A Dataset Details and experimental setup

- 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 dateset 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*.
- CelebA (Liu et al., 2015): CelebA dataset contains 202,599 samples with 40 binary attributes. We choose gender as target label, and the sensitive attribute in this dataset is *age*.

The classifier is chosen as ResNet-18 for CelebA and MLP for the other three datasets, 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, 0.01 for German dataset and 0.1 for CelebA dataset, where the the perturbation level is empirically determined to achieve the largest perturbation while still ensuring convergence.

B Empirical verification of Theoretical results

We empirically validate our discussion regarding the relationship between DI and EOd attack as in Corollary 1. As shown in Fig 2, under a successful DI attack, EOd always reaches its maximum, and a successful DI attack also leads to a successful EOd attack. We also empirically verify the effectiveness of upper-bounds stated in Theorem 1. The following results on CelebA dataset shows the change of cross-entropy loss for samples from different groups by baseline and fair adversarial training under different perturbation levels:

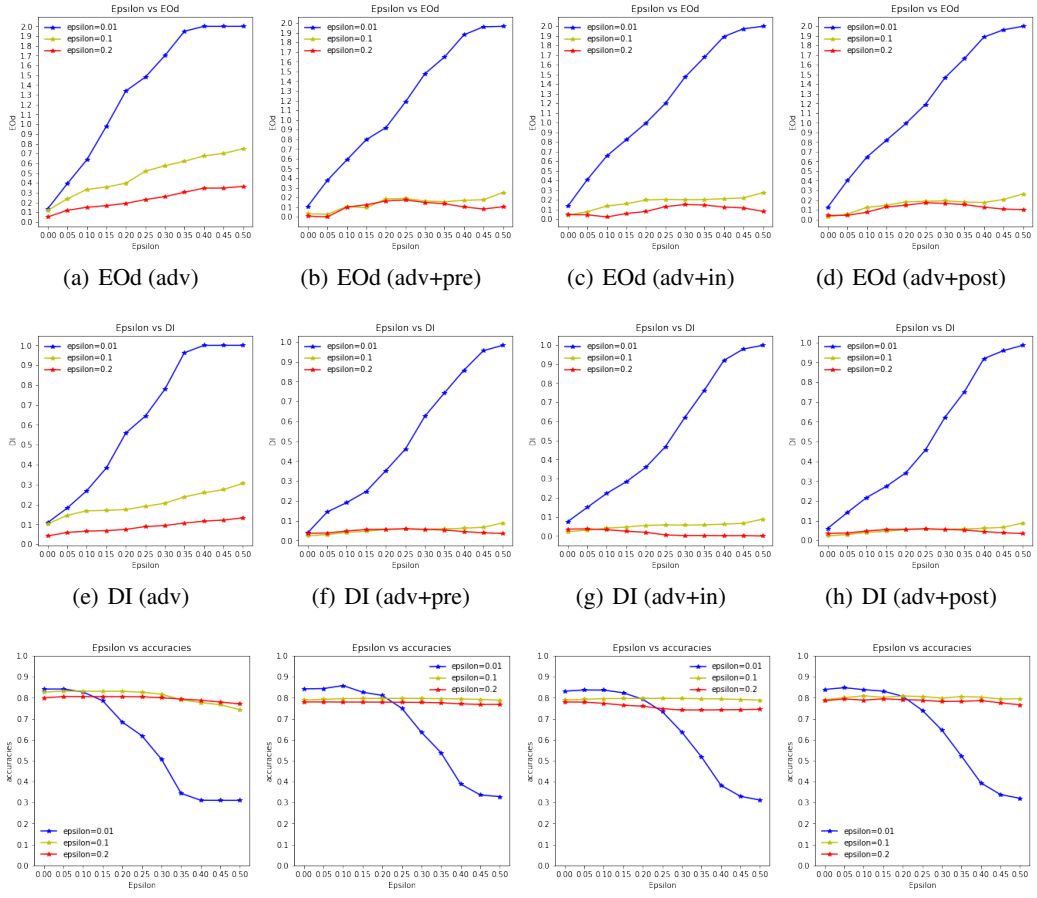
ϵ	Method	$D^{\text{Fair}}_{\text{FN,male}}$	$D^{\text{Fair}}_{\text{FN,female}}$
0.1	Baseline	0.16±0.03	0.18±0.02
0.1	Adversarial training (preprocessing)	**0.07±0.02**	**0.09±0.02**
0.1	Adversarial training (in-processing)	0.07±0.02	0.11±0.02
0.1	Adversarial training (post-processing)	0.08±0.01	0.09±0.02
0.3	Baseline	0.23±0.02	0.26±0.03
0.3	Adversarial training (preprocessing)	**0.09±0.02**	0.11±0.02
0.3	Adversarial training (in-processing)	0.10±0.02	0.12±0.02
0.3	Adversarial training (post-processing)	0.10±0.02	**0.09±0.01**

