TO BE ROBUST AND TO BE FAIR: ALIGNING FAIRNESS WITH ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial training has been shown to be reliable in improving robustness against adversarial samples. However, the problem of adversarial training in terms of fairness has not yet been properly studied, and the relationship between fairness and accuracy attack still remains unclear. Can we simultaneously improve robustness w.r.t. both fairness and accuracy? To tackle this topic, in this paper, we study the problem of adversarial training and adversarial attack w.r.t. both metrics. We propose a unified structure for fairness attack which bring together common notions in group fairness, and we theoretically prove the equivalence of fairness attack against different notions. We show the alignment of fairness and accuracy attack in disadvantaged groups, and we theoretically demonstrate that robustness of samples w.r.t. adversarial attack against one metric also benefit from robustness of samples w.r.t. adversarial attack against the other metric. Our work unifies adversarial training and attack w.r.t. fairness and accuracy, where both metrics benefit from robustness of the other metric under adversarial attack. Our study suggests a novel way to incorporate adversarial training with fairness, and experimental results show that our proposed method achieves better performance in terms of robustness w.r.t. both fairness and accuracy.

1 INTRODUCTION

As machine learning systems have been increasingly applied in social fields, it is imperative that machine learning models do not reflect real-world discrimination. However, machine learning models have been shown to have biased predictions against disadvantaged groups on several real-world tasks (Larson et al., 2016; Dressel & Farid, 2018; Mehrabi et al., 2021a). In order to improve fairness and reduce discrimination of machine learning systems, a variety of work has been proposed to quantify and rectify bias in machine learning systems (Hardt et al., 2016; Kleinberg et al., 2016; Mitchell et al., 2018).

Despite the emerging interest in fairness, the topic of adversarial fairness attack has not yet been properly discussed. Most of current literature on adversarial training has been focusing on improving robustness against accuracy attack (Chakraborty et al., 2018), while the problem of adversarial attack and adversarial training w.r.t. fairness has been rarely addressed. Although adversarial training has been widely discussed in fairness literature, much of them have been focusing on applying adversarial learning as means to unlearn the impact of sensitive attributes to achieve fairness (Madras et al., 2018; Creager et al., 2019). Mehrabi et al. (2021b) and Solans et al. (2020) made the first attempt to propose novel ways to generate adversarial samples taking into account fairness objectives to disturb the training process and exacerbate bias on clean testing data. However, the relationship between fairness and accuracy attack is understudied, and it remains unclear whether it is possible to incorporate adversarial training with fairness to improve robustness against fairness attack.

In light of current limitations on adversarial attack and adversarial training w.r.t. fairness, in this work, we propose a general framework for fairness attack, where we show that different notions in group fairness have similar targets in terms of adversarial attack, and specifically, in terms of gradient-based attack, these notions are equivalent. Based on this unified framework, we discuss the relationship between fairness and accuracy attack, and we show theoretically how robustness w.r.t. fairness and accuracy can benefit from the other, i.e., the alignment between the two notions in terms of robustness. Our discussion suggest a novel framework, *fair adversarial training*, to

incorporate fair classification with adversarial training to improve robustness against fairness attack. We summarize our contribution as follows:

- We propose a unified framework for adversarial attack fairness, which brings together different notions in group fairness.
- We theoretically demonstrate the alignment between robustness of fairness and accuracy, and we propose a *fair adversarial training* approach that incorporates adversarial training with fair classification.
- We empirically validate the superiority of our method under adversarial attack, and the connection between robustness w.r.t. fairness and accuracy on three benchmark datasets.

2 Related Work

2.1 Fairness in machine learning

Fairness has gained much attention in machine learning society. Different notions have been proposed to quantify discrimination of machine learning models, including individual fairness (Lahoti et al., 2019; John et al., 2020; Mukherjee et al., 2020), group fairness (Feldman et al., 2015; Hardt et al., 2016; Zafar et al., 2017) and counterfactual fairness (Kusner et al., 2017). Our work is most closely related with group fairness notion. Works on group fairness generally falls into three categories: preprocessing (Creager et al., 2019; Jiang & Nachum, 2020; Jang et al., 2021), where the goal is to adjust training distribution to reduce discrimination; inprocessing (Zafar et al., 2017; Jung et al., 2021; Roh et al., 2021), where the goal is to impose fairness constraint during training by reweighing or adding relaxed fairness regularization; and postprocessing (Hardt et al., 2016; Jang et al., 2022), where the goal is to adjust the decision threshold in each sensitive group to achieve expected fairness parity.

