}}_0^T \mathbf{b}\|_2^2 \leq \eta^2 \|\mathbf{b}\|_2^2 = \eta^2 (1 + \|\boldsymbol{\beta}\|_2^2). \quad (24)$$

Then we notice that

$$\begin{aligned} L_1 &\stackrel{(d)}{\leq} \max \left\{ \left( \frac{\lambda}{m+1} + \frac{1}{n+1} \right) \eta^2 (1 + \|\boldsymbol{\beta}\|_2^2) \right. \\ &\quad + \left( \frac{\lambda}{m+1} + \frac{1}{n+1} \right) \|\mathbf{y}_1 - \mathbf{X}_1 \boldsymbol{\beta}\|_2^2 \\ &\quad + \left( -\frac{\lambda}{n-m} + \frac{1}{n+1} \right) \|\mathbf{y}_2 - \mathbf{X}_2 \boldsymbol{\beta}\|_2^2, \\ &\quad \max \left\{ 0, -\frac{\lambda}{m+1} + \frac{1}{n+1} \right\} \eta^2 (1 + \|\boldsymbol{\beta}\|_2^2) \\ &\quad + \left( -\frac{\lambda}{m+1} + \frac{1}{n+1} \right) \|\mathbf{y}_1 - \mathbf{X}_1 \boldsymbol{\beta}\|_2^2 \\ &\quad \left. + \left( \frac{\lambda}{n-m} + \frac{1}{n+1} \right) \|\mathbf{y}_2 - \mathbf{X}_2 \boldsymbol{\beta}\|_2^2 \right\} \\ &= \max\{g_1(\boldsymbol{\beta}), h_1(\boldsymbol{\beta})\}, \end{aligned}$$

where (d) is from (24). Then we verify the achievability of the equality in (d). Define a set  $B_1 := \{\boldsymbol{\beta} : g_1(\boldsymbol{\beta}) \geq h_1(\boldsymbol{\beta}) = \{\boldsymbol{\beta} : \frac{1}{m+1} \|\mathbf{y}_1 - \mathbf{X}_1 \boldsymbol{\beta}\|_2^2 - \frac{1}{n-m} \|\mathbf{y}_2 - \mathbf{X}_2 \boldsymbol{\beta}\|_2^2 \geq \max \left\{ -\frac{1}{2(m+1)} - \frac{1}{2\lambda(n+1)}, -\frac{1}{m+1} \right\} \cdot \eta^2 (1 + \|\boldsymbol{\beta}\|_2^2)\}$ . In the

sequel, we will verify the achievability of the equality in (d) with two cases:  $\beta \in B_1$  and  $\beta \in B_1^c$ .

Case 1:  $\beta \in B_1$ : By taking  $\tilde{\mathbf{x}}_0 = \eta \frac{\mathbf{b}}{\|\mathbf{b}\|_2}$ , we have

$$\begin{aligned} & \frac{1}{m+1} (\|y_0 - \mathbf{x}_0^T \beta\|_2^2 + \|\mathbf{y}_1 - \mathbf{X}_1 \beta\|_2^2) \\ & - \frac{1}{n-m} \|\mathbf{y}_2 - \mathbf{X}_2 \beta\|_2^2 \\ = & \frac{1}{m+1} [\eta^2 (1 + \|\beta\|_2^2) + \|\mathbf{y}_1 - \mathbf{X}_1 \beta\|_2^2] \\ & - \frac{1}{n-m} \|\mathbf{y}_2 - \mathbf{X}_2 \beta\|_2^2 \\ \stackrel{(e)}{\geq} & \max \left\{ \frac{1}{2(m+1)} - \frac{1}{2\lambda(n+1)}, 0 \right\} \eta^2 (1 + \|\beta\|_2^2), \quad (25) \end{aligned}$$

where (e) is from the definition of set  $B_1$ . Then we have  $L_1 \stackrel{(f)}{=} g_1(\beta)$ , where (f) follows from (25). Therefore, for  $\beta \in B_1$ , we have  $h_1(\beta) \leq g_1(\beta)$  and  $L_1 \leq g_1(\beta)$ , in which the equality can be achieved for  $\tilde{\mathbf{x}}_0 = \eta \frac{\mathbf{b}}{\|\mathbf{b}\|_2}$ .

Case 2:  $\beta \in B_1^c$ : On the one hand, if  $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} \leq 0$ , by taking  $\tilde{\mathbf{x}}_0 = \eta \frac{\mathbf{b}}{\|\mathbf{b}\|_2}$ , we have

$$\begin{aligned} & \frac{1}{m+1} (\|y_0 - \mathbf{x}_0^T \beta\|_2^2 + \|\mathbf{y}_1 - \mathbf{X}_1 \beta\|_2^2) \\ & - \frac{1}{n-m} \|\mathbf{y}_2 - \mathbf{X}_2 \beta\|_2^2 \\ \stackrel{(g)}{<} & \max \left\{ \frac{1}{2(m+1)} - \frac{1}{2\lambda(n+1)}, 0 \right\} \eta^2 (1 + \|\beta\|_2^2) \\ \stackrel{(h)}{=} & 0, \quad (26) \end{aligned}$$

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

$$\begin{aligned} L_1 \stackrel{(j)}{=} & \frac{1}{n+1} [\eta^2 (1 + \|\beta\|_2^2) + \|\mathbf{y}_1 - \mathbf{X}_1 \beta\|_2^2 \\ & + \|\mathbf{y}_2 - \mathbf{X}_2 \beta\|_2^2] - \lambda \left[ \frac{1}{m+1} \eta^2 (1 + \|\beta\|_2^2) + \frac{1}{m+1} \|\mathbf{y}_1 \right. \\ & \left. - \mathbf{X}_1 \beta\|_2^2 - \frac{1}{n-m} \|\mathbf{y}_2 - \mathbf{X}_2 \beta\|_2^2 \right] \stackrel{(k)}{=} h_1(\beta), \end{aligned}$$

where (j) is from (26) and (k) is true because  $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} \leq 0$ . Therefore, for  $\beta \in B_1^c$  and  $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} \leq 0$ , we have  $h_1(\beta) \geq g_1(\beta)$  and  $L_1 \leq h_1(\beta)$ , in which the equality can be achieved for  $\tilde{\mathbf{x}}_0 = \eta \frac{\mathbf{b}}{\|\mathbf{b}\|_2}$ .