Table 2: Change of cross-entropy loss for FN samples on CelebA dataset under fairness attacks with $\epsilon = 0.1$ and $\epsilon = 0.3$. Experiments are repeated three times.

As shown in Table 2, under fair adversarial training, both advantaged and disadvantaged groups show improvements in D^{Fair} compared with the baseline, which validates our theoretical results, that is, the alignment between fairness robustness and accuracy robustness.

C Results on varying ϵ

Results of varying ϵ on Adult, COMPAS, German and CelebA dataset can be found in Fig. 4-7. As shown in the figures, larger perturbation levels result in classifiers that are more robust to adversarial perturbations against fairness for both vanilla adversarial training and fair adversarial training during testing.



(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) Accuracy (adv) (j) Accuracy (adv+pre) (k) Accuracy (adv+in) (l) Accuracy (adv+post)

Figure 4: Change of accuracy, DI and EOd under DI attack with varying training perturbation ϵ on Adult dataset.

D Proof of Corollary 1

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

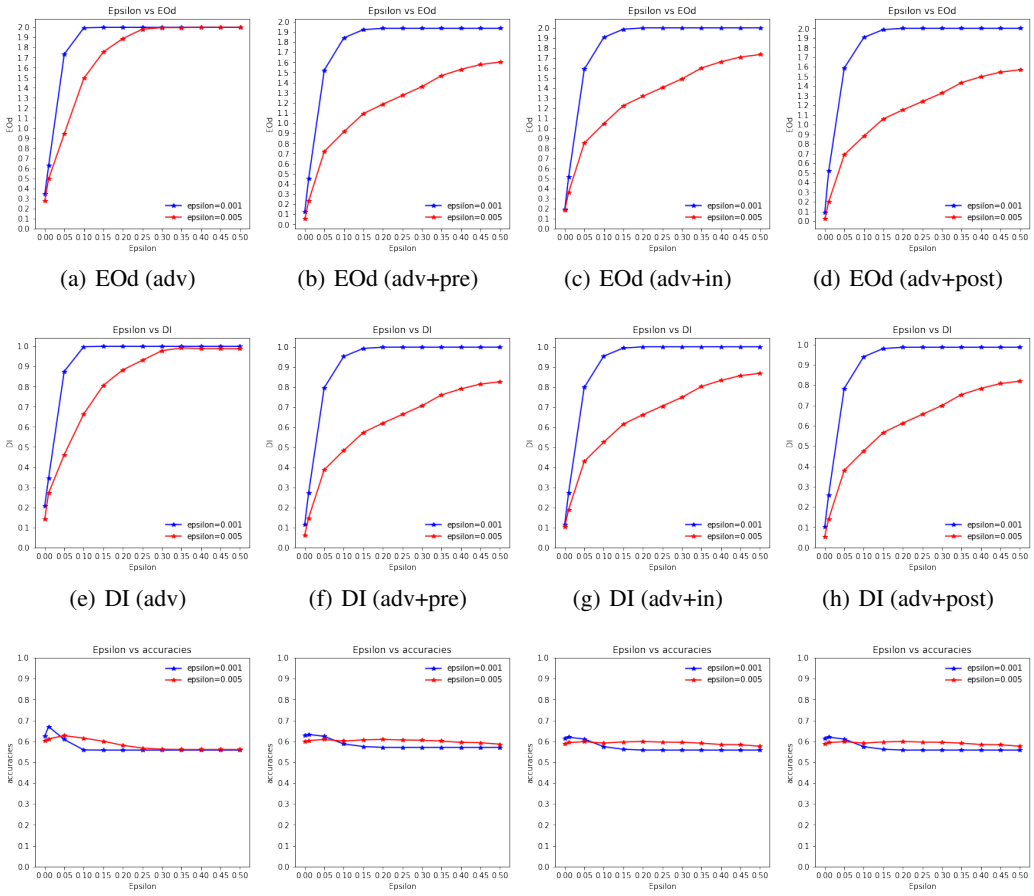
$$
\begin{aligned}
L_{\text{EOd}} &= \left| \sum_{x_i \in \mathbb{S}_{00}} \frac{f(x_i)}{|\mathbb{S}_{00}|} - \sum_{x_i \in \mathbb{S}_{01}} \frac{f(x_i)}{|\mathbb{S}_{01}|} \right| + \left| \sum_{x_i \in \mathbb{S}_{10}} \frac{f(x_i)}{|\mathbb{S}_{10}|} - \sum_{x_i \in \mathbb{S}_{11}} \frac{f(x_i)}{|\mathbb{S}_{11}|}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

