

## A APPENDIX

### A.1 DATASETS

Adult (Dua & Graff, 2017): The Adult dataset contains 65,123 samples with 14 attributes. The goal is to predict whether an individual’s annual income exceeds 50K, and the sensitive attribute is chosen as *race*.

COMPAS (Larson et al., 2016): The ProPublica COMPAS dataset contains 7,215 samples with 10 attributes. The goal is to predict whether a defendant re-offend within two years. Following the protocol in earlier fairness methods Zafar et al. (2017), we only select white and black individuals in COMPAS dataset, which contains 6,150 samples in total. The sensitive attribute in this dataset is *race*.

German (Dua & Graff, 2017): The German credit risk dataset contains 1,000 samples with 9 attributes. The goal is to predict whether a client is highly risky, and the sensitive attribute in this dataset is *sex*.

### A.2 FULL RESULTS ON VARYING $\epsilon$

Results of varying  $\epsilon$  on COMPAS and German dataset can be found in Fig. 5, 6. As shown in the figures, larger perturbation levels result in classifiers that are more robust to adversarial perturbations.

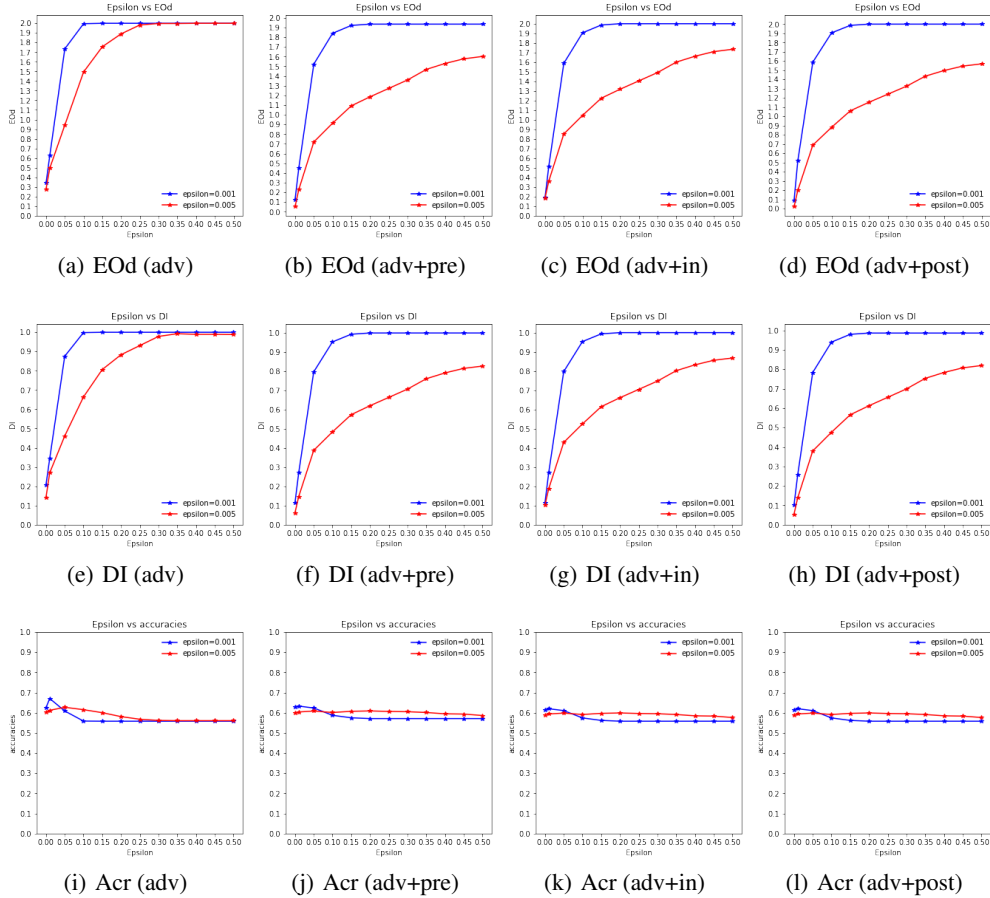


Figure 5: Change of accuracy and EOD under EOD attack with varying training perturbation  $\epsilon$  on German dataset.

