# ADVERSARIALLY ROBUST FAIRNESS-AWARE REGRESSION

*Yulu Jin and Lifeng Lai*

Department of ECE, University of California, Davis
Email:{yuljin,lflai}@ucdavis.edu

## ABSTRACT

Fairness and robustness are critical elements of trustworthy machine learning systems that need to be addressed. Using a minimax framework, in this paper, we aim to design an adversarially robust fair regression model that achieves optimal performance in the presence of an attacker who is able to 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. Based on two real-world datasets, numerical results illustrate that the proposed adversarially robust fair model has better performance on the poisoned dataset than other fair machine learning models in both prediction accuracy and group-based fairness measure.

## 1. 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 [1], 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 [2, 3, 4, 5].

Since a large body of work has shown that machine learning models are vulnerable to various types of attacks [6, 7], 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 [8].

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, in this paper, we focus on regression problems and design a fair regression model that is robust to rank-one adversarial attacks. The design of such robust model is formulated as a minimax game between a defender aiming to minimize the accuracy loss and bias, and an attacker aiming to maximize these objectives. The attacker is assumed to be able to observe the entire training dataset and carefully modify the training data so as to reduce the accuracy and fairness of the regression model. To characterize both the prediction and fairness performance of a model, the objective function is selected to include both 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 simultaneously.

One major challenge of the work is that the proposed minimax problem is nonsmooth nonconvex-nonconcave, which may not have a local saddle point in general [9]. Although there exist many iterative methods for finding stationary point or local optima of nonconvex-nonconcave minimax problems [10, 11], the non-smooth terms they consider usually have special forms. However, the proposed realistic minimax problem does not satisfy assumptions made in these papers. To solve the complicated minimax problem in hand, we first investigate the inner maximization problem and then the outer minimization problem. Through various transformations, the original nonconvex-nonconcave minimax problem for two vectors will be converted into a weakly-convex-weakly-concave minimax problem for one vector and one scalar, which can be approximately solved using existing algorithms such as [12].

Using 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 fairness. On the other hand, for other fair regression models, at least one performance metric will be severely affected by the rank-one attack.

This work was supported in part by National Science Foundation under Grants CCF-1717943, CNS-1824553, CCF-1908258 and ECCS-2000415.

## 2. 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, [8] provides three online attacks based on different group-based fairness measures, and [13] 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 [14, 15]. 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 present a similar issue [16].

**Intersection of fairness and robustness.** Fairness and robustness are critical elements of trustworthy AI that need to be addressed together. 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 [17, 18]. For example, [17] 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 sample selection-based method is shown to obtain more fair standard and robust classifiers [19]. Thirdly, [20] investigates the class-wise robustness and proposes methods to improve the robustness of the most vulnerable class, so as to obtain a fairer robust model.

## 3. PROBLEM FORMULATION

Using a set of training samples $\{\boldsymbol{x}_i, y_i, G_i\}_{i=1}^n \coloneqq \{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{G}\}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the feature vector, $y_i$ is the response variable and $G_i$ indicates the group membership or sensitive status (for example, race, gender), we aim to develop a model that can predict the value of a target variable $Y$ from the input variables $X$. In this 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. We denote

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}.$$

To build a robust model, we assume that there is an 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. In particular, the attacker will carefully design a rank-one feature modification matrix $\boldsymbol{\Delta}$ and add it to the original feature matrix $\boldsymbol{X}$, so as to obtain the modified feature matrix $\hat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta}$. Since $\boldsymbol{\Delta}$ is of rank 1, we assume that $\boldsymbol{\Delta} = \boldsymbol{c}\boldsymbol{d}^T$, where $\boldsymbol{c} \in \mathbb{R}^n$ and $\boldsymbol{d} \in \mathbb{R}^p$. Moreover, recall that there are samples from two groups, we denote the modification matrix of the first group as $\boldsymbol{\Delta}_1$, i.e., the first $m$ rows of $\boldsymbol{\Delta}$, and assume that $\boldsymbol{\Delta}_1 = \boldsymbol{c}_1\boldsymbol{d}^T$, where $\boldsymbol{c}_1$ consists of the first $m$ components of $\boldsymbol{c}$. Similarly, for the second group, the modification matrix is $\boldsymbol{\Delta}_2 = \boldsymbol{c}_2\boldsymbol{d}^T$. Then the modified feature matrices for two groups are $\hat{\boldsymbol{X}}_1 = \boldsymbol{X}_1 + \boldsymbol{\Delta}_1$ and $\hat{\boldsymbol{X}}_2 = \boldsymbol{X}_2 + \boldsymbol{\Delta}_2$.