On the other hand, if  $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} > 0$ , by taking  $\tilde{\mathbf{x}}_0$  to be a vector such that  $\tilde{\mathbf{x}}_0 \perp \mathbf{b}$ , we have

$$\begin{aligned} & \frac{1}{m+1} (\|y_0 - \mathbf{x}_0^T \beta\|_2^2 + \|\mathbf{y}_1 - \mathbf{X}_1 \beta\|_2^2) \\ & - \frac{1}{n-m} \|\mathbf{y}_2 - \mathbf{X}_2 \beta\|_2^2 \\ = & \frac{1}{m+1} \|\mathbf{y}_1 - \mathbf{X}_1 \beta\|_2^2 - \frac{1}{n-m} \|\mathbf{y}_2 - \mathbf{X}_2 \beta\|_2^2 \\ \stackrel{(l)}{<} & \max \left\{ -\frac{1}{2(m+1)} - \frac{1}{2\lambda(n+1)}, -\frac{1}{m+1} \right\} \eta^2 (1 + \|\beta\|_2^2) \\ < & 0, \quad (27) \end{aligned}$$

where (l) is from the definition of set  $B_1$ . Then we have

$$\begin{aligned} L_1 \stackrel{(s)}{=} & \frac{1}{n+1} (\|\mathbf{y}_1 - \mathbf{X}_1 \beta\|_2^2 + \|\mathbf{y}_2 - \mathbf{X}_2 \beta\|_2^2) \\ & - \lambda \left[ \frac{1}{m+1} \|\mathbf{y}_1 - \mathbf{X}_1 \beta\|_2^2 - \frac{1}{n-m} \|\mathbf{y}_2 - \mathbf{X}_2 \beta\|_2^2 \right] \\ \stackrel{(t)}{=} & h_1(\beta), \end{aligned}$$

where (s) is from (27) and (t) is because  $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} > 0$ . Therefore, for  $\beta \in B_1^c$  and  $\frac{1}{m+1} - \frac{1}{\lambda(n+1)} > 0$ , we have  $h_1(\beta) \geq g_1(\beta)$  and  $L_1 \leq h_1(\beta)$ , in which the equality can be achieved when  $\tilde{\mathbf{x}}_0 \perp \mathbf{b}$ .

## APPENDIX B

### PROOF OF PROPOSITION 2

First, we summarize the process of finding  $\bar{\beta}$  as follows.

1. Check whether  $A = \{\alpha_1 : \mathbf{M}_{g_1} - \alpha_1(\mathbf{M}_{g_1} - \mathbf{M}_{h_1}) \succeq 0\} \neq \emptyset$ . If  $A = \emptyset$ , there does not exist a global minimizer in this case.
2. By randomly selecting an  $\alpha_1^* \in A_{g_1 h_1} := \{\alpha : \mathbf{M}_{g_1} - \alpha(\mathbf{M}_{g_1} - \mathbf{M}_{h_1}) \succ 0\}$ , we solve the optimization problem

$$\begin{aligned} \min_{\beta} \quad & k(\beta) = g_1(\beta) - \alpha_1^* [g_1(\beta) - h_1(\beta)], \\ \text{s.t.} \quad & C_1(\beta) = g_1(\beta) - h_1(\beta) = 0, \quad (28) \end{aligned}$$

where  $k(\beta)$  is positive-definite and the choice of  $\alpha_1^*$  does not affect the solution to the problem.

3. For the solution to (28), check whether  $\alpha_1 > 0$ , (11), (12) and (13) are satisfied.

Now we explore the details of steps 1, 2 and 3.

In step 1, the assumption  $\eta^2 \geq \eta_{\min}^2 = \max \left\{ \frac{(n+1)v_{X_1,p}}{m(m+1)}, \frac{(n+1)v_{X_2,p}}{(n-m+1)(n-m)} \right\}$  will guarantee that  $A$  is nonempty. To be exact, we denote  $A_{g_1 g_2} = \{\alpha : \mathbf{M}_{g_1} - \alpha(\mathbf{M}_{g_1} - \mathbf{M}_{g_2}) \succ 0\}$ ,  $A_{g_1 h_2} = \{\alpha : \mathbf{M}_{g_1} - \alpha(\mathbf{M}_{g_1} - \mathbf{M}_{h_2}) \succ 0\}$ ,  $A_{h_1 g_2} = \{\alpha : \mathbf{M}_{h_1} - \alpha(\mathbf{M}_{h_1} - \mathbf{M}_{g_2}) \succ 0\}$ ,  $A_{h_1 h_2} = \{\alpha : \mathbf{M}_{h_1} - \alpha(\mathbf{M}_{h_1} - \mathbf{M}_{h_2}) \succ 0\}$ ,  $A_{g_2 h_2} = \{\alpha : \mathbf{M}_{g_2} - \alpha(\mathbf{M}_{g_2} - \mathbf{M}_{h_2}) \succ 0\}$ . Then under the assumption 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 are able to derive that  $A_{g_1 h_1} \neq \emptyset$ ,  $A_{g_1 g_2} \neq \emptyset$ ,  $A_{g_1 h_2} \neq \emptyset$ ,  $A_{h_1 g_2} \neq \emptyset$ ,  $A_{h_1 h_2} \neq \emptyset$ ,  $A_{g_2 h_2} \neq \emptyset$ . The detailed proof is omitted here. Particularly, in this case study, we have  $A_{g_1 h_1} \subset A$  and  $A \neq \emptyset$ .