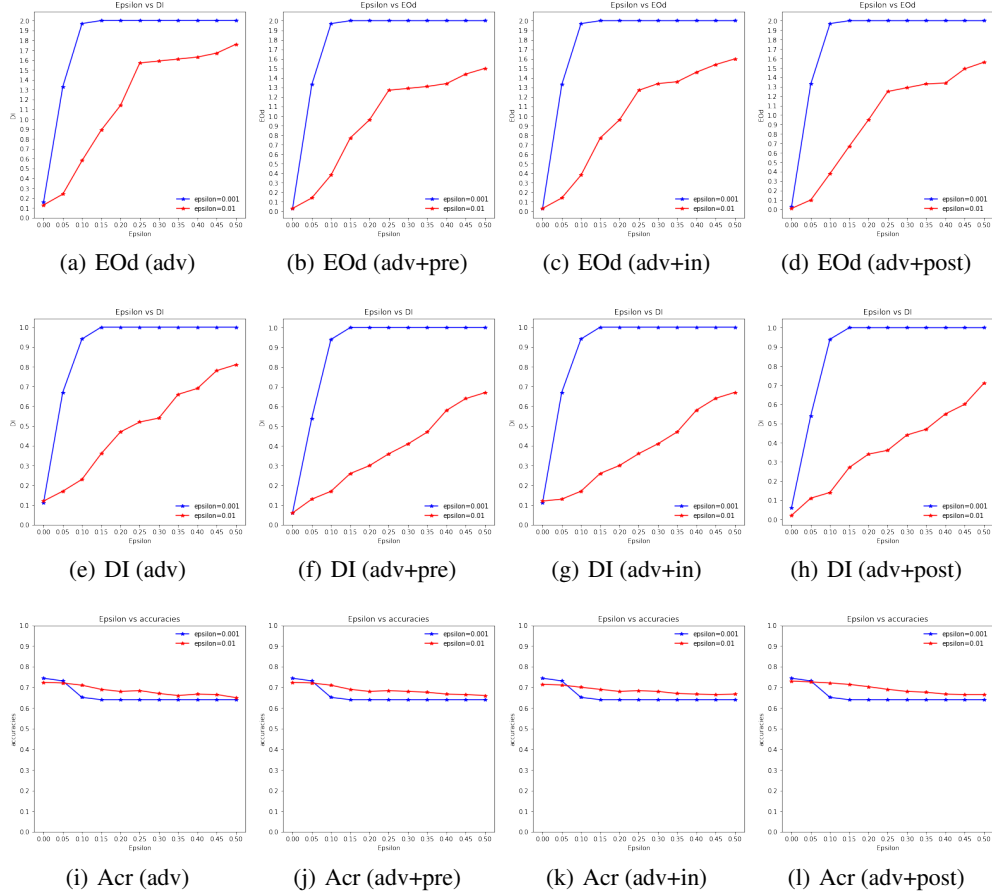


Figure 6: Change of accuracy and EOd under EOd attack with varying training perturbation  $\epsilon$  on German dataset.

### A.3 PROOF OF COROLLARY 11

*Proof.* The objective for EOd attack can be written as the following form:

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

Without loss of generality, assume  $a = 1$  the advantaged group, we have

$$\begin{aligned} L_{EOd} &= \sum_{x_i \in \mathbb{S}_{01}} \frac{f(x_i)}{|\mathbb{S}_{01}|} - \sum_{x_i \in \mathbb{S}_{00}} \frac{f(x_i)}{|\mathbb{S}_{00}|} + \sum_{x_i \in \mathbb{S}_{11}} \frac{f(x_i)}{|\mathbb{S}_{11}|} - \sum_{x_i \in \mathbb{S}_{10}} \frac{f(x_i)}{|\mathbb{S}_{10}|} \\ &= \sum_{x_i \in \mathbb{S}_{01}} \frac{|\mathbb{S}_{01}| + |\mathbb{S}_{11}|}{|\mathbb{S}_{01}|} \frac{f(x_i)}{|\mathbb{S}_{01}| + |\mathbb{S}_{11}|} + \sum_{x_i \in \mathbb{S}_{11}} \frac{|\mathbb{S}_{01}| + |\mathbb{S}_{11}|}{|\mathbb{S}_{11}|} \frac{f(x_i)}{|\mathbb{S}_{01}| + |\mathbb{S}_{11}|} \\ &\quad - \sum_{x_i \in \mathbb{S}_{00}} \frac{|\mathbb{S}_{10}| + |\mathbb{S}_{00}|}{|\mathbb{S}_{00}|} \frac{f(x_i)}{|\mathbb{S}_{10}| + |\mathbb{S}_{00}|} - \sum_{x_i \in \mathbb{S}_{10}} \frac{|\mathbb{S}_{10}| + |\mathbb{S}_{00}|}{|\mathbb{S}_{10}|} \frac{f(x_i)}{|\mathbb{S}_{10}| + |\mathbb{S}_{00}|} \\ &= \sum_{x_i \in \mathbb{S}_{01}} \frac{|\mathbb{S}_{.1}|}{|\mathbb{S}_{01}|} \frac{f(x_i)}{|\mathbb{S}_{.1}|} + \sum_{x_i \in \mathbb{S}_{11}} \frac{|\mathbb{S}_{.1}|}{|\mathbb{S}_{11}|} \frac{f(x_i)}{|\mathbb{S}_{.1}|} - \sum_{x_i \in \mathbb{S}_{00}} \frac{|\mathbb{S}_{.0}|}{|\mathbb{S}_{00}|} \frac{f(x_i)}{|\mathbb{S}_{.0}|} - \sum_{x_i \in \mathbb{S}_{10}} \frac{|\mathbb{S}_{.0}|}{|\mathbb{S}_{10}|} \frac{f(x_i)}{|\mathbb{S}_{.0}|}, \end{aligned} \tag{8}$$

where  $\mathbb{S}_k$  refers to the set of testing samples with  $a_i = k$ . Similarly, for DI attack, we have the following objective:

$$\begin{aligned} L_{DI} &= \left| \sum_{x_i \in \mathbb{S}_{.0}} \frac{f(x_i)}{|\mathbb{S}_{.0}|} - \sum_{x_i \in \mathbb{S}_{.1}} \frac{f(x_i)}{|\mathbb{S}_{.1}|} \right| \\ &= \sum_{x_i \in \mathbb{S}_{01}} \frac{f(x_i)}{|\mathbb{S}_{.1}|} + \sum_{x_i \in \mathbb{S}_{11}} \frac{f(x_i)}{|\mathbb{S}_{.1}|} - \sum_{x_i \in \mathbb{S}_{00}} \frac{f(x_i)}{|\mathbb{S}_{.0}|} - \sum_{x_i \in \mathbb{S}_{10}} \frac{f(x_i)}{|\mathbb{S}_{.0}|}. \end{aligned}$$