This shows that the EOd attack is lower-bounded by the weighted DI attack as in equation 1. Specifically, under a successful DI attack, we have $f(x) = 1, \forall x \in \mathbb{S}_{.a}$ and $f(x) = 0, \forall x \in \mathbb{S}_{.a'}$, and the lower bound can be simplified as

$$
L_{\text{EOd}} \geq \left| \sum_{x \in \mathbb{S}_{0a}} \frac{|\mathbb{S}_{.a}|}{|\mathbb{S}_{0a}|} \frac{1}{|\mathbb{S}_{.a}|} + \sum_{x \in \mathbb{S}_{1a}} \frac{|\mathbb{S}_{.a}|}{|\mathbb{S}_{1a}|} \frac{1}{|\mathbb{S}_{.a}|} \right| = 2, \tag{5}
$$

which shows that a successful DI attack always implies a successful EOd attack.

Remark 3. A successful EOd attack does not always imply a successful DI attack. Assume $\sum_{x_i \in \mathbb{S}_{00}} \frac{f(x_i)}{|\mathbb{S}_{00}|} \leq \sum_{x_i \in \mathbb{S}_{01}} \frac{f(x_i)}{|\mathbb{S}_{01}|}$ and $\sum_{x_i \in \mathbb{S}_{10}} \frac{f(x_i)}{|\mathbb{S}_{10}|} \geq \sum_{x_i \in \mathbb{S}_{11}} \frac{f(x_i)}{|\mathbb{S}_{11}|}$, under a successful EOd