2.2 adversarial machine learning

Adversarial training and adversarial attack have been widely studied in trustworthy machine learning. Goodfellow et al. (2014) proposes a simple one-step gradient-based attack to adversarially perturb the predicted label. Madry et al. (2017) extends the one-step attack to an iterative attack strategy, and shows that iterative strategy is better at finding adversarial samples. Accordingly, different methods on adversarial attack and defenses have been proposed (Shafahi et al., 2019; Wong et al., 2020; Xie et al., 2020; Cui et al., 2021; Jia et al., 2022) to improve robustness of classifier against accuracy attack. However, few literature has addressed adversarial training and attack against fairness. Some work discusses the problem of fairness posioning attack during training (Solans et al., 2020; Mehrabi et al., 2021b); however, it is not clear how fairness attack would influence the predicted soft labels, and the relationship between fairness and accuracy attack/robustness remains unclear.

3 Problem Definition

3.1 adversarial attack w.r.t. accuracy

We start by formulating adversarial attack against accuracy. Denote $x \in \mathbb{R}^n$ as the input feature, $y \in \{0, 1\}$ as the label, and $a \in \{0, 1\}$ as the one-hot encoding sensitive attribute[1]. Let $f : \mathbb{R}^n \to [0, 1]$ be the function of classifier, the objective of adversarial attack against accuracy for each sample (x, y, a) can be formulated as

$$\arg\max_{\epsilon} L_{CE}(f(x + \epsilon), y), \|\epsilon\| \leq \epsilon_0,$$

where $\|\epsilon\|$ refers to the L^p norm of ϵ with a general choice of perturbation constraint in L^∞ norm, and L_{CE} is the cross-entropy loss. A common way to obtain adversarial samples is through projected gradient descent (PGD) attack, where the adversarial sample is updated in each step based on

[1]We formulate the problem under binary classification and binary sensitive attribute for simplicity; however, our method can be readily generalized to multi-class and multi-sensitive-attribute scenario.

the signed gradient:

$$x^{t+1} = \Pi_{x+S}\left(x^t + \alpha \operatorname{sign}\left(\nabla_x L_{CE}(x, y)\right)\right),$$

where α is the step size, and $S := \{\epsilon, \|\epsilon\| \leq \epsilon_0\}$ is the set of allowed perturbation. PGD attack has been shown to be effective in finding adversarial samples compared with one-step adversarial attack (Madry et al., 2017).

3.2 ADVERSARIAL ATTACK W.R.T. FAIRNESS

Fairness adversarial attack have yet been less studied in current literature. In light of the formulation of accuracy adversarial attack, we propose to formulate fairness adversarial attack as follows:

$$\arg\max_{\epsilon} L(f(x+\epsilon), y), \|\epsilon\| \leq \epsilon_0,$$

where L is some relaxed fairness constraint. Below we discuss the specific formulation of L for two widely adopted group fairness notions: disparate impact (DI) and equalized odds (EOd).

Consider a testing set $\mathbb{S} = \{(x_i, y_i, a_i), 1 \leq i \leq N\}$ and denote $\mathbb{S}_{jk} = \{x_i | y_i = j, a_i = k\}$, and $\mathbb{S}_{.k} = \{x_i | a_i = k\}$. The relaxations (Madras et al., 2018; Wang et al., 2022) for fairness attack corresponding to DI and EOd can be formulated as:

$$L_{DI} = \left| \sum_{x_i \in \mathbb{S}_{.1}} \frac{f(x_i)}{|\mathbb{S}_{.1}|} - \sum_{x_i \in \mathbb{S}_{.0}} \frac{f(x_i)}{|\mathbb{S}_{.0}|} \right|, \tag{1}$$

$$L_{EOd} = \left| \sum_{x_i \in \mathbb{S}_{00}} \frac{f(x_i)}{|\mathbb{S}_{00}|} - \sum_{x_i \in \mathbb{S}_{01}} \frac{f(x_i)}{|\mathbb{S}_{01}|} \right| + \left| \sum_{x_i \in \mathbb{S}_{10}} \frac{f(x_i)}{|\mathbb{S}_{10}|} - \sum_{x_i \in \mathbb{S}_{11}} \frac{f(x_i)}{|\mathbb{S}_{11}|} \right|, \tag{2}$$

And fair adversarial samples can be obtained similarly via PGD attack:

$$x^{t+1} = \Pi_{x+S}\left(x^t + \alpha \operatorname{sign}\left(\nabla_x L(x, y)\right)\right).$$

4 FAIR ADVERSARIAL TRAINING

4.1 EQUIVALENCE BETWEEN EOD AND DI ATTACK

We now discuss the detailed relationship between DI and EOd attack. The following corollary states the compatibility of the two objectives:

Corollary 1. *The adversarial objective of EOd and DI attack are equivalent per sample up to an multiplicative constant.*

We defer the proof to appendix. Corollary 1 shows the equivalence between adversarial attack against different group fairness notions, where attack targeting at group fairness perturb the predicted soft labels against sensitive attributes. Specifically, in the context of gradient-based attack, we have EOd and DI attack equivalent.