For  $i$ -th sample, the adversarial objective between EOd and DI only differs by a constant  $\frac{|\mathbb{S}_{.a_i}|}{|\mathbb{S}_{y_i a_i}|}$ , and the constant value is determined by the base rate of data. Specifically, in terms of gradient-based attack, the adversarial objective of EOd and DI are equivalent.  $\square$

#### A.4 PROOF OF COROLLARY 2

*Proof.* Without loss of generality, assume the positive label is the favorable outcome for classification, the objective for accuracy attack for sample  $x_i$  can be written as

$$\max_{\delta} L((x_i + \epsilon), y_i), \|\epsilon\| \leq \epsilon_0, \quad (9)$$

Consider the EOd attack in equation 8, we have the objective for EOd attack as follows:

$$\max_{\delta} \alpha_i \frac{|\mathbb{S}_{.a_i}|}{|\mathbb{S}_{y_i a_i}|} \frac{f(x_i + \epsilon)}{|\mathbb{S}_{.a_i}|}, \|\epsilon\| \leq \epsilon',$$

where  $\alpha_i = -1$  for  $a_i = 0$  and  $\alpha_i = 1$  for  $a_i = 1$ . For positive samples, we can further write equation 9 as

$$\max_{\delta} -\log(f(x_i + \epsilon)), \|\epsilon\| \leq \epsilon_0,$$

where the perturbation is expected to minimize the predicted soft label, which is in alignment with the objective for EOd when  $\alpha_i = -1$ , i.e., for TP and FN disadvantaged samples, the two attack are in alignment. Similarly, for negative samples, we have equation 9 as

$$\max_{\delta} -\log(1 - f(x + \epsilon)), \|\epsilon\| \leq \epsilon',$$

where the perturbation is expected to maximize the predicted soft label, which is in alignment with the objective for EOd when  $\alpha_i = 1$ , i.e., for TN and FP advantaged samples, the two attack are in alignment.. Specifically, for gradient-based attack, we have the two kinds of attack equivalent.  $\square$

#### A.5 PROOF OF THEOREM 1

*Proof.* Let  $f$  be the function of classifier under adversarial training w.r.t. accuracy, consider the positive testing set  $\{(x_i, 1, a_i), 1 \leq i \leq N\}$  for simplicity, at  $t - 1$ -th iteration, we have the linear approximation of testing CE loss under EOd attack as follows:

$$L_{CE}(x^t) = -\log(x^t) = -\log(f(x^{t-1}) - \delta^{t-1}) = -\log(f(x^{t-1})) + \frac{\delta^{t-1}}{f(x^{t-1})} + r_L(x^{t-1}), \quad (10)$$

where  $\delta^{t-1}$  is the change of soft label induced by EOd attack at  $t - 1$ -th iteration, and  $r_L(x)$  is the remainder of Taylor's expansion. For gradient-based attack, the predicted soft label for adversarial sample can be formulated as

$$\begin{aligned} &f(x^t) \\ &= f(x^{t-1} + \alpha \text{sign}(\nabla_{x^{t-1}} L_{EOd})) \\ &= f(x^{t-1}) + \alpha (\nabla_{x^{t-1}} f(x^{t-1}))^T \text{sign}(\nabla_{x^{t-1}} L_{EOd}) + r_f(x^{t-1}), \end{aligned} \quad (11)$$

where  $\epsilon$  is the magnitude of perturbation,  $L_{EOd}$  is the relaxed EOd loss, and  $r_f(x)$  is the remainder of Taylor's expansion. Let  $D(x^t) := |L(x^t) - L(x^{t-1})|$  be the change of CE loss under EOd attack

at  $t$ -th iteration, according to equation [10](#) and equation [11](#) we have

$$\begin{aligned} D(x^t) &= |L_{CE}(x^t) - L_{CE}(x^{t-1})| \\ &= |-\log(f(x^{t-1})) + \frac{\delta^{t-1}}{f(x^{t-1})} + r_L(x) + \log(f(x))| \\ &\approx \frac{|\alpha(\nabla_{x^{t-1}} f(x^{t-1}))^T \text{sign}(\nabla_{x^{t-1}} L_{EOd})|}{f(x^{t-1})}. \end{aligned}$$