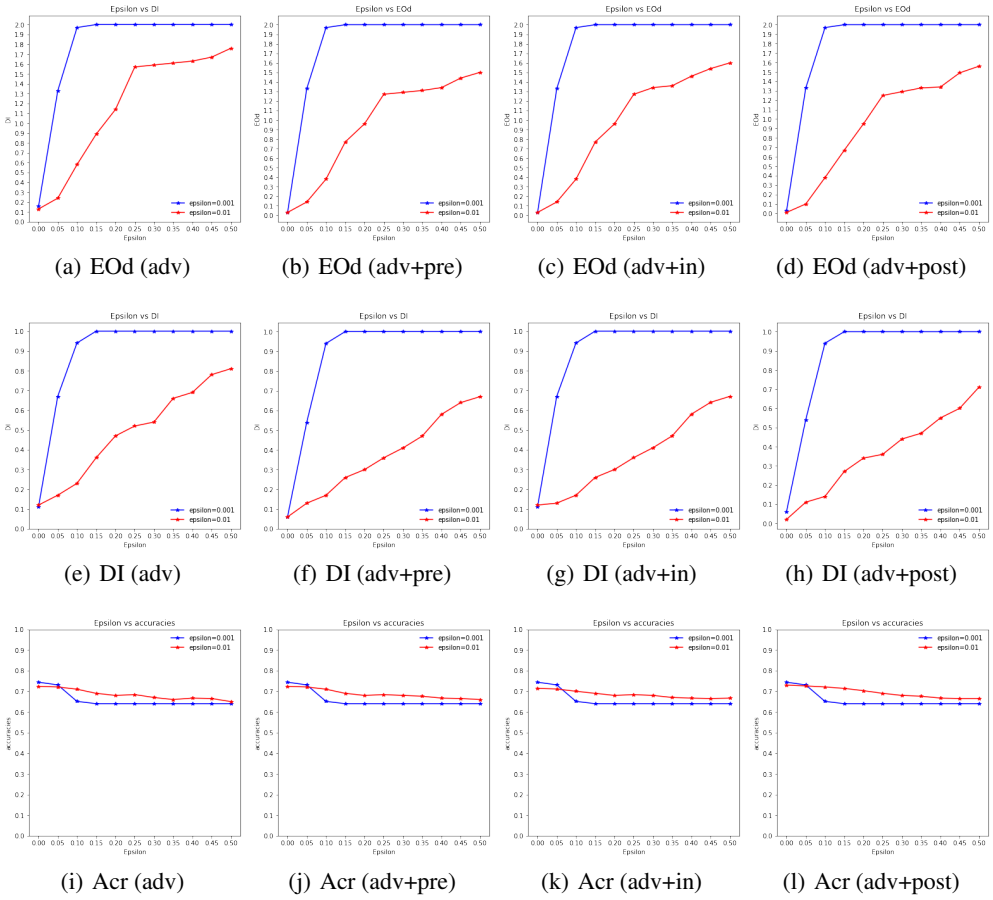


(a) EOd (adv) (b) EOd (adv+pre) (c) EOd (adv+in) (d) EOd (adv+post)

(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 5: Change of accuracy, DI and EOd under DI attack with varying training perturbation ϵ on COMPAS dataset.



(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 6: Change of accuracy, DI and EOd under DI attack with varying training perturbation ϵ on German dataset.

attack, all the predictions in the disadvantaged group will become correct, while all the predictions in the advantaged group will become incorrect, and the disparate impact will not be maximized as both groups contain positive predictions.

□

E Proof of Corollary 2

Proof. The objective for accuracy attack for sample x_i can be written as

$$\max_{\delta} L_{\text{CE}}((x_i + \epsilon), y_i), \|\epsilon\| \leq \epsilon_0, \tag{6}$$

Consider the DI attack in equation 1, we have the objective for DI attack as follows:

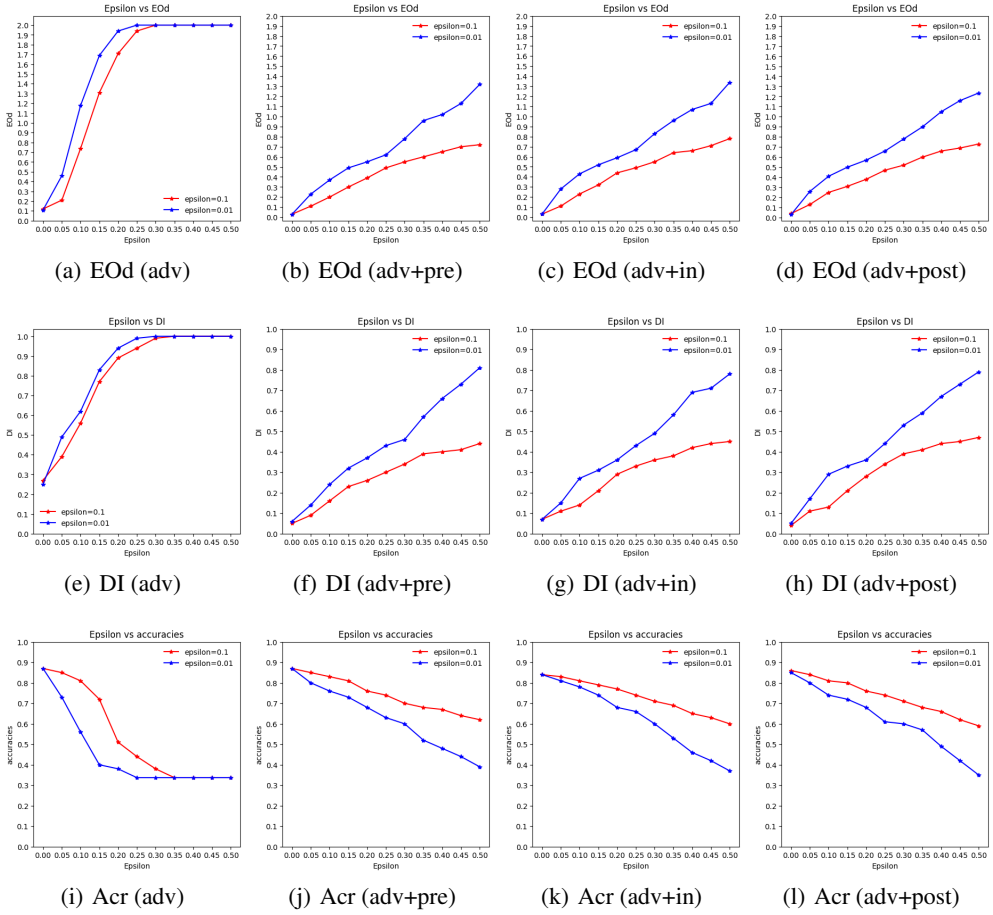
$$\max_{\delta} \alpha_i \frac{f(x_i + \epsilon)}{|\mathbb{S}_{.a_i}|}, \|\epsilon\| \leq \epsilon',$$