From the poisoned dataset $\{\hat{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{G}\}$, we aim to design a robust fairness-aware regression model. In order to characterize both the prediction and the fairness performance, we consider the following objective function

$$L = f(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) + \lambda F(\boldsymbol{\beta}, \hat{\boldsymbol{X}}), \tag{1}$$

where $\boldsymbol{\beta}$ is the regression coefficient, $f(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ corresponds to the prediction accuracy loss, $F(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ corresponds to the group fairness gap and $\lambda$ is the trade-off parameter. The goal of the adversary is to design $\boldsymbol{\Delta}$ 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 rank-one attack and use the Frobenius norm to measure the energy of the modification matrix $\boldsymbol{\Delta}$. Thus, we have the minimax problem

$$\min_{\boldsymbol{\beta}} \max_{\|\Delta\|_F \leq \eta} f(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) + \lambda F(\boldsymbol{\beta}, \hat{\boldsymbol{X}}), \tag{2}$$

where $\eta$ is the energy budget. To measure the prediction accuracy, we consider the mean-squared error (MSE), $f(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = \mathbb{E}[(Y - \hat{Y})^2]$, where $\hat{Y}$ is the prediction result. For the group fairness gap, we consider a measure that is closely related to the accuracy parity criterion [21] and use $F(\boldsymbol{\beta}, \hat{\boldsymbol{X}}) = |\mathbb{E}[(Y - \hat{Y})^2|G = 1] - \mathbb{E}[(Y - \hat{Y})^2|G = 2]|$ to measure the severity of violations.

## 4. PROPOSED METHOD

For the objective function of the formulated minimax problem (2), we have

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

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

**Lemma 4.1.** *For $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ and $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$, we have that*

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

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

Based on Lemma 4.1, we can solve the maximization problem in (2). First, note that

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

which indicates that the maximization problem can be separated into two sub-problems. For $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$, we have

- if $a' \geq 0$, the maximum value of $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ is $h_a(\eta_{c_1}, \boldsymbol{\beta}) = a'(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 + b'(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta^2 - \eta_{c_1}^2}\|\boldsymbol{\beta}\|_2)^2$;
- if $a' < 0$, the maximum value of $h(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$ is

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

with $h_{b_1}(\eta_{c_1}, \boldsymbol{\beta}) = b'(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta^2 - \eta_{c_1}^2}\|\boldsymbol{\beta}\|_2)^2$, $h_{b_2}(\eta_{c_1}, \boldsymbol{\beta}) = a'(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 - \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 + b'(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta^2 - \eta_{c_1}}\|\boldsymbol{\beta}\|_2)^2$.

Similar results apply to $g(\boldsymbol{\beta}, \hat{\boldsymbol{X}})$.

Subsequently, the minimax problem (2) can be transformed to a minimax problem for one vector and one scalar with a piece-wise max-type ojective function. For example, if $b \geq 0$ and $a' < 0$, (2) can be written as

$$\min_{\beta} \max_{0 \leq \eta_{c_1} \leq \eta} \max\{g_a(\eta_{c_1}, \boldsymbol{\beta}), h_b(\eta_{c_1}, \boldsymbol{\beta})\}, \tag{3}$$

where

$$\begin{aligned}g_a(\eta_{c_1}, \boldsymbol{\beta}) \quad = \quad & a(\|\boldsymbol{y}_1 - \boldsymbol{X}_1\boldsymbol{\beta}\|_2 + \eta_{c_1}\|\boldsymbol{\beta}\|_2)^2 \\ & + b(\|\boldsymbol{y}_2 - \boldsymbol{X}_2\boldsymbol{\beta}\|_2 + \sqrt{\eta^2 - \eta_{c_1}^2}\|\boldsymbol{\beta}\|_2)^2.\end{aligned}$$

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

**Lemma 4.2.** *If the norm of $\boldsymbol{\beta}$ is bounded, i.e. $\|\boldsymbol{\beta}\|_2 \leq B_\beta$, then we have*

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

*(ii) $h_b$ 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 $\boldsymbol{\beta}$ and weakly-convex in $\boldsymbol{\beta}$ for any given $\eta_{c_1}$.*

**Lemma 4.3.** *For any given $\boldsymbol{\beta}$, $g_a$ and $h_{b_2}$ are unimodal functions with respect to $\eta_{c_1}$ that increase first and then decrease.*

Note that (3) can be transformed to three sub-problems:

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

For the sub-problem 1, the maximization on $\eta_{c_1}$ 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 $\|\boldsymbol{\beta}\|_2 \leq B_\beta$, which is reasonable in reality, and have that: 1) the feasible set $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2 \leq B_\beta\} \times [0, \eta]$ is convex and compact; 2) the objective function is weakly-convex-weakly-concave by Lemma 4.2; 3) the saddle-point exists by Lemma 4.3. Based on those properties, we are able to apply a first-order algorithms proposed by [12] to solve the non-convex non-concave minimax problem as in sub-problem 2 and derive the nearly $\epsilon$-stationary solution. In particular, define $\mathcal{Z} = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_2 \leq B_\beta\} \times [0, \eta]$ and the mapping $H(\boldsymbol{z}) \coloneqq (\partial_{\boldsymbol{\beta}} h_{b_2}(\eta_{c_1}, \boldsymbol{\beta}), \partial_{\eta_{c_1}}[-h_{b_2}(\eta_{c_1}, \boldsymbol{\beta})])^T$, where $\boldsymbol{z} = (\boldsymbol{\beta}, \eta_{c_1})$. The minty variational inequality (MVI) problem corresponding to the saddle-point problem in sub-problem 2 is to find $\boldsymbol{z}^* \in \mathcal{Z}$ such that $\langle \boldsymbol{\xi}, \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0, \forall \boldsymbol{z} \in \mathcal{Z}, \forall \boldsymbol{\xi} \in H(\boldsymbol{z})$. Then the saddle-point problem can be solved through the lens of MVI. In [12], the proposed inexact proximal point method consists of approximately solving a sequence of strongly monotone MVIs constructed by adding a strongly monotone mapping to $H(\boldsymbol{z})$ with a sequentially updated proximal center. Thus, the complex non-convex non-concave minmax problem can be decomposed into a sequence of easier strongly-convex strongly-concave problems.

## 5. NUMERICAL RESULTS

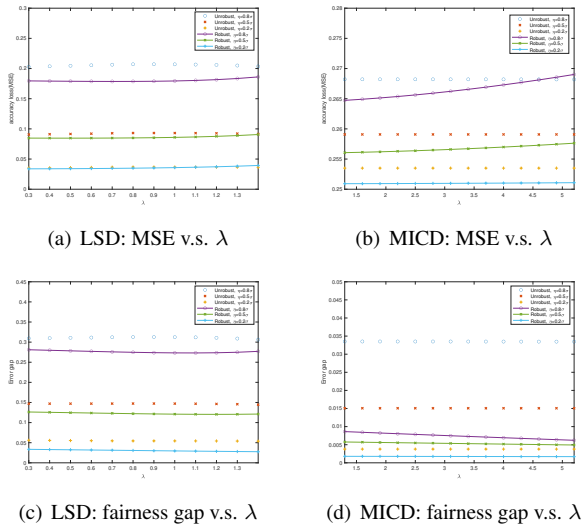
In this section, we test our proposed adversarially robust models on practical regression problems. We conduct experiments on two real-world datasets: Law School Dataset (LSD) [22] and Medical Insurance Cost Dataset (MICD) [23].

For comparison purpose, we introduce an unrobust fair regression model that does not consider the existence of the adversary. In particular, the unrobust fair model is

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

Moreover, 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 [24]. The optimal regression coefficient for each model is derived by fitting the model on the original dataset $\{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{G}\}$. To obtain the performance of each model on the poisoned dataset, we apply the derived optimal regression coefficient on the poisoned dataset, $\{\hat{\boldsymbol{X}}, \boldsymbol{y}, \boldsymbol{G}\}$, and calculate the MSE as well as the group fairness gap.

In the first experiment, for our proposed adversarially robust fair model and unrobust fair model, 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. 1, 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. 1(b), 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 for 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.

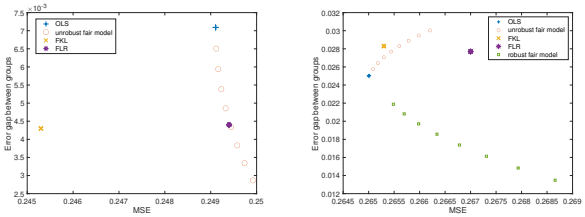


(a) LSD: MSE v.s. $\lambda$ (b) MICD: MSE v.s. $\lambda$

(c) LSD: fairness gap v.s. $\lambda$ (d) MICD: fairness gap v.s. $\lambda$

**Fig. 1**. Effects of $\lambda$ and $\eta$ on MSE and fairness gap

In the second experiment, we compare our proposed adversarially robust fair model with other fair regression models. In Fig. 2, we provide the performance of different regression models on the original dataset as well as the poisoned dataset with $\eta = 0.5\sigma$. For the unrobust fair model and adversarially robust fair model, since the choice of the trade-off parameter $\lambda$ will affect the model performance, we explore models with various choices of $\lambda$. As shown in Fig. 2(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. 2(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 the MSE locates in the lower left corner and is convex. Thus, by appropriately choosing the value of $\lambda$, a model that performs well in terms of both fairness and prediction accuracy can be obtained.