Consider marginal TP sample  $x_{TP,0}$  from disadvantaged group and marginal FN sample  $x_{FN,1}$  from advantaged group, since the gradient of  $f$  w.r.t.  $x$  is Lipschitz with constant  $K$ , we have the difference of change in CE loss under EOD attack at  $t$ -th iteration as follows:

$$\begin{aligned} &|D(x_{FN,1}^t) - D(x_{TP,0}^t)| \\ &= \alpha \left| \frac{|(\nabla_{x_{FN,1}^{t-1}} f(x_{FN,1}^{t-1}))^T \text{sign}(\nabla_{x_{FN,1}^{t-1}} L_f)|}{f(x_{FN,1}^{t-1})} - \frac{|(\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1}))^T \text{sign}(\nabla_{x_{TP,0}^{t-1}} L_f)|}{f(x_{TP,0}^{t-1})} \right| \\ &= \alpha \left| \frac{(\nabla_{x_{FN,1}^{t-1}} f(x_{FN,1}^{t-1}))^T \text{sign}(\nabla_{x_{FN,1}^{t-1}} L_f)}{f(x_{FN,1}^{t-1})} + \frac{(\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1}))^T \text{sign}(\nabla_{x_{TP,0}^{t-1}} L_f)}{f(x_{TP,0}^{t-1})} \right| \\ &= \alpha \left| \frac{(\nabla_{x_{FN,1}^{t-1}} f(x_{FN,1}^{t-1}))^T \text{sign}(\frac{1}{N_1} \nabla_{x_{FN,1}^{t-1}} f(x_{FN,1}^{t-1}))}{f_\theta(x_{FN,1}^{t-1})} - \frac{(\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1}))^T \text{sign}(\frac{1}{N_0} \nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1}))}{f(x_{TP,0}^{t-1})} \right| \\ &= \alpha \left| \frac{\sum_{j=1}^n |\partial_{x_j} f(x_{FN,1}^{t-1})|}{f(x_{FN,1}^{t-1})} - \frac{\sum_{j=1}^n |\partial_{x_j} f(x_{TP,0}^{t-1})|}{f(x_{TP,0}^{t-1})} \right| \\ &= \alpha \left| \frac{\|\nabla_{x_{FN,1}^{t-1}} f(x_{FN,1}^{t-1})\|_1}{f(x_{FN,1}^{t-1})} - \frac{\|\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1})\|_1}{f(x_{TP,0}^{t-1})} \right|, \end{aligned} \tag{12}$$

where  $n$  is the dimension of input feature. Since  $\nabla_x f(x)$  is Lipschitz, we have

$$\|\nabla_x f(x_1)\|_2 - \|\nabla_x f(x_0)\|_2 \leq \|\nabla_x f(x_1) - \nabla_x f(x_0)\|_2 \leq Kd(x_1, x_0),$$

where the first sign is due to triangle inequality. By Jensen's inequality we have  $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$ , and

$$\|\nabla_x f(x_1)\|_1 - \|\nabla_x f(x_0)\|_1 \leq \|\nabla_x f(x_0) - \nabla_x f(x_1)\|_1 \leq \sqrt{n}Kd(x_1, x_0). \tag{13}$$

Assume  $\frac{\|\nabla_{x_{FN,1}^{t-1}} f(x_{FN,1}^{t-1})\|_1}{f(x_{FN,1}^{t-1})} \geq \frac{\|\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1})\|_1}{f(x_{TP,0}^{t-1})}$ , plugging equation [13](#) back into equation [12](#), we have

$$\begin{aligned} &|D(x_{FN,1}^{t-1}) - D(x_{TP,0}^{t-1})| \\ &= \alpha \left| \frac{\|\nabla_{x_{FN,1}^{t-1}} f(x_{FN,1}^{t-1})\|_1}{f(x_{FN,1}^{t-1})} - \frac{\|\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1})\|_1}{f(x_{TP,0}^{t-1})} \right| \\ &\leq \alpha \left| \frac{\sqrt{n}Kd(x_{FN,1}^{t-1}, x_{TP,0}^{t-1}) + \|\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1})\|_1}{f(x_{FN,1}^{t-1})} - \frac{\|\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1})\|_1}{f(x_{TP,0}^{t-1})} \right| \\ &\leq \frac{\sqrt{n}\alpha Kd(x_{FN,1}^{t-1}, x_{TP,0}^{t-1})}{f(x_{FN,1}^{t-1})} + \left| \frac{\alpha \|\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1})\|_1}{f(x_{FN,1}^{t-1})} - \frac{\alpha \|\nabla_{x_{TP,0}^{t-1}} f(x_{TP,0}^{t-1})\|_1}{f(x_{TP,0}^{t-1})} \right|, \end{aligned} \tag{14}$$

where  $d(x, y) := \|x - y\|_2$  is the distance between the two feature. Taking the summation over  $T$  iterations, we have

$$|D(x_{FN,1}) - D(x_{TP,0})| \leq \sum_{t=1}^T \left[ \frac{\sqrt{n}\alpha Kd(x_{FN,1}^{t-1}, x_{TP,0}^{t-1})}{f(x_{FN,1}^{t-1})} + \alpha \left| \frac{f(x_{TP,0}^{t-1}) - f(x_{FN,1}^{t-1})}{f(x_{FN,1}^{t-1})f(x_{TP,0}^{t-1})} \right| \delta_{TP,0}^{t-1} \right]. \tag{15}$$