where $\alpha_i = -1$ for $a_i = 0$ and $\alpha_i = 1$ for $a_i = 1$. For positive samples, we can further write equation 6 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 DI when $\alpha_i = -1$, i.e., for TP and FN disadvantaged samples, the two attacks are in alignment. Similarly, for negative samples, we have equation 6 as

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



(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 7: Change of accuracy, DI and EOd under DI attack with varying training perturbation ϵ on CelebA dataset.

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

F Proof of Theorem 1

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

$$L_{\text{CE}}(x^t) = -\log(f(x^t)) = -\log(f(x^{t-1}) - \delta^{t-1,\text{Fair}}) = -\log(f(x^{t-1})) + \frac{\delta^{t-1,\text{Fair}}}{f(x^{t-1})} + r_L(x^{t-1}), \tag{7}$$

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

$$\begin{aligned} & f(x^t) \\ =& f(x^{t-1} + \alpha \text{sign}(\nabla_{x^{t-1}} L_{\text{DI}})) \\ =& f(x^{t-1}) + \alpha (\nabla_{x^{t-1}} f(x^{t-1}))^T \text{sign}(\nabla_{x^{t-1}} L_{\text{DI}}) + r_f(x^{t-1}), \end{aligned} \tag{8}$$

where L_{DI} is the relaxed DI loss and $r_f(x)$ is the remainder of Taylor's expansion. Let $D^{t,\text{Fair}} := |L(x^t) - L(x^{t-1})|$ be the change of CE loss under the fairness attack at t-th iteration, according to

equation 7 and equation 8 we have

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

Consider FN sample $x_{\text{FN},0}$ from disadvantaged group and FN sample $x_{\text{FN},1}$ 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 DI attack at t-th iteration as follows:

$$\begin{aligned} &|D^{t,\text{Fair}}_{\text{FN},1} - D^{t,\text{Fair}}_{\text{FN},0}| \\ =&\alpha \left| \frac{|(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} f(x^{t-1,\text{Fair}}_{\text{FN},1}))^T \text{sign}(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} L_{\text{DI}})|}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} - \frac{|(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0}))^T \text{sign}(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} L_{\text{DI}})|}{f(x^{t-1,\text{Fair}}_{\text{FN},0})} \right| \\ =&\alpha \left| \frac{(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} f(x^{t-1,\text{Fair}}_{\text{FN},1}))^T \text{sign}(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} L_{\text{DI}})}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} + \frac{(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f_\theta(x^{t-1,\text{Fair}}_{\text{FN},0}))^T \text{sign}(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} L_{\text{DI}})}{f(x^{t-1,\text{Fair}}_{\text{FN},0})} \right| \\ =&\alpha \left| \frac{(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} f(x^{t-1,\text{Fair}}_{\text{FN},1}))^T \text{sign}(\frac{1}{N_1} \nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} f(x^{t-1,\text{Fair}}_{\text{FN},1}))}{f_\theta(x^{t-1,\text{Fair}}_{\text{FN},1})} - \frac{(\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0}))^T \text{sign}(\frac{1}{N_0} \nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0}))}{f(x^{t-1,\text{Fair}}_{\text{FN},0})} \right| \\ =&\alpha \left| \frac{\sum_{j=1}^n |\partial_{x_j} f(x^{t-1,\text{Fair}}_{\text{FN},1})|}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} - \frac{\sum_{j=1}^n |\partial_{x_j} f(x^{t-1,\text{Fair}}_{\text{FN},0})|}{f(x^{t-1,\text{Fair}}_{\text{FN},0})} \right| \\ =&\alpha \left| \frac{\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} f(x^{t-1,\text{Fair}}_{\text{FN},1})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} - \frac{\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},0})} \right|, \end{aligned} \tag{9}$$