Next, we take EOd attack as an example to demonstrate the behavior of fairness attack in different sensitive groups. Without loss of generality, assume $a = 1$ is the advantaged group, and assume the average positive predictions are biased towards the advantaged group, that is, the average positive prediction is greater in the advantaged group:

$$\sum_{x_i \in \mathbb{S}_{y0}} \frac{f(x_i)}{|\mathbb{S}_{y0}|} \leq \sum_{x_i \in \mathbb{S}_{y1}} \frac{f(x_i)}{|\mathbb{S}_{y1}|}, \; y \in \{0, 1\}. \tag{3}$$

Consider the adversarial EOd loss in equation 2. For a sample (x_j, y_j, a_j) in advantaged group (i.e., $a_j = 1$), based on equation 3 we can rewrite L_{EOd} as:

$$\begin{aligned} L_{EOd} &= \left| \sum_{x_i \in \mathbb{S}_{00}} \frac{f(x_i)}{|\mathbb{S}_{00}|} - \sum_{x_i \in \mathbb{S}_{01}} \frac{f(x_i)}{|\mathbb{S}_{01}|} \right| + \left| \sum_{x_i \in \mathbb{S}_{10}} \frac{f(x_i)}{|\mathbb{S}_{10}|} - \sum_{x_i \in \mathbb{S}_{11}} \frac{f(x_i)}{|\mathbb{S}_{11}|} \right| \\ &= \frac{f(x_j)}{|\mathbb{S}_{y_j 1}|} + \sum_{x_i \in \mathbb{S}_{y_j 1} \setminus \{x_j\}} \frac{f(x_i)}{|\mathbb{S}_{y_j 1}|} - \sum_{x_i \in \mathbb{S}_{y_j 0}} \frac{f(x_i)}{|\mathbb{S}_{y_j 0}|} + \left| \sum_{x_i \in \mathbb{S}_{\bar{y}_j 0}} \frac{f(x_i)}{|\mathbb{S}_{\bar{y}_j 0}|} - \sum_{x_i \in \mathbb{S}_{\bar{y}_j 1}} \frac{f(x_i)}{|\mathbb{S}_{\bar{y}_j 1}|} \right| \\ &= \frac{f(x_j)}{|\mathbb{S}_{y_j 1}|} + C_j, \end{aligned} \tag{4}$$

where $\bar{y}_j = |1 - y_j|$ and C_j is a constant w.r.t. x_j thus does not affect $\frac{\partial L_{EOd}}{\partial x_j}$. Based on equation 4 we can derive that the adversarial perturbation w.r.t. L_{EOd} is expected to *maximize* the predicted soft labels for samples in the advantaged group.

Similarly, for sample (x_k, y_k, a_k) in the disadvantaged group ($a_k = 0$), we have:

$$\begin{aligned} L_{EOd} &= \left| \sum_{x_i \in \mathbb{S}_{00}} \frac{f(x_i)}{|\mathbb{S}_{00}|} - \sum_{x_i \in \mathbb{S}_{01}} \frac{f(x_i)}{|\mathbb{S}_{01}|} \right| + \left| \sum_{x_i \in \mathbb{S}_{10}} \frac{f(x_i)}{|\mathbb{S}_{10}|} - \sum_{x_i \in \mathbb{S}_{11}} \frac{f(x_i)}{|\mathbb{S}_{11}|} \right| \\ &= -\frac{f(x_k)}{|\mathbb{S}_{y_k 0}|} + C_k, \end{aligned} \tag{5}$$

where C_k is a constant w.r.t. x_k thus does not affect $\frac{\partial L_{EOd}}{\partial x_k}$. And based on equation 5 we can derive that adversarial perturbation w.r.t. L_{EOd} is expected to *minimize* the predicted soft labels in disadvantaged group, which is the opposite from samples in advantage group.

4.2 Alignment between EOd and accuracy attack

We move on to discuss the connection between EOd and accuracy attack. Without loss of generality, assume the positive label is the favorable outcome for classification, we have the following corollary regarding the connection between fairness and accuracy attack:

Corollary 2. *The fairness adversarial attack and accuracy adversarial attack are in alignment regarding true negative (TN) and false positive (FP) samples in advantaged group and true positive (TP) and false negative (FN) samples in disadvantaged group.*

We defer the detailed proof to appendix. It is worth noticing that the fairness adversarial attack and accuracy adversarial attack operates towards the opposite direction for the other two pairs. That is, for TP samples in advantaged group and TN samples in disadvantaged group: as in equation 4 and equation 5, the adversarial attack regarding fairness is expected to maximize the predicted confidence for samples in the two sensitive groups (i.e., maximizing the predicted soft labels for TP samples in advantage group, and minimizing the predicted soft labels for TN samples in disadvantaged group), while adversarial attack regarding accuracy is expected to minimize the predicted confidence for the two sensitive groups. Similarly, for FN samples in the advantaged group and FP samples in the disadvantaged group: the fairness attack is expected to 'correct' the predicted soft labels, i.e., the predicted adversarial labels is expected to be in alignment with ground-truth labels, while accuracy attack is expected to exacerbate the error.