Since the above inequality holds true for all disadvantaged TP samples, we can further write equation [15](#) as

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

where  $\delta_{\text{TP},0}^{t-1} := \delta \|\nabla_{x_{\text{TP},0}^{t-1}} f_{\theta}(x_{\text{TP},0}^{t-1})\|_1$  is the change of  $x_{\text{TP},0}$ 's predicted label under  $\epsilon$ -level accuracy attack at  $t-1$ -th iteration. This shows that under EOd attack, the difference of change in performance regarding marginal TP samples and marginal FN samples are upper-bounded by the robustness of marginal TP samples up to an additive constant. For  $f_{\theta'}$  that is under normal training, we have similar upper-bound except that we now have  $\delta_{\text{TP},0}^{\prime t-1} \geq \delta_{\text{TP},0}^{t-1}$ , which indicates that the adversarial

classifier achieves tighter upper-bound than that of a normal classifier. For  $\frac{\|\nabla_{x_{\text{FN},1}^{t-1}} f(x_{\text{FN},1}^{t-1})\|_1}{f(x_{\text{FN},1}^{t-1})} \leq \frac{\|\nabla_{x_{\text{TP},0}^{t-1}} f(x_{\text{TP},0}^{t-1})\|_1}{f(x_{\text{TP},0}^{t-1})}$ , we have same upper-bound:

$$\begin{aligned} & |D(x_{\text{FN},1}^t) - D(x_{\text{TP},0}^t)| \\ &= \alpha \left| \frac{\|\nabla_{x_{\text{FN},1}^{t-1}} f(x_{\text{FN},1}^{t-1})\|_1}{f(x_{\text{FN},1}^{t-1})} - \frac{\|\nabla_{x_{\text{TP},0}^{t-1}} f(x_{\text{TP},0}^{t-1})\|_1}{f(x_{\text{TP},0}^{t-1})} \right| \\ &\leq \alpha \left| \frac{\|\nabla_{x_{\text{TP},0}^{t-1}} f(x_{\text{TP},0}^{t-1})\|_1}{f(x_{\text{TP},0}^{t-1})} - \frac{\|\nabla_{x_{\text{TP},0}^{t-1}} f(x_{\text{TP},0}^{t-1})\|_1 - \sqrt{n}\alpha K d(x_{\text{FN},1}^{t-1}, x_{\text{TP},0}^{t-1})}{f(x_{\text{TP},0}^{t-1})} \right| \\ &\leq \frac{\sqrt{n}\alpha K d(x_{\text{FN},1}^{t-1}, x_{\text{TP},0}^{t-1})}{f(x_{\text{FN},1}^{t-1})} + \left| \frac{\alpha \|\nabla_{x_{\text{TP},0}^{t-1}} f(x_{\text{TP},0}^{t-1})\|_1}{f(x_{\text{FN},1}^{t-1})} - \frac{\alpha \|\nabla_{x_{\text{TP},0}^{t-1}} f(x_{\text{TP},0}^{t-1})\|_1}{f(x_{\text{TP},0}^{t-1})} \right|. \end{aligned}$$

□

## A.6 PROOF OF THEOREM [2](#)

*Proof.* Let  $f$  be the function of classifier under adversarial training w.r.t. EOd, consider TP sample  $x_{\text{TP},0}$  in the disadvantaged group, we have the predicted soft label for sample  $x_{\text{TP},0}$  under accuracy attack at  $t-1$ -th iteration as follows:

$$\begin{aligned} & f(x_{\text{TP},0,t}) \\ &= f(x_{\text{TP},0,t-1} + \alpha \text{sign}(\nabla_{x_{\text{TP},0,t-1}} L)) \\ &\approx f(x_{\text{TP},0,t-1}) + \alpha (\nabla_{x_{\text{TP},0,t-1}} f(x_{\text{TP},0,t-1}))^T \text{sign}(-\frac{1}{f(x_{\text{TP},0,t-1})} \nabla_{x_{\text{TP},0,t-1}} f(x_{\text{TP},0,t-1})) \\ &= f(x_{\text{TP},0,t-1}) + \alpha (\nabla_x f(x_{\text{TP},0,t-1}))^T \text{sign}(\nabla_{x_{\text{TP},0,t-1}} L) \\ &= f(x_{\text{TP},0,t-1}) - \alpha \|\nabla_{x_{\text{TP},0,t-1}} f(x_{\text{TP},0,t-1})\|_1 \\ &= f(x_{\text{TP},0,t-1}) - \delta_{\text{TP},t-1}^0, \end{aligned}$$

where  $\delta_{\text{TP},t-1}^0 := \alpha \|\nabla_{x_{\text{TP},0,t-1}} f(x_{\text{TP},0,t-1})\|_1$  is the change of  $x_{\text{TP},0}$ 's predicted label under  $\epsilon$ -level EOd attack at  $t-1$ -th iteration. This shows that disadvantaged TP samples that attains  $\delta$ -level robustness under  $\epsilon$ -level EOd attack also attains similar robustness w.r.t. accuracy attack.