In step 2, (28) is a strictly convex quadratic optimization problem with one quadratic equality constraint, which has been discussed in [41]. Define the Lagrangian function of (28) as

$$\begin{aligned} \mathcal{L}(\beta, \gamma) &= k(\beta) - \gamma C_1(\beta) \\ &= g_1(\beta) - (\alpha_1^* + \gamma)(g_1(\beta) - h_1(\beta)) \\ &= (1 - \alpha_1^* - \gamma)g_1(\beta) + (\alpha_1^* + \gamma)h_1(\beta), \end{aligned}$$

where  $\gamma$  is the Lagrangian multiplier. According to [41], the global minimizer  $\bar{\beta}$  and the corresponding multiplier  $\gamma^*$  of (28) satisfy first-order, second-order and the

constraint conditions

$$\frac{\partial \mathcal{L}}{\partial \beta} \Big|_{\check{\beta}} = (1 - \alpha_1^* - \gamma) \frac{\partial g_1(\beta)}{\partial \beta} \Big|_{\check{\beta}} + (\alpha_1^* + \gamma) \frac{\partial h_1(\beta)}{\partial \beta} \Big|_{\check{\beta}} = 0, \quad (29)$$

$$\frac{\partial^2 \mathcal{L}}{\partial \beta^2} = 2[(1 - \alpha_1^* - \gamma) \mathbf{M}_{g_1} + (\alpha_1^* + \gamma^*) \mathbf{M}_{h_1}] \succeq \mathbf{0}, \quad (30)$$

$$C_1(\check{\beta}) = 0. \quad (30)$$

From (29), we have

$$\check{\beta} = [(1 - \alpha_1^* - \gamma^*) \mathbf{M}_{g_1} + (\alpha_1^* + \gamma^*) \mathbf{M}_{h_1}]^{-1} \cdot [(1 - \alpha_1^* - \gamma^*) \mathbf{E}_{g_1} - (\alpha_1^* + \gamma^*) \mathbf{E}_{h_1}]. \quad (31)$$

Substituting (31) into (30), we derive an equation for  $\gamma$ ,  $K(\gamma) = C_1(\check{\beta}) = 0$ , whose root is  $\gamma^*$ . By plugging  $\gamma = \gamma^*$  back into (31), the exact solution for  $\check{\beta}$  is obtained.

For step 3, if  $C_2(\check{\beta}) \geq 0, C_3(\check{\beta}) \geq 0$ , then we have that:

1. if  $\alpha_1^* + \gamma^* > 0$ ,  $\beta = \check{\beta}$  is a global minimizer satisfying (11), (12), (13) with  $\alpha_1 = \alpha_1^* + \gamma^*$ ;
2. if  $\alpha_1^* + \gamma^* = 0$ ,  $\beta = \check{\beta}$  is a global minimizer in *Case I* that satisfies (7), (8), (9);
3. if  $\alpha_1^* + \gamma^* < 0$ ,  $\beta = \check{\beta}$  satisfies global optimality conditions for the minimization of  $h_1(\beta)$  with multipliers  $\alpha'_1 = 1 - \alpha_1^* - \gamma^*, \alpha'_2 = \alpha'_3 = 0$ .

#### APPENDIX C PROOF OF PROPOSITION 3

First, we summarize the process of finding  $\hat{\beta}$  as follows.

1. Check  $AA = \{(\alpha_1, \alpha_2) : \mathbf{M}_{g_1} - \alpha_1(\mathbf{M}_{g_1} - \mathbf{M}_{h_1}) - \alpha_2(\mathbf{M}_{g_1} - \mathbf{M}_{g_2}) \succeq 0\} \neq \emptyset$ . Under the assumption made in Proposition 2 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  $A_{g_1 h_1} = \{\alpha : \mathbf{M}_{g_1} - \alpha(\mathbf{M}_{g_1} - \mathbf{M}_{h_1}) \succ 0\} \neq \emptyset$  and  $A_{g_1 g_2} = \{\alpha : \mathbf{M}_{g_1} - \alpha(\mathbf{M}_{g_1} - \mathbf{M}_{g_2}) \succ 0\} \neq \emptyset$ , which implies  $AA \neq \emptyset$ .

2. Solve the optimization problem

$$\begin{aligned} \min_{\beta} \quad & k(\beta) = g_1(\beta) - \alpha_1^*[g_1(\beta) - h_1(\beta)], \\ \text{s.t.} \quad & C_1(\beta) = C_2(\beta) = 0. \end{aligned} \quad (32)$$

3. For the solution to (32), check whether  $\alpha_1 > 0, \alpha_2 > 0$ , and (17) are satisfied.

We now provide more details of steps 2 and 3. In step 2, define the Lagrangian function of (32) as

$$\begin{aligned} \mathcal{L}(\beta, \gamma_i) &= k(\beta) - \gamma_1 C_1(\beta) - \gamma_2 C_2(\beta) \\ &= (1 - \alpha_1^* - \gamma_1 - \gamma_2) g_1(\beta) + (\alpha_1^* + \gamma_1) h_1(\beta) \\ &\quad + \gamma_2 h_1(\beta). \end{aligned}$$

Then the derived optimal solution  $\check{\beta}$  and the corresponding Lagrangian multipliers  $\gamma_1^*, \gamma_2^*$  satisfy first-order, second-order and the constraint conditions