4.3 Alignment between EOd and accuracy robustness

We now discuss the alignment between robustness w.r.t. fairness and accuracy. Although the relationship between robustness w.r.t. the two metrics seem straightforward on certain groups, the relationships on other groups are not clear. Before we state the detailed relationship, we first state the assumption we need to prove the relationship:

Assumption 1. *The gradient of f w.r.t. input feature x is Lipschitz with constant K.*

Under Assumption 1, we have the following guarantee for robustness against EOd attack under adversarial training w.r.t. accuracy:

Theorem 1. *For classifier f under accuracy adversarial training, let $D(x) := |L(f(x_{adv}^{EOd}), y) - L(f(x), y)|$ be the change of cross-entropy loss for sample x under EOd attack, consider FN sample $x_{FN,1}$ in the advantaged group and TP sample $x_{TP,0}$ in the disadvantaged group, let $\delta_{TP,0}^{t-1}$ be the change of $x_{TP,0}$'s predicted label under ϵ-level accuracy attack at $(t-1)$-th iteration, let x^t be the adversarial sample obtained at $(t-1)$-th iteration during EOd attack, the difference of robustness against ϵ-level EOd attack with step size α and up to T iterations between the two samples are upper-bounded by the robustness of TP sample against accuracy attack up to an addictive constant:*

$$|D(x_{FN,1}) - D(x_{TP,0})| \leq \min_{x_{TP,0} \in \mathbb{S}_{10}} \alpha \sum_{t=1}^{T} \left[\frac{\sqrt{n} K d(x_{FN,1}^{t-1}, x_{TP,0}^{t-1})}{f(x_{FN,1}^{t-1})} + \left| \frac{f(x_{TP,0}^{t-1}) - f(x_{FN,1}^{t-1})}{f(x_{FN,1}^{t-1}) f(x_{TP,0}^{t-1})} \right| \delta_{TP,0}^{t-1} \right].$$

Detailed proof can be found in the appendix. As discussed in Section 4.2, fairness robustness of TP samples in the disadvantaged group benefit directly from adversarial training w.r.t. accuracy, while robustness of FN samples in the advantaged group are not clearly related with adversarial training w.r.t. accuracy. Instead, we compare robustness of disadvantaged TP sample and advantaged FN sample to provide robustness guarantee for FN samples against EOd attack. Besides, as fairness attack regarding disadvantaged TP samples and advantaged FN samples are towards the opposite directions, Theorem 1 also shows the benefit of robustness against accuracy attack to EOd attack, where both samples contribute to lower change of EOd under fairness attack. For a classifier f' under normal training, we have similar upper-bound, except that we now have $\delta_{\text{TP},0}^{'t-1} \geq \delta_{\text{TP},0}^{t-1}$, which indicates a looser upper-bound compared with classifiers under adversarial training. For marginal FN samples which are more vulnerable under EOd attack, we have their robustness bounded by marginal TP samples, and smaller $\delta_{\text{TP},0}$, or tighter bound indicates better robustness. In this way, classifiers under accuracy adversarial training also achieve improvement in fairness robustness. Similar inequality in Theorem 1 also holds for FP samples in the disadvantaged group and TN samples in the advantaged group:

Remark 1. *For FP sample $x_{FP,0}$ in the disadvantaged group and TN sample $x_{TN,1}$ in the advantaged group, we have similar inequality regarding the upper-bound of robustness difference:*

$$|D(x_{FP,0}) - D(x_{TN,1})| \leq \min_{x_{TN,1} \in \mathbb{S}_{01}} \alpha \sum_{t=1}^{T} \left[\frac{\sqrt{n} K d(x_{FP,0}^{t-1}, x_{TN,1}^{t-1})}{f(x_{FP,0}^{t-1})} + \left| \frac{f(x_{FP,0}^{t-1}) - f(x_{TN,1}^{t-1})}{f(x_{TN,1}^{t-1}) f(x_{FP,0}^{t-1})} \right| \delta_{TN,0}^{t-1} \right].$$

For the reversed direction, under Assumption 1, we have the following guarantee for robustness against accuracy attack under adversarial training w.r.t. EOd:

Theorem 2. *For classifier f under EOd adversarial training, let $F(x) := |f(x_{adv}^{ACC}) - f(x)|$ be the change of predicted soft label under ϵ-level accuracy attack, consider TP samples $x_{TP,1}$ in the advantaged group and its neighbor TP sample $x_{TP,0}$ in the disadvantaged group, let δ_{TP}^{0} be the change of soft label of $x_{TP,0}$ under ϵ-level EOd attack with step size α and up to T iterations, let x_t be the adversarial sample under accuracy attack at $(t-1)$-th iteration, the robustness of TP samples $x_{TP,1}$ in the advantaged group against ϵ-level accuracy attack are upper-bounded by robustness of $x_{TP,0}$ against ϵ-level EOd attack up to an addictive constant:*

$$F(x_{TP,0}) \leq \min_{x_{TP,0} \in \mathbb{S}_{10}} \delta_{TP}^{0} + \sum_{t=1}^{T} \sqrt{n} \alpha K d(x_{TP,0,t-1}, x_{TP,0,t-1}).$$