(a) Original dataset (b) Poisoned dataset

**Fig. 2**. MICD: fairness gap v.s. MSE.

## 6. CONCLUSION

In this paper, we have proposed a minimax framework to characterize the best attacker that generates the optimal rank-one attack on 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 provided a method to solve the proposed nonsmooth nonconvex-nonconcave minimax problem. Moreover, we have performed numerical experiments on two real-world datasets and shown that the proposed adversarially robust fair model can achieve better performance in prediction accuracy and fairness guarantee than other fair models with a proper choice of $\lambda$.

## 7. REFERENCES


[1] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, Jul. 2018.

[2] N. Goel, M. Yaghini, and B. Faltings, "Non-discriminatory machine learning through convex fairness criteria," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 32, (New Orleans, LA), Apr. 2018.

[3] H. Zhao and G. Gordon, "Inherent tradeoffs in learning fair representations," in *Proc. Advances in Neural Information Processing Systems*, vol. 32, (Vancouver, Canada), Dec. 2019.

[4] J. Fitzsimons, A. Al Ali, M. Osborne, and S. Roberts, "A general framework for fair regression," *Entropy*, vol. 21, no. 8, p. 741, Aug. 2019.

[5] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Fair regression with wasserstein barycenters," in *Proc. Advances in Neural Information Processing Systems*, vol. 33, pp. 7321–7331, Dec. 2020.

[6] B. Li, Y. Wang, A. Singh, and Y. Vorobeychik, "Data poisoning attacks on factorization-based collaborative filtering," in *Proc. Advances in neural information processing systems*, vol. 29, (Barcelona, Spain), Dec. 2016.

[7] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Proc. Advances in neural information processing systems*, vol. 31, (Montréal, Canada), Dec. 2018.

[8] M.-H. Van, W. Du, X. Wu, and A. Lu, "Poisoning attacks on fair machine learning," *arXiv preprint arXiv:2110.08932*, Oct. 2021.

[9] J. Jiang and X. Chen, "Optimality conditions for nonsmooth nonconvex-nonconcave min-max problems and generative adversarial networks," *arXiv preprint arXiv:2203.10914*, Mar. 2022.

[10] T. Lin, C. Jin, and M. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *Proc. International Conference on Machine Learning*, pp. 6083–6093, Jul. 2020.

[11] D. M. Ostrovskii, B. Barazandeh, and M. Razaviyayn, "Nonconvex-nonconcave min-max optimization with a small maximization domain," *arXiv preprint arXiv:2110.03950*, Oct. 2021.

[12] 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.

[13] H. Chang, T. D. Nguyen, S. K. Murakonda, E. Kazemi, and R. Shokri, "On adversarial bias and the robustness of fair machine learning," *arXiv preprint arXiv:2006.08669*, Jun. 2020.

[14] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, Jun. 2017.

[15] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. International conference on machine learning*, (Long Beach, CA), pp. 7472–7482, Jun. 2019.

[16] 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. International Conference on Machine Learning*, pp. 11492–11501, Jul. 2021.

[17] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: an application to recidivism prediction," *arXiv preprint arXiv:1807.00199*, Jun. 2018.

[18] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in *Proc. International Conference on Machine Learning*, (Stockholm, Sweden), pp. 3384–3393, Jul. 2018.

[19] Y. Roh, K. Lee, S. Whang, and C. Suh, "Sample selection for fair and robust training," in *Proc. Advances in Neural Information Processing Systems*, vol. 34, pp. 815–827, Dec. 2021.

[20] V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson, "Fairness through robustness: Investigating robustness disparity in deep learning," in *Proc. ACM Conference on Fairness, Accountability, and Transparency*, pp. 466–477, Mar. 2021.

[21] J. Chi, Y. Tian, G. J. Gordon, and H. Zhao, "Understanding and mitigating accuracy disparity in regression," *arXiv preprint arXiv:2102.12013*, Feb. 2021.

[22] L. F. Wightman, "Lsac national longitudinal bar passage study. lsac research report series.," 1998.

[23] B. Lantz, *Machine learning with R: expert techniques for predictive modeling*. Packt Publishing ltd, 2019.

[24] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, "Fair kernel learning," in *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, (Skopje, Macedonia), pp. 339–355, Springer, Sep. 2017.