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \beta} \Big|_{\check{\beta}} &= (1 - \alpha_1^* - \gamma_1^* - \gamma_2^*) \frac{\partial g_1(\beta)}{\partial \beta} \Big|_{\check{\beta}} \\ &\quad + (\alpha_1^* + \gamma_1^*) \frac{\partial h_1(\beta)}{\partial \beta} \Big|_{\check{\beta}} + \gamma_2^* \frac{\partial g_2(\beta)}{\partial \beta} \Big|_{\check{\beta}} = \mathbf{0}, \end{aligned} \quad (33)$$

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial \beta^2} &= 2[(1 - \alpha_1^* - \gamma_1^* - \gamma_2^*) \mathbf{M}_{g_1} \\ &\quad + (\alpha_1^* + \gamma_1^*) \mathbf{M}_{h_1} + \gamma_2^* \mathbf{M}_{g_2}] \succeq \mathbf{0}, \end{aligned} \quad (34)$$

$$C_1(\check{\beta}) = 0, C_2(\check{\beta}) = 0. \quad (35)$$

From (33), we have

$$\begin{aligned} \mathbf{0} &= [(1 - \alpha_1^* - \gamma_1^* - \gamma_2^*) \mathbf{M}_{g_1} + (\alpha_1^* + \gamma_1^*) \mathbf{M}_{h_1} \\ &\quad + \gamma_2^* \mathbf{M}_{g_2}] \check{\beta} - (1 - \alpha_1^* - \gamma_1^* - \gamma_2^*) \mathbf{E}_{g_1} \\ &\quad - (\alpha_1^* + \gamma_1^*) \mathbf{E}_{h_1} - \gamma_2^* \mathbf{E}_{g_2}, \end{aligned}$$

where  $\mathbf{E}_{g_2} = C_{g_2} \mathbf{X}_1^T \mathbf{y}_1 + D_{g_2} \mathbf{X}_2^T \mathbf{y}_2$ . Then we have

$$\begin{aligned} \check{\beta} &= [(1 - \alpha_1^* - \gamma_1^* - \gamma_2^*) \mathbf{M}_{g_1} + (\alpha_1^* + \gamma_1^*) \mathbf{M}_{h_1} + \gamma_2^* \mathbf{M}_{g_2}]^{-1} \\ &\quad \cdot [(1 - \alpha_1^* - \gamma_1^* - \gamma_2^*) \mathbf{E}_{g_1} + (\alpha_1^* + \gamma_1^*) \mathbf{E}_{h_1} + \gamma_2^* \mathbf{E}_{g_2}]. \end{aligned} \quad (36)$$

Plugging (36) into (35), we have

$$K_1(\gamma_1, \gamma_2) = C_1(\check{\beta}) = 0, K_2(\gamma_1, \gamma_2) = C_2(\check{\beta}) = 0,$$

with solution  $(\gamma_1^*, \gamma_2^*)$ . By substituting  $\gamma_1 = \gamma_1^*, \gamma_2 = \gamma_2^*$  into (36), we obtain the solution for  $\check{\beta}$ .

For step 3, the verification process is given as follows.

1. If  $\alpha_1^* + \gamma_1^* > 0$  and  $\gamma_2^* > 0$ , (14), (15), (16) are satisfied for  $\alpha_1 = \alpha_1^* + \gamma_1^*, \alpha_2 = \gamma_2^*$  and  $\beta = \check{\beta}$  based on (33), (34), (35). If we further have  $C_3(\check{\beta}) \geq 0$ , then  $\check{\beta}$  is a global minimizer of (5).
2. If  $\alpha_1^* + \gamma_1^* < 0$ , we could consider the minimization of  $h_1(\beta)$ .
3. If  $\gamma_2^* < 0$ , we consider the minimization of  $g_2(\beta)$ .

#### APPENDIX D PROOF OF LEMMA 2

Note that  $\mathbf{c}_1$  and  $\mathbf{c}_2$  are independent without considering the optimization on  $\eta_{c_1}$ . In particular, the first term in  $g(\beta, \hat{\mathbf{X}})$  only involves  $\mathbf{c}_1$  and the second term in  $g(\beta, \hat{\mathbf{X}})$  only involves  $\mathbf{c}_2$ . Thus, we firstly focus on the first term in  $g(\beta, \hat{\mathbf{X}})$  and solve the maximization with respect to  $\mathbf{c}_1$ .

$$\begin{aligned} \max_{\eta_{c_1}} \max_{\|\mathbf{d}\|_2 \leq 1} \max_{\|\mathbf{c}_1\|_2 = \eta_{c_1}} \|\mathbf{y}_1 - \mathbf{X}_1 \beta - \mathbf{c}_1 \mathbf{d}^T \beta\|_2^2 \\ = \max_{\eta_{c_1}} \max_{\|\mathbf{d}\|_2 \leq 1} \max_{\|\mathbf{c}_1\|_2 = \eta_{c_1}} (\mathbf{d}^T \beta)^2 \|\mathbf{e}_1\|_2^2 \\ = \max_{\eta_{c_1}} \max_{\|\mathbf{d}\|_2 \leq 1} (\mathbf{d}^T \beta)^2 \max_{\|\mathbf{c}_1\|_2 = \eta_{c_1}} \|\mathbf{e}_1\|_2^2, \end{aligned}$$

in which  $\mathbf{e}_1 = \mathbf{f}_1 - \mathbf{c}_1$  with  $\mathbf{f}_1 = \frac{1}{\mathbf{d}^T \beta} (\mathbf{y}_1 - \mathbf{X}_1 \beta)$ . For the maximization problem on  $\mathbf{c}_1$ , we have

$$\begin{aligned} \max_{\mathbf{c}_1} \|\mathbf{e}_1\|_2^2, \quad \text{s.t. } \|\mathbf{c}_1\|_2 = \eta_{c_1}, \\ \iff \min_{\mathbf{e}_1} -\|\mathbf{e}_1\|_2^2, \quad \text{s.t. } \|\mathbf{f}_1 - \mathbf{e}_1\|_2^2 = \eta_{c_1}^2. \end{aligned} \quad (37)$$