Here the adversarial attack against fairness and accuracy are in alignment regarding disadvantaged TP samples, which we use to upper-bound robustness of advantaged TP samples under accuracy attack. Theorem 2 shows that adversarial training w.r.t. fairness also benefits robustness w.r.t. accuracy. For a classifier f' under normal training, we have similar inequality except that we now have $\delta_{\text{TP}}^{0'} \geq \delta_{\text{TP}}^{0}$. Similar upper-bound also holds for TN samples:

Remark 2. *For TN sample $x_{TN,1}$ in the advantaged group and TN sample $x_{TN,0}$ in the disadvantaged group, we have similar inequality regarding the upper-bound of robustness against accuracy attack:*

$$F(x_{TN,1}) \leq \min_{x_{TN,1} \in \mathbb{S}_{01}} \delta_{TN}^{0} + \sum_{t=1}^{T} \sqrt{n} \alpha K d(x_{TN,1,t-1}, x_{TN,1,t-1}).$$

4.4 FAIR ADVERSARIAL TRAINING

One direct result regarding Theorem 1 is to incorporate adversarial samples w.r.t. accuracy attack during training to obtain a fair classifier that is also robust to adversarial fairness perturbations. Consider the relaxed EOd loss under fairness perturbation:

$$\begin{aligned} L'_{EOd} &= \left| \sum_{x_i \in \mathbb{S}_{00}} \frac{(f(x_i) - \delta_i)}{|\mathbb{S}_{00}|} - \sum_{x_i \in \mathbb{S}_{01}} \frac{(f(x_i) + \delta_i)}{|\mathbb{S}_{01}|} \right| + \left| \sum_{x_i \in \mathbb{S}_{10}} \frac{(f(x_i) - \delta_i)}{|\mathbb{S}_{10}|} - \sum_{x_i \in \mathbb{S}_{11}} \frac{(f(x_i) + \delta_i)}{|\mathbb{S}_{11}|} \right| \\ &\leq \left| \sum_{x_i \in \mathbb{S}_{00}} \frac{f(x_i)}{|\mathbb{S}_{00}|} - \sum_{x_i \in \mathbb{S}_{01}} \frac{f(x_i)}{|\mathbb{S}_{01}|} \right| + \left| \sum_{x_i \in \mathbb{S}_{10}} \frac{f(x_i)}{|\mathbb{S}_{10}|} - \sum_{x_i \in \mathbb{S}_{11}} \frac{f(x_i)}{|\mathbb{S}_{11}|} \right| \\ &+ \sum_{x_i \in \mathbb{S}_{00}} \frac{\delta_i}{|\mathbb{S}_{00}|} + \sum_{x_i \in \mathbb{S}_{01}} \frac{\delta_i}{|\mathbb{S}_{01}|} + \sum_{x_i \in \mathbb{S}_{10}} \frac{\delta_i}{|\mathbb{S}_{10}|} + \sum_{x_i \in \mathbb{S}_{11}} \frac{\delta_i}{|\mathbb{S}_{11}|} \\ &= L_{EOd} + \sum_{x_i \in \mathbb{S}_{00}} \frac{\delta_i}{|\mathbb{S}_{00}|} + \sum_{x_i \in \mathbb{S}_{01}} \frac{\delta_i}{|\mathbb{S}_{01}|} + \sum_{x_i \in \mathbb{S}_{10}} \frac{\delta_i}{|\mathbb{S}_{10}|} + \sum_{x_i \in \mathbb{S}_{11}} \frac{\delta_i}{|\mathbb{S}_{11}|}, \end{aligned} \tag{6}$$

where the relaxed EOd loss under adverasrial perturbation is upper-bounded by the relaxed EOd loss without adversarial perturbation and robustness of samples against EOd adversarial attack. A direct implication of this formulation is to improve robustness of classifier w.r.t. fairness by incorporating adversarial samples and fairness constraints during training:

$$\arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L_{CE}(f(x_{\text{adv},i}^{\text{ACC}}), y_i), \ s.t. \ L \leq \gamma, \tag{7}$$

where L can be specified by fairness relaxations as regularization during training, or can be implicitly specified as preprocessing or postprocessing techniques.