For TP sample  $x_{\text{TP},1}$  in the advantaged group, let  $F(x_{\text{TP},1,t}) = |f(x_{\text{TP},1,t}) - f(x_{\text{TP},1,t-1})|$ , we have its predicted soft label under accuracy attack at  $t-1$ -th iteration as follows:

$$\begin{aligned} & F(x_{\text{TP},1,t}) \\ &= |f(x_{\text{TP},1,t}) - f(x_{\text{TP},1,t-1})| \\ &= |f(x_{\text{TP},1,t-1} + \alpha \text{sign}(\nabla_{x_{\text{TP},1,t-1}} L)) - f(x_{\text{TP},1,t-1})| \\ &\approx \alpha (\nabla_{x_{\text{TP},1,t-1}} f(x_{\text{TP},1,t-1}))^T \text{sign}(\nabla_{x_{\text{TP},1,t-1}} L_{CE}) \\ &= \alpha \|\nabla_{x_{\text{TP},1,t-1}} f(x_{\text{TP},1,t-1})\|_1 \\ &\leq \delta_{\text{TP},t-1}^0 + \sqrt{n}\alpha K d(x_{\text{TP},1,t-1}, x_{\text{TP},1,t-1}). \end{aligned} \tag{16}$$

Taking the summation over all iterations, we have

$$F(x_{\text{TP},1}) \leq \delta_{\text{TP}}^0 + \sum_{t=1}^T \sqrt{n\alpha} K d(x_{\text{TP},0,t-1}, x_{\text{TP},1,t-1}), \quad (17)$$

where  $\delta_{\text{TP}}^0$  is the change of predicted soft label of sample  $x_{\text{TP},0}$  under  $\epsilon$ -level PGD attack. Since the inequality hold true for all  $x_{\text{TP},0}$ , we can further write equation 17 as

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

And the lower bound  $F(x_{\text{TP},1}) \geq 0$  naturally holds true for samples under accuracy attack. This shows that for samples in the advantaged group, the change of predicted soft label under accuracy attack is lower-bounded by the robustness of its neighbor sample(s) in the disadvantaged group up to an additive constant. For  $f'$  that is under normal training, we have similar upper-bound except that we now have  $\delta_{\text{TP}}^0 \geq \delta_{\text{TP}}^0$ , which indicates that the adversarial classifier achieves tighter upper-bound than that of a normal classifier.  $\square$

#### A.7 RESULTS OF ROBUSTNESS AGAINST EOD ATTACK

We include the results of fair adversarial training in Tab. 1-21 to better distinguish between different fairness methods.

#### A.8 MORE RESULTS ON ROBUSTNESS AGAINST ACCURACY ATTACK

We show the results on robustness against accuracy attack on COMPAS and GERMAN datasets in Fig. 7 and 8.

M	adv+pre	adv+in	adv+post
0.000	0.800	0.800	0.800
0.050	0.790	0.800	0.790
0.100	0.795	0.795	0.790
0.150	0.794	0.794	0.790
0.200	0.794	0.794	0.790
0.250	0.784	0.794	0.784
0.300	0.788	0.788	0.781
0.350	0.771	0.781	0.771
0.400	0.778	0.778	0.771
0.450	0.776	0.776	0.774
0.500	0.771	0.771	0.772

Table 1: results of accuracy for adversarial fair training on Adult dataset under EOd attack.

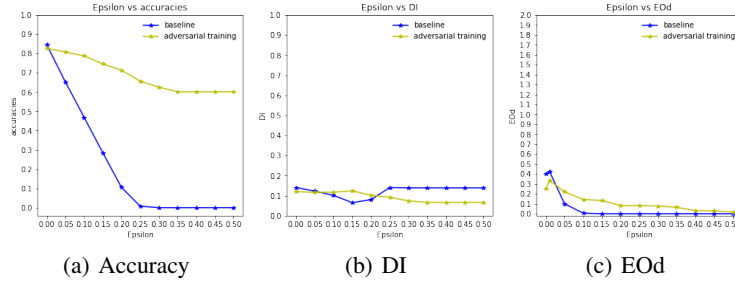


Figure 7: Results of a classifier adversarial trained w.r.t. EOd. Change of accuracy, DI and EOd under accuracy attack on COMPAS dataset.

M	adv+pre	adv+in	adv+post
0.000	0.029	0.039	0.016
0.050	0.117	0.137	0.098
0.100	0.129	0.111	0.119
0.150	0.128	0.108	0.108
0.200	0.114	0.104	0.114
0.250	0.123	0.093	0.123
0.300	0.114	0.074	0.104
0.350	0.099	0.059	0.090
0.400	0.086	0.046	0.096
0.450	0.103	0.073	0.113
0.500	0.152	0.152	0.132

Table 2: results of EOd for adversarial fair training on Adult dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.050	0.050	0.050
0.050	0.067	0.067	0.067
0.100	0.066	0.056	0.063
0.150	0.066	0.054	0.066
0.200	0.070	0.050	0.070
0.250	0.077	0.047	0.072
0.300	0.068	0.043	0.068
0.350	0.080	0.040	0.087
0.400	0.090	0.040	0.090
0.450	0.087	0.047	0.087
0.500	0.088	0.058	0.083