Although (37) is not a convex optimization problem, we can first investigate its KKT necessary conditions. The Lagrangian function of (37) is  $\mathcal{L}(\mathbf{e}_1, \gamma_{e_1}) = -\|\mathbf{e}_1\|_2^2 + \gamma_{e_1} (\|\mathbf{f}_1 - \mathbf{e}_1\|_2^2 - \eta_{c_1}^2)$ , where  $\gamma_{e_1}$  is the Lagrangian multiplier. According to the

KKT conditions, we have

$$\begin{aligned} \partial \mathcal{L}(e_1, \gamma_{e_1}) &= -2e_1^T - 2\gamma_{e_1}(\mathbf{f}_1 - e_1)^T = \mathbf{0}, \\ \|\mathbf{f}_1 - e_1\|_2^2 &= \eta_{c_1}^2, \end{aligned}$$

from which we can derive that the solution to (37) is  $e_1^* = \mathbf{f}_1 + \frac{\eta_{c_1}}{\|\mathbf{f}_1\|_2} \mathbf{f}_1$ , and the maximum value is

$$\max_{\|e_1\|_2 = \eta_{c_1}} \|e_1\|_2^2 = \|e_1^*\|_2^2 = \left(1 + \frac{\eta_{c_1}}{\|\mathbf{f}_1\|_2}\right)^2 \|\mathbf{f}_1\|_2^2.$$

Then we focus on the second term in  $g(\beta, \hat{\mathbf{X}})$ , solve the maximization on  $\eta_{c_2}$ , and derive the formulation for  $g_{m_1}(\eta_{c_1}, \beta, \mathbf{d})$ .

#### APPENDIX E

##### PROOF OF PROPOSITION 4

We observe that  $g_{m_1}(\eta_{c_1}, \beta, \mathbf{d})$  is a quadratic function with respect to  $\mathbf{d}^T \beta$ , i.e.

$$g_{m_1}(\eta_{c_1}, \beta, \mathbf{d}) = A(\mathbf{d}^T \beta)^2 + B(\mathbf{d}^T \beta) + C, \quad (38)$$

in which  $A, B, C$  are three coefficients. In particular, we have

$$\begin{aligned} A &= (C_g - D_g)\eta_{c_1}^2 + D_g\eta_c^2, \\ B &= 2[C_g\eta_{c_1}\|\mathbf{y}_1 - \mathbf{X}_1\beta\|_2 + D_g\eta_{c_2}\|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2] \geq 0, \\ C &= C_g\|\mathbf{y}_1 - \mathbf{X}_1\beta\|_2^2 + D_g\|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2^2. \end{aligned} \quad (39)$$

$$C = C_g\|\mathbf{y}_1 - \mathbf{X}_1\beta\|_2^2 + D_g\|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2^2. \quad (40)$$

Since  $A > 0$ ,  $-\frac{B}{2A} \leq 0$  and  $\mathbf{d}^T \beta \in [-\|\beta\|_2, \|\beta\|_2]$ , we can conclude that the maxima of  $g_{m_1}(\mathbf{d}, \eta_{c_1}, \beta)$  is attained when  $\mathbf{d}^T \beta = \|\beta\|_2$  and the maximum value is

$$\begin{aligned} \max_{\|\mathbf{d}\|_2 \leq 1} g_{m_1}(\eta_{c_1}, \beta, \mathbf{d}) &= C_g(\|\mathbf{y}_1 - \mathbf{X}_1\beta\|_2 + \eta_{c_1}\|\beta\|_2)^2 \\ &\quad + D_g(\|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2 + \sqrt{\eta_c^2 - \eta_{c_1}^2}\|\beta\|_2)^2, \end{aligned}$$

which provides the form of  $g_a$ .

#### APPENDIX F

##### PROOF OF LEMMA 3

In this case, the analysis for the first term in  $g(\beta, \hat{\mathbf{X}})$  remains the same. However, for the second term, we have

$$\begin{aligned} \min_{\eta_{c_2}} \min_{\|\mathbf{d}\|_2 \leq \frac{\eta}{\eta_{c_2}}} \min_{\|c_2\|_2 \leq \eta_{c_2}} \|\mathbf{y}_2 - \mathbf{X}_2\beta - c_2\mathbf{d}^T \beta\|_2^2 \\ = \min_{\eta_{c_2}} \min_{\|\mathbf{d}\|_2 \leq \frac{\eta}{\eta_{c_2}}} (\mathbf{d}^T \beta)^2 \min_{\|\mathbf{f}_2 - e_2\|_2 \leq \eta_{c_2}} \|e_2\|_2^2, \end{aligned}$$

where  $\eta_{c_2} = \sqrt{\eta_c^2 - \eta_{c_1}^2}$ ,  $\mathbf{f}_2 = \frac{1}{\mathbf{d}^T \beta}(\mathbf{y}_2 - \mathbf{X}_2\beta)$  and  $e_2 = \mathbf{f}_2 - c_2$ . Thus, the minimization on  $e_2$  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  $e_2$  is

$$\mathcal{L}(e_2, \gamma_{e_2}) = \|e_2\|_2^2 + \gamma_{e_2}(\|\mathbf{f}_2 - e_2\|_2^2 - \eta_{c_2}^2),$$

in which  $\gamma_{e_2}$  is the Lagrangian multiplier. By exploring the KKT conditions, we have

$$\nabla \mathcal{L}(e_2, \gamma_{e_2}) = 2e_2^T - 2\gamma_{e_2}(\mathbf{f}_2 - e_2)^T = 0, \quad (41)$$

$$\|\mathbf{f}_2 - e_2\|_2^2 \leq \eta_{c_2}^2,$$

$$\gamma_{e_2}(\|\mathbf{f}_2 - e_2\|_2^2 - \eta_{c_2}^2) = 0, \quad (42)$$

$$\gamma_{e_2} \geq 0.$$

By inspecting the complementary slackness condition (42), we consider two cases based on the value of  $\gamma_{e_2}$ .