5 EXPERIMENTS

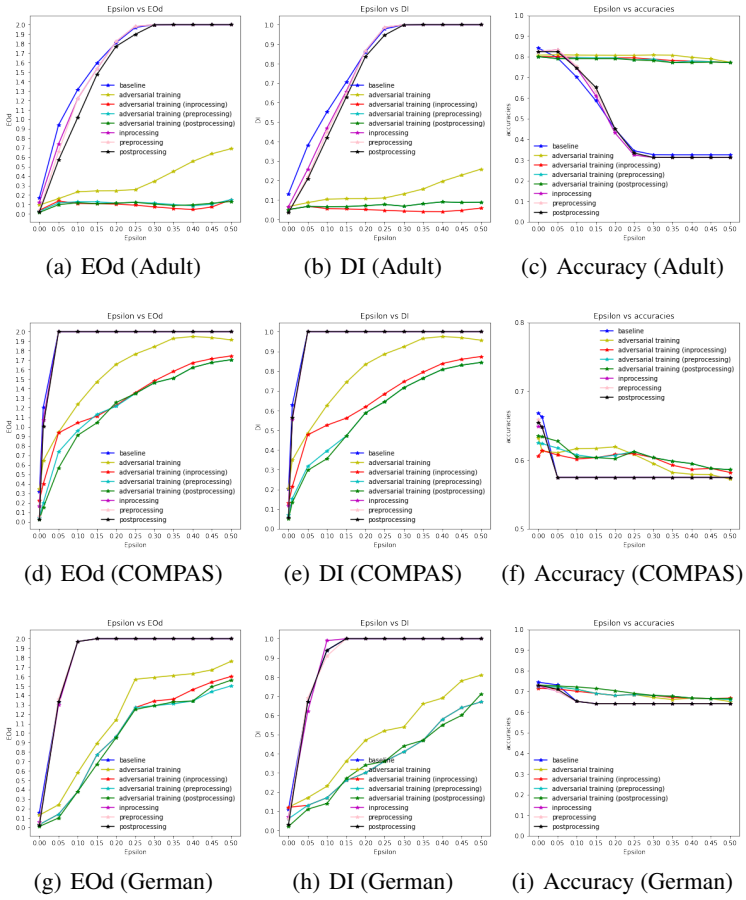
We evaluate our method on three datasets: Adult (Dua & Graff, 2017), COMPAS (Larson et al., 2016) and German (Dua & Graff, 2017). The sensitive attributes are chosen as *race* for Adult and COMPAS and *sex* for German. Details of the datasets are in the Appendix.

We use accuracy as performance evaluation, and disparate impact (DI) and equalized odds (EOd) as fairness metric. The classifier for all compared methods is chosen as MLP, and all methods are trained under the same data partition. During adversarial training, the perturbation level is set as 0.2 for Adult dataset, 0.005 for COMPAS dataset and 0.01 for German dataset, where the the perturbation level is empirically determined to achieve largest perturbation while still ensuring convergence.

5.1 ROBUSTNESS AGAINST EOD ATTACK

We compare five different methods with our fair adversarial training method. Specifically, we consider three different versions for our fair adversarial training method (preprocessing, inprocessing and postprocessing). The three versions differ in the fairness regularization L in equation 7.

- Baseline: MLP model under normal training.
- Preprocessing: MLP model under normal training with label processed by Jiang & Nachum (2020).
- Postprocessing: MLP model under normal training with postprocessing technique by Jang et al. (2022).
- Inprocessing: MLP model under normal training with relaxed EOd constraint by Wang et al. (2022).
- Adversarial training: MLP model under adversarial training w.r.t. accuracy.
- Adversarial training (preprocessing): MLP model under adversarial training w.r.t. accuracy with training label processed by Jiang & Nachum (2020).



(a) EOd (Adult) (b) DI (Adult) (c) Accuracy (Adult)

(d) EOd (COMPAS) (e) DI (COMPAS) (f) Accuracy (COMPAS)

(g) EOd (German) (h) DI (German) (i) Accuracy (German)