Table 3: results of DI for adversarial fair training on Adult dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.268	0.275	0.282
0.050	0.286	0.286	0.286
0.100	0.243	0.243	0.246
0.150	0.235	0.231	0.235
0.200	0.227	0.226	0.227
0.244	0.225	0.225	0.225
0.300	0.205	0.205	0.211
0.350	0.183	0.188	0.186
0.400	0.184	0.184	0.181
0.450	0.195	0.193	0.193
0.500	0.203	0.203	0.207

Table 4: results of white TPR for adversarial fair training on Adult dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.973	0.973	0.973
0.050	0.970	0.970	0.970
0.100	0.977	0.977	0.977
0.150	0.979	0.979	0.979
0.200	0.982	0.982	0.982
0.250	0.983	0.983	0.983
0.300	0.981	0.981	0.981
0.350	0.967	0.977	0.967
0.400	0.964	0.974	0.964
0.450	0.957	0.967	0.957
0.500	0.958	0.958	0.958

Table 5: results of white TNR for adversarial fair training on Adult dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.268	0.248	0.262
0.050	0.208	0.168	0.201
0.100	0.168	0.148	0.168
0.150	0.141	0.141	0.151
0.200	0.134	0.134	0.134
0.250	0.141	0.141	0.141
0.300	0.131	0.144	0.135
0.350	0.141	0.140	0.143
0.400	0.154	0.151	0.158
0.450	0.141	0.141	0.143
0.500	0.074	0.074	0.094

Table 6: results of black TPR for adversarial fair training on Adult dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.989	0.989	0.984
0.050	0.989	0.981	0.987
0.100	0.990	0.992	0.992
0.150	0.993	0.997	0.995
0.200	0.993	0.993	0.993
0.250	0.991	0.991	0.991
0.300	0.990	0.986	0.991
0.350	0.991	0.993	0.990
0.400	0.986	0.990	0.990
0.450	0.988	0.984	0.988
0.500	0.978	0.982	0.976

Table 7: results of black TNR for adversarial fair training on Adult dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.625	0.627	0.635
0.010	0.624	0.609	0.634
0.050	0.617	0.601	0.627
0.100	0.607	0.607	0.604
0.150	0.603	0.610	0.603
0.200	0.607	0.610	0.602
0.250	0.612	0.606	0.612
0.300	0.603	0.592	0.603
0.350	0.598	0.579	0.598
0.400	0.595	0.567	0.595
0.450	0.588	0.558	0.588
0.500	0.586	0.551	0.586

Table 8: results of accuracy for adversarial fair training on COMPAS dataset under EOd attack.

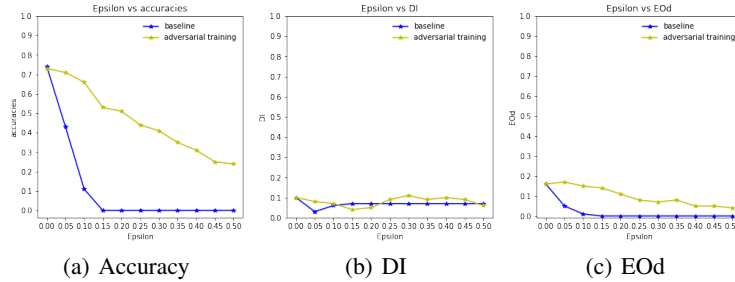


Figure 8: Results of a classifier adversarial trained w.r.t. EOd. Change of accuracy, DI and EOd under accuracy attack on German dataset.

M	adv+pre	adv+in	adv+post
0.000	0.044	0.240	0.024
0.010	0.197	0.584	0.147
0.050	0.735	0.979	0.565
0.100	0.960	1.146	0.910
0.150	1.131	1.231	1.041
0.200	1.214	1.289	1.254
0.250	1.348	1.387	1.348
0.300	1.463	1.502	1.463
0.350	1.513	1.598	1.513
0.400	1.623	1.645	1.623
0.450	1.676	1.665	1.676
0.500	1.705	1.710	1.705

Table 9: results of EOd for adversarial fair training on COMPAS dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.070	0.133	0.050
0.010	0.154	0.302	0.134
0.050	0.317	0.488	0.297
0.100	0.396	0.572	0.356
0.150	0.471	0.614	0.471
0.200	0.588	0.643	0.588
0.250	0.645	0.692	0.645
0.300	0.716	0.750	0.716
0.350	0.765	0.798	0.765
0.400	0.809	0.822	0.809
0.450	0.830	0.832	0.830
0.500	0.844	0.855	0.844

Table 10: results of DI for adversarial fair training on COMPAS dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.311	0.336	0.321
0.010	0.267	0.229	0.297
0.050	0.072	0.014	0.172
0.100	0.000	0.000	0.021
0.150	0.000	0.000	0.003
0.200	0.000	0.000	0.000
0.250	0.000	0.000	0.000
0.300	0.000	0.000	0.000
0.350	0.000	0.000	0.000
0.400	0.000	0.000	0.000
0.450	0.000	0.000	0.000
0.500	0.000	0.000	0.000