*Case 1:*  $\gamma_{e_2} = 0$ . In this case, we have  $e_2 = 0$  according to (41), which can be true when  $\|\mathbf{f}_2\|_2 \leq \eta_{c_2}$ . Moreover, note that

$$\begin{aligned} \|\mathbf{f}_2\|_2 \leq \eta_{c_2} &\iff \|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2 \leq |\mathbf{d}^T \beta| \eta_{c_2} \\ &\stackrel{(a)}{\leq} \frac{\eta}{\eta_{c_2}} \|\beta\|_2 \eta_{c_2} = \eta \|\beta\|_2, \end{aligned}$$

where the equality in (a) is achieved if  $\mathbf{d} = \frac{\eta}{\eta_{c_2}\|\beta\|_2} \beta$ . Thus, if  $\|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2 \leq \eta \|\beta\|_2$ , the minimum value of  $\|e_2\|_2^2$  is 0.

*Case 2:*  $\gamma_{e_2} > 0$ . If there is no feasible solution in *Case 1*, we can conclude that  $\|\mathbf{f}_2\|_2 > \eta_{c_2}$ . Moreover, by (41) and (42), we have  $e_2^* = \frac{\gamma_{e_2}^* \mathbf{f}_2}{\gamma_{e_2}^* + 1}$ ,  $\eta_{c_2} = \|\mathbf{f}_2 - e_2^*\|_2 = \frac{1}{\gamma_{e_2}^* + 1} \|\mathbf{f}_2\|_2$ , which implies  $\gamma_{e_2}^* = \frac{\|\mathbf{f}_2\|_2}{\eta_{c_2}} - 1$ ,  $e_2^* = \mathbf{f}_2 - \frac{\eta_{c_2}}{\|\mathbf{f}_2\|_2} \mathbf{f}_2$ . Then we

have  $\min_{\|\mathbf{f}_2 - e_2\|_2 \leq \eta_{c_2}} \|e_2\|_2^2 = \|e_2^*\|_2^2 = \left(1 - \frac{\eta_{c_2}}{\|\mathbf{f}_2\|_2}\right)^2 \|\mathbf{f}_2\|_2^2$ . By combining these two cases, Lemma 3 is proved.

#### APPENDIX G

##### PROOF OF PROPOSITION 5

Now we solve the maximization problem on  $\mathbf{d}$ . Firstly, consider the case when  $\|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2 \leq \eta \|\beta\|_2$ . In this case, we notice that as long as  $\eta_{c_1} \neq 0$ ,  $g_{m_2}(\eta_{c_1}, \beta, \mathbf{d})$  is a quadratic function for  $\mathbf{d}^T \beta$  with  $A = C_g\eta_{c_1}^2 > 0$ ,  $B = 2C_g\eta_{c_1}\|\mathbf{y}_1 - \mathbf{X}_1\beta\|_2 \geq 0$  and  $-\frac{B}{2A} \leq 0$ . Thus, the maxima is attained when  $\mathbf{d}^T \beta = \|\beta\|_2$  and the maximum value of  $g(\beta, \hat{\mathbf{X}})$  is

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

For  $\eta_{c_1} = 0$ , the attacker only changes the feature matrix of the second group and the maximum value of  $g(\beta, \hat{\mathbf{X}})$  can also be derived as  $g_{b_1}(\eta_{c_1}, \beta)$ .

Secondly, consider the case when  $\|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2 > \eta \|\beta\|_2$ . In this case,  $g_{m_2}(\eta_{c_1}, \beta, \mathbf{d})$  can also be written in the form of (38) with coefficients  $A, B, C$ . In particular,  $A$  and  $C$  are defined the same as (39) and (40), and  $B$  is defined as  $B = 2C_g\eta_{c_1}\|\mathbf{y}_1 - \mathbf{X}_1\beta\|_2 - 2D_g\eta_{c_2}\|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2 \geq 0$ . 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  $\|\mathbf{y}_2 - \mathbf{X}_2\beta\|_2 > \eta \|\beta\|_2$ , the maximum value of  $g(\beta, \hat{\mathbf{X}})$  is  $g_{b_2}(\eta_{c_1}, \beta)$ .

If  $A > 0$ , we have  $-\frac{B}{2A} \leq 0$  and the maxima is attained when  $\mathbf{d}^T \beta = \|\beta\|_2$  with the maximum value

to be  $\max_{\|\mathbf{d}\|_2 \leq 1} g_{m_2}(\eta_{c_1}, \boldsymbol{\beta}, \mathbf{d}) = C_g(\|\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 + D_g(\|\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\|_2 - \eta_{c_2}\|\boldsymbol{\beta}\|_2)^2$ , which implies that

$$\begin{aligned} & \max_{\|\mathbf{d}^T\|_F \leq \eta} g(\boldsymbol{\beta}, \hat{\mathbf{X}}) \\ &= \max_{0 < \eta_c \leq \eta} \max_{0 \leq \eta_{c_1} \leq \eta_c} \max_{\|\mathbf{d}\|_2 \leq 1} g_{m_2}(\eta_{c_1}, \boldsymbol{\beta}, \mathbf{d}) \\ &\stackrel{(a)}{=} \max_{0 \leq \eta_{c_1} \leq \eta} [C_g(\|\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 \\ &\quad + D_g(\|\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\|_2 - \max_{0 < \eta_c \leq \eta} \eta_{c_2}\|\boldsymbol{\beta}\|_2)^2] \\ &= \max_{0 \leq \eta_{c_1} \leq \eta} g_{b_2}(\eta_{c_1}, \boldsymbol{\beta}), \end{aligned}$$

where (a) follows from the fact that  $D_g < 0$  and  $\|\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\|_2 > \eta\|\boldsymbol{\beta}\|_2 \geq \eta_{c_2}\|\boldsymbol{\beta}\|_2$ .