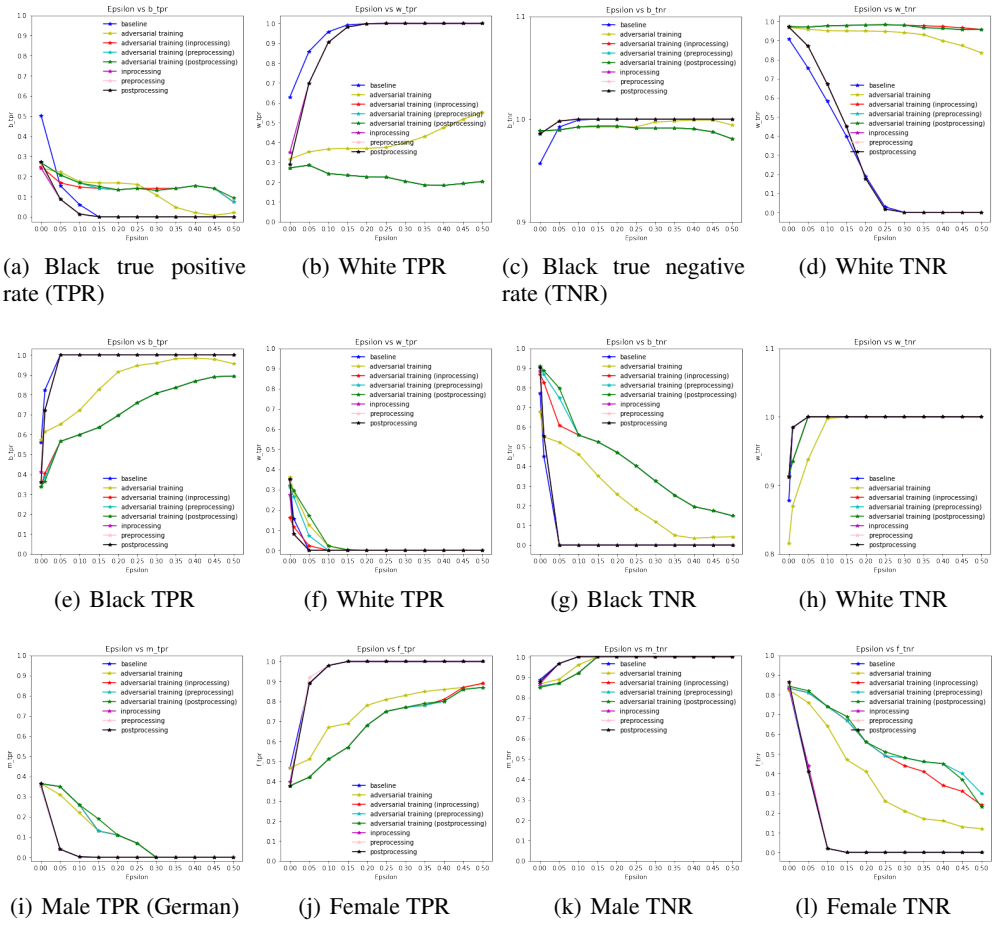
Figure 1: Change in accuracy, DI and EOd under EOd attack on three datasets. Our adversarial training methods (preprocessing, inprocessing, postprocessing) obtain improved fairness (lower EOd and DI) and accuracy with significant margin.

- Adversarial training (inprocessing): MLP model under adversarial training w.r.t. accuracy with relaxed Eod constraint by Wang et al. (2022).
- Adversarial training (postprocessing): MLP model under adversarial training w.r.t. accuracy with predicted label postprocessed by Jang et al. (2022).

Results on classifiers under EOd attack are shown in Fig. 1 - 2. The EOd attack enforces biased predictions against testing samples based on the sensitive information, and under a successful attack (the EOd reaches its maximum), the DI also reaches its maximum, while the accuracy under adversarial attack is determined by the base rate of sensitive groups. Compared with methods under adversarial training, methods under normal training show a sharp increase in DI and EOd under adversarial attack, and improvement in fairness of classifiers under normal training do not help with the robustness under adversarial perturbation. In comparison, classifiers under adversarial training w.r.t. accuracy show improvement in terms of robustness against fairness attack, and classifiers under fair adversarial training show further remarkable improvement in terms of robustness against fairness attack [2]. This shows that improving robustness against accuracy attack also improves robustness against fairness attack and is in line with our discussion in equation 6.

Furthermore, in Fig. 3 show the effect of perturbation levels during training on Adult dataset, where the results show that larger perturbation level during training indicates better robustness against

[2]We defer the detailed values for fair adversarial training in Appendix, as part of the fair adversarial training results are overlapped with each other.



(a) Black true positive rate (TPR) (b) White TPR (c) Black true negative rate (TNR) (d) White TNR

(e) Black TPR (f) White TPR (g) Black TNR (h) White TNR

(i) Male TPR (German) (j) Female TPR (k) Male TNR (l) Female TNR

Figure 2: Change of TPR and TNR under EOd attack on three datasets.

fairness adversarial attack for both vanilla adversarial training and fair adversarial training during testing. We defer full results of varying training perturbation levels to appendix.

5.2 ROBUSTNESS AGAINST ACCURACY ATTACK

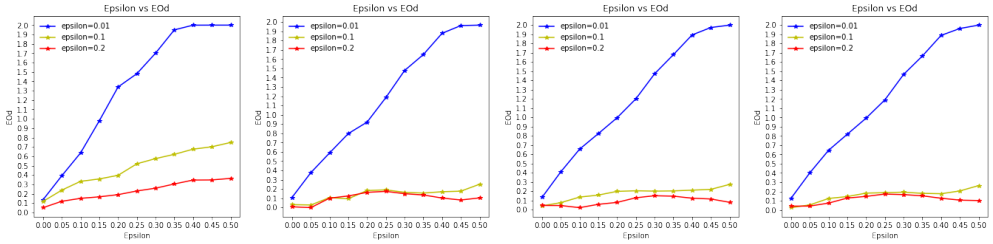
We move on to discuss the improvement of robustness w.r.t. accuracy under adversarial training w.r.t. fairness. We compare two different methods:

- baseline: MLP model under normal training.
- Adversarial training (EOd): MLP model under adversarial training w.r.t. relaxed EOd.