Table 11: results of white TPR for adversarial fair training on COMPAS dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.914	0.788	0.914
0.010	0.935	0.864	0.935
0.050	1.000	0.983	1.000
0.100	1.000	1.000	1.000
0.150	1.000	1.000	1.000
0.200	1.000	1.000	1.000
0.250	1.000	1.000	1.000
0.300	1.000	1.000	1.000
0.350	1.000	1.000	1.000
0.400	1.000	1.000	1.000
0.450	1.000	1.000	1.000
0.500	1.000	1.000	1.000

Table 12: results of white TNR for adversarial fair training on COMPAS dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.339	0.525	0.339
0.010	0.385	0.573	0.365
0.050	0.565	0.596	0.565
0.100	0.599	0.672	0.599
0.150	0.635	0.720	0.635
0.200	0.695	0.749	0.695
0.250	0.760	0.793	0.760
0.300	0.808	0.828	0.808
0.350	0.836	0.858	0.836
0.400	0.868	0.862	0.868
0.450	0.890	0.858	0.890
0.500	0.894	0.870	0.894

Table 13: results of black TPR for adversarial fair training on COMPAS dataset under EOD attack.

M	adv+pre	adv+in	adv+post
0.000	0.908	0.736	0.908
0.010	0.866	0.625	0.886
0.050	0.748	0.586	0.798
0.100	0.559	0.525	0.559
0.150	0.524	0.489	0.524
0.200	0.471	0.460	0.471
0.250	0.401	0.406	0.401
0.300	0.325	0.327	0.325
0.350	0.253	0.260	0.253
0.400	0.195	0.217	0.195
0.450	0.174	0.193	0.174
0.500	0.148	0.160	0.148

Table 14: results of black TNR for adversarial fair training on COMPAS dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.724	0.714	0.730
0.050	0.721	0.711	0.726
0.100	0.710	0.700	0.721
0.150	0.690	0.690	0.714
0.200	0.680	0.680	0.703
0.250	0.684	0.684	0.690
0.300	0.680	0.680	0.680
0.350	0.676	0.670	0.676
0.400	0.667	0.667	0.667
0.450	0.665	0.665	0.665
0.500	0.660	0.667	0.665

Table 15: results of accuracy for adversarial fair training on German dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.030	0.030	0.010
0.050	0.140	0.140	0.100
0.100	0.380	0.380	0.380
0.150	0.770	0.770	0.670
0.200	0.960	0.960	0.950
0.250	1.270	1.270	1.250
0.300	1.290	1.340	1.290
0.350	1.310	1.360	1.330
0.400	1.340	1.460	1.340
0.450	1.440	1.540	1.490
0.500	1.500	1.600	1.560

Table 16: results of EOd for adversarial fair training on German dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.060	0.120	0.020
0.050	0.130	0.130	0.110
0.100	0.170	0.170	0.140
0.150	0.260	0.260	0.270
0.200	0.300	0.300	0.340
0.250	0.360	0.360	0.360
0.300	0.410	0.410	0.440
0.350	0.470	0.470	0.470
0.400	0.580	0.580	0.550
0.450	0.640	0.640	0.600
0.500	0.670	0.670	0.710

Table 17: results of DI for adversarial fair training on German dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.364	0.364	0.364
0.050	0.350	0.350	0.350
0.100	0.260	0.260	0.260
0.150	0.130	0.130	0.190
0.200	0.110	0.110	0.110
0.250	0.070	0.070	0.070
0.300	0.000	0.000	0.000
0.350	0.000	0.000	0.000
0.400	0.000	0.000	0.000
0.450	0.000	0.000	0.000
0.500	0.000	0.000	0.000

Table 18: results of male TPR for adversarial fair training on German dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.857	0.857	0.850
0.050	0.870	0.870	0.870
0.100	0.920	0.920	0.920
0.150	1.000	1.000	1.000
0.200	1.000	1.000	1.000
0.250	1.000	1.000	1.000
0.300	1.000	1.000	1.000
0.350	1.000	1.000	1.000
0.400	1.000	1.000	1.000
0.450	1.000	1.000	1.000
0.500	1.000	1.000	1.000

Table 19: results of male TNR for adversarial fair training on German dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.377	0.377	0.377
0.050	0.420	0.420	0.420
0.100	0.510	0.510	0.510
0.150	0.570	0.570	0.570
0.200	0.680	0.680	0.680
0.250	0.750	0.750	0.750
0.300	0.770	0.770	0.770
0.350	0.780	0.780	0.790
0.400	0.800	0.810	0.800
0.450	0.860	0.870	0.860
0.500	0.870	0.890	0.870

Table 20: results of female TPR for adversarial fair training on German dataset under EOd attack.

M	adv+pre	adv+in	adv+post
0.000	0.833	0.833	0.843
0.050	0.810	0.810	0.820
0.100	0.740	0.740	0.740
0.150	0.670	0.670	0.690
0.200	0.560	0.560	0.560
0.250	0.490	0.490	0.510
0.300	0.480	0.440	0.480
0.350	0.460	0.410	0.460
0.400	0.450	0.340	0.450
0.450	0.400	0.310	0.370
0.500	0.300	0.240	0.230

Table 21: results of female TNR for adversarial fair training on German dataset under EOd attack.