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  $\mathbf{d}^T \boldsymbol{\beta}$  and the maxima is attained when  $\mathbf{d}^T \boldsymbol{\beta} = \|\boldsymbol{\beta}\|_2$  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  $\mathbf{d}^T \boldsymbol{\beta}$  with  $-\frac{B}{2A} > \frac{(\frac{\lambda}{n-m} - \frac{1}{n})\eta_{c_2}\|\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\|_2}{-(\frac{\lambda}{n-m} - \frac{1}{n})\eta_{c_1}^2 + (\frac{\lambda}{n-m} - \frac{1}{n})\eta_c^2} \stackrel{(g)}{>} \frac{\eta_{c_2}\eta\|\boldsymbol{\beta}\|_2}{\eta_{c_2}^2} \geq \|\boldsymbol{\beta}\|_2$ , in which (g) is from the fact that  $\|\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\|_2 > \eta\|\boldsymbol{\beta}\|_2$ . Thus, the maxima is attained when  $\mathbf{d}^T \boldsymbol{\beta} = \|\boldsymbol{\beta}\|_2$  and the maximum value is also  $g_{b_2}(\eta_{c_1}, \boldsymbol{\beta})$ .

#### APPENDIX H PROOF OF LEMMA 4

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  $\eta_{c_1}$ , we have  $\frac{\partial^2 g_a(\eta_{c_1}, \boldsymbol{\beta})}{\partial \eta_{c_1}^2} = 2 \left( \frac{\lambda}{m} + \frac{\lambda}{n-m} \right) \|\boldsymbol{\beta}\|_2^2 - 2D_g \frac{\eta_c^2}{\eta_{c_2}^2} \|\boldsymbol{\beta}\|_2 \|\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\|_2$ . 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  $\eta_{c_1}$ .

For  $\boldsymbol{\beta}$ , we have

$$\begin{aligned} \frac{\partial^2 g_a(\eta_{c_1}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} &\geq 2C_g \left[ \eta_{c_1} \left( \eta_{c_1} - 2 \frac{\text{Tr}(\mathbf{X}_1^T \mathbf{X}_1)}{\|\mathbf{X}_1\|_F} \right) \mathbf{I} + \mathbf{X}_1^T \mathbf{X}_1 \right] \\ &\quad + 2D_g \left[ \eta_{c_2} \left( \eta_{c_2} - 2 \frac{\text{Tr}(\mathbf{X}_2^T \mathbf{X}_2)}{\|\mathbf{X}_2\|_F} \right) \mathbf{I} + \mathbf{X}_2^T \mathbf{X}_2 \right]. \end{aligned}$$

Since  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are feature matrices with finite norm, there always exist  $\rho_2 < \infty$  such that  $\frac{\partial^2 g_a(\eta_{c_1}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \succeq -\rho_2 \mathbf{I}$ , which indicates that  $g_a$  is weakly-convex in  $\boldsymbol{\beta}$ .

#### APPENDIX I PROOF OF LEMMA 5

For  $g_a$ , we have

$$\begin{aligned} \frac{\partial g_a(\eta_{c_1}, \boldsymbol{\beta})}{\partial \eta_{c_1}} &= 2C_g \|\boldsymbol{\beta}\|_2 \|\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\|_2 + 2(C_g - D_g) \eta_{c_1} \|\boldsymbol{\beta}\|_2^2 \\ &\quad - 2D_g \frac{\eta_{c_1}}{\eta_{c_2}} \|\boldsymbol{\beta}\|_2 \|\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\|_2 = 0, \end{aligned}$$

which implies

$$\begin{aligned} & \left( \eta_{c_1} \|\boldsymbol{\beta}\|_2 + \frac{C_g}{C_g - D_g} \|\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\|_2 \right) \\ & \cdot \left( \eta_{c_2} \|\boldsymbol{\beta}\|_2 - \frac{D_g}{C_g - D_g} \|\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\|_2 \right) \\ &= - \frac{C_g D_g}{\left( \frac{\lambda}{n-m} + \frac{\lambda}{m} \right)^2} \|\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}\|_2 \|\mathbf{y}_2 - \mathbf{X}_2\boldsymbol{\beta}\|_2. \quad (43) \end{aligned}$$

From (43), we note that  $\eta_{c_1}$  and  $\eta_{c_2}$  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 (43) (which can be seen geometrically), denoted as  $\eta_{c_1}^*$ . Moreover, we have

- $\eta_{c_1} < \eta_{c_1}^*$ , left hand side of (43) is positive;
- $\eta_{c_1} > \eta_{c_1}^*$ , left hand side of (43) 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.