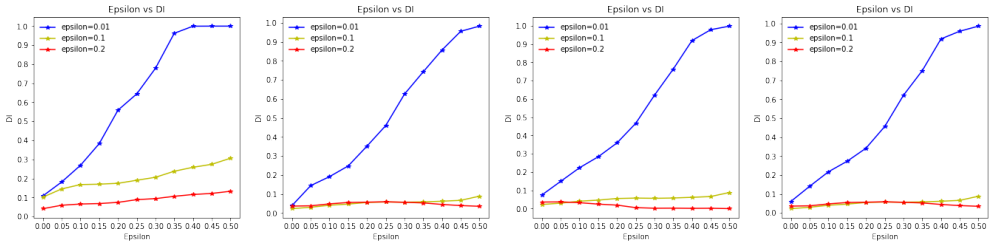
We show results on classifiers under accuracy attack on Adult dataset in Fig. 4. Under a successful accuracy attack (the accuracy reaches its minimum), the EOd also becomes zero, while DI does not necessarily vanishes due to distributional disparities across sensitive groups. Compared with baseline classifier, classifier under adversarial training w.r.t. EOd shows remarkable improvement in robustness against accuracy attack. This shows that robustness against accuracy attack also benefits from adversarial training against fairness. Results on other datasets are shown in the appendix.

6 CONCLUSION

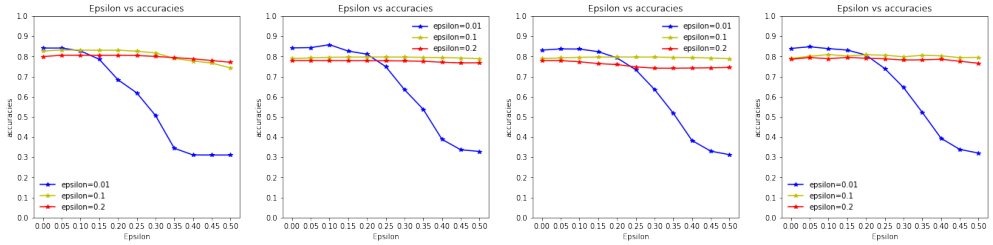
Fairness attack and fairness adversarial training is an important yet not properly addressed problem. In this paper, we propose a unified framework for fairness attack against group fairness notions,



(a) EOd (adv) (b) EOd (adv+pre) (c) EOd (adv+in) (d) EOd (adv+post)

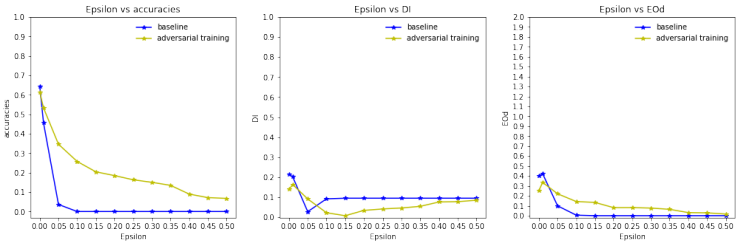


(e) DI (adv) (f) DI (adv+pre) (g) DI (adv+in) (h) DI (adv+post)



(i) Acr (adv) (j) Acr (adv+pre) (k) Acr (adv+in) (l) Acr (adv+post)

Figure 3: Change of accuracy and EOd under EOd attack with varying training perturbation ϵ on Adult dataset.



(a) Accuracy (b) DI (c) EOd

Figure 4: Results of a classifier adversarial trained w.r.t. EOd. Change of accuracy, DI and EOd under accuracy attack on Adult dataset.

where we show theoretically the alignment of attack against different notions, and we demonstrate the connections between fairness attack and conventional accuracy attack. We show theoretically the alignment between accuracy robustness and fairness robustness, and we propose a fair adversarial training structure, where the goal is to improve adversarial robustness w.r.t. accuracy while ensuring fairness. We show from experiments that our method achieves better robustness under fairness adversarial attack, and we show from experiments the alignment between robustness w.r.t. fairness and accuracy. Future directions include finding alternative relaxations for fairness attack, and alternative training strategies for fair adversarial training.

REFERENCES

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15721–15730, 2021.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Taeuk Jang, Feng Zheng, and Xiaoqian Wang. Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7908–7916, 2021.

Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6988–6995, 2022.

Xiaojun Jia, Yong Zhang, Baoyuan Wu, Jue Wang, and Xiaochun Cao. Boosting fast adversarial training with learnable adversarial initialization. *IEEE Transactions on Image Processing*, 31: 4417–4430, 2022.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. Verifying individual fairness in machine learning models. In *Conference on Uncertainty in Artificial Intelligence*, pp. 749–758. PMLR, 2020.

Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12115–12124, 2021.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439*, 2019.

J Larson, S Mattu, L Kirchner, and J Angwin. Compas analysis. *GitHub, available at: https://github.com/propublica/compas-analysis*, 2016.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021a.

Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8930–8938, 2021b.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pp. 7097–7107. PMLR, 2020.

Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems*, 34:815–827, 2021.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 162–177. Springer, 2020.

Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding instance-level impact of fairness constraints. In *International Conference on Machine Learning*, pp. 23114–23130. PMLR, 2022.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.