#### REFERENCES

- [1] Y. Jin and L. Lai, "Adversarially robust fairness-aware regression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
- [2] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," Jul. 2018, *arXiv:1808.00023*.
- [3] N. Martinez, M. Bertran, and G. Sapiro, "Minimax pareto fairness: A multi objective perspective," in *Proc. Int. Conf. Mach. Learn.*, Vienna, Austria, Nov. 2020, pp. 6755–6764.
- [4] A. Agarwal, M. Dudík, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, May 2019, pp. 120–129.
- [5] E. Chzhen, C. Denis, M. Hebrici, L. Oneto, and M. Pontil, "Fair regression with Wasserstein barycenters," in *Proc. Advances Neural Inf. Process. Syst.*, Dec. 2020, vol. 33, pp. 7321–7331.
- [6] G. Zalcberg and A. Wiesel, "Fair principal component analysis and filter design," *IEEE Trans. Signal Process.*, vol. 69, pp. 4835–4842, 2021.
- [7] J. Chi, Y. Tian, G. J. Gordon, and H. Zhao, "Understanding and mitigating accuracy disparity in regression," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 1866–1876.
- [8] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," Dec. 2017, *arXiv:1712.05526*.
- [9] Y. Jin and L. Lai, "On the adversarial robustness of hypothesis testing," *IEEE Trans. Signal Process.*, vol. 69, pp. 515–530, 2021.
- [10] F. Li, L. Lai, and S. Cui, "Optimal feature manipulation attacks against linear regression," *IEEE Trans. Signal Process.*, vol. 69, pp. 5580–5594, 2021.
- [11] D. Solans, B. Biggio, and C. Castillo, "Poisoning attacks on algorithmic fairness," Apr. 2020, *arXiv:2004.07401*.
- [12] N. Mehrabi, M. Naveed, F. Morstatter, and A. Galstyan, "Exacerbating algorithmic bias through fairness attacks," Dec. 2020, *arXiv:2012.08723*.
- [13] M.-H. Van, W. Du, X. Wu, and A. Lu, "Poisoning attacks on fair machine learning," Oct. 2021, *arXiv:2110.08932*.
- [14] H. Chang, T. D. Nguyen, S. K. Murakonda, E. Kazemi, and R. Shokri, "On adversarial bias and the robustness of fair machine learning," Jun. 2020, *arXiv:2006.08669*.
- [15] J. Jiang and X. Chen, "Optimality conditions for nonsmooth nonconvex-nonconcave min-max problems and generative adversarial networks," Mar. 2022, *arXiv:2203.10914*.
- [16] T. Lin, C. Jin, and M. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2020, pp. 6083–6093.
- [17] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen, "Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications," *IEEE Trans. Signal Process.*, vol. 68, pp. 3676–3691, 2020.

- [18] S. Lee and D. Kim, "Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems," in *Proc. Advances Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 22588–22600.
- [19] D. M. Ostrovskii, B. Barzandeh, and M. Razaviyayn, "Nonconvex-nonconcave min-max optimization with a small maximization domain," Oct. 2021, *arXiv:2110.03950*.
- [20] Z.-Q. Luo, N. D. Sidiropoulos, P. Tseng, and S. Zhang, "Approximation bounds for quadratic optimization with homogeneous quadratic constraints," *SIAM J. Optim.*, vol. 18, no. 1, pp. 1–28, Jan. 2007.
- [21] K. Huang and N. D. Sidiropoulos, "Consensus-ADMM for general quadratically constrained quadratic programming," *IEEE Trans. Signal Process.*, vol. 64, no. 20, pp. 5297–5310, Jul. 2016.
- [22] M. Liu, H. Rafique, Q. Lin, and T. Yang, "First-order convergence theory for weakly-convex-weakly-concave min-max problems," *J. Mach. Learn. Res.*, vol. 22, pp. 169–1, Jan. 2021.
- [23] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," Nov. 2016, *arXiv:1611.01236*.
- [24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," Jun. 2017, *arXiv:1706.06083*.
- [25] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, Long Beach, CA, USA, Jun. 2019, pp. 7472–7482.
- [26] F. Li, L. Lai, and S. Cui, "On the adversarial robustness of subspace learning," *IEEE Trans. Signal Process.*, vol. 68, pp. 1470–1483, 2020.
- [27] H. Xu, X. Liu, Y. Li, A. Jain, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 11492–11501.
- [28] Y. Roh, K. Lee, S. Whang, and C. Suh, "Sample selection for fair and robust training," in *Proc. Advances Neural Inf. Process. Syst.*, Dec. 2021, vol. 34, pp. 815–827.
- [29] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: An application to recidivism prediction," Jun. 2018, *arXiv:1807.00199*.
- [30] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 3384–3393.
- [31] P. Benz, C. Zhang, S. Ham, A. Karjauv, G. Cho, and I. S. Kweon, "Trade-off between accuracy, robustness, and fairness of deep classifiers," presented at the Workshop Adversarial Mach. Learn. Real-World Comput. Vision Syst. Online Challenges, Nashville, TN, USA, Jun. 2021.
- [32] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson, "Fairness through robustness: Investigating robustness disparity in deep learning," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Mar. 2021, pp. 466–477.
- [33] Q. Tian, K. Kuang, K. Jiang, F. Wu, and Y. Wang, "Analysis and applications of class-wise robustness in adversarial training," in *Proc. ACM Conf. Knowl. Discovery Data Mining*, Virtual Event, Singapore, Aug. 2021, pp. 1561–1570.
- [34] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Sydney, Australia, Aug. 2015, pp. 259–268.
- [35] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Advances Neural Inf. Process. Syst.*, Barcelona, Spain, vol. 29, Dec. 2016, pp. 3323–3331.
- [36] A. Shah et al., "Selective regression under fairness criteria," Oct. 2021, *arXiv:2110.15403*.
- [37] V. Jeyakumar, A. M. Rubinov, and Z.-Y. Wu, "Non-convex quadratic minimization problems with quadratic constraints: Global optimality conditions," *Math. Program.*, vol. 110, no. 3, pp. 521–541, Sep. 2007.
- [38] L. F. Wightman, "LSAC national longitudinal bar passage study. LSAC research report series," 1998.
- [39] B. Lantz, *Machine Learning with R: Expert Techniques for Predictive Modeling*. Birmingham, U.K.: Packt Publishing, 2019.
- [40] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, "Fair kernel learning," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Skopje, Macedonia, Springer, Sep. 2017, pp. 339–355.
- [41] H. Hmam, "Quadratic optimisation with one quadratic equality constraint," Defence Science and Technology Organisation Edinburgh Electronic Warfare and Radar Division, Canberra, Australia, Tech. Rep. 2416, Jun. 2010.