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

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

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

$$\|\nabla_x f(x_1)\|_1 - \|\nabla_x f(x_0)\|_1 \leq \|\nabla_x f(x_0) - \nabla_x f(x_1)\|_1 \leq \sqrt{n} K d(x_1, x_0). \tag{10}$$

Assume $\frac{\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} f(x^{t-1,\text{Fair}}_{\text{FN},1})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} \geq \frac{\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},0})}$, plugging equation 10 back into equation 9, we have

$$\begin{aligned} &|D^{t,\text{Fair}}_{\text{FN},1} - D^{t,\text{Fair}}_{\text{FN},0}| \\ =&\alpha \left| \frac{\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} f(x^{t-1,\text{Fair}}_{\text{FN},1})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} - \frac{\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},0})} \right| \\ \leq&\alpha \left| \frac{\sqrt{n} K d(x^{t-1,\text{Fair}}_{\text{FN},1}, x^{t-1,\text{Fair}}_{\text{FN},0}) + \|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} - \frac{\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},0})} \right| \\ \leq&\frac{\sqrt{n}\alpha K d(x^{t-1,\text{Fair}}_{\text{FN},1}, x^{t-1,\text{Fair}}_{\text{FN},0})}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} + \left| \frac{\alpha\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} - \frac{\alpha\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},0})} \right|, \end{aligned} \tag{11}$$

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

$$|D^{\text{Fair}}_{\text{FN},1} - D^{\text{Fair}}_{\text{FN},0}| \leq \sum_{t=1}^{T} \left[\frac{\sqrt{n}\alpha K d(x^{t-1,\text{Fair}}_{\text{FN},1}, x^{t-1,\text{Fair}}_{\text{FN},0})}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} + \alpha \left| \frac{f(x^{t-1,\text{Fair}}_{\text{FN},0}) - f(x^{t-1,\text{Fair}}_{\text{FN},1})}{f(x^{t-1,\text{Fair}}_{\text{FN},1}) f(x^{t-1,\text{Fair}}_{\text{FN},0})} \right| \delta^{t-1,\text{Acc}}_{\text{FN},0} \right], \tag{12}$$

where $\delta^{t-1,\text{Acc}}_{\text{FN},0} \coloneqq \|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f_\theta(x^{t-1,\text{Fair}}_{\text{FN},0})\|_1$ is the change of $x_{\text{FN},0}$'s predicted label under ϵ-level accuracy attack at t-th iteration since both are equivalent regarding $x_{\text{FN},0}$. Since the above inequality holds true for all disadvantaged TP samples and $D^{\text{Acc}}_{\text{FN},1} = D^{\text{Fair}}_{\text{FN},1}$, we can further write equation 12 as

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

This shows that under the fairness attack, the difference of change in performance regarding marginal advantaged FN samples are upper-bounded by the robustness of marginal disadvantaged FN samples up to an additive constant. For f under normal training and f' under normal training, we have similar upper-bound except that we now have $\delta'^{t-1,\text{Acc}}_{\text{FN},0} \geq \delta^{t-1,\text{Acc}}_{\text{FN},0}$, which indicates that the adversarial classifier achieves tighter upper-bound than that of a normal classifier. For $\frac{\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},1}} f(x^{t-1,\text{Fair}}_{\text{FN},1})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},1})} \leq \frac{\|\nabla_{x^{t-1,\text{Fair}}_{\text{FN},0}} f(x^{t-1,\text{Fair}}_{\text{FN},0})\|_1}{f(x^{t-1,\text{Fair}}_{\text{FN},0})}$, we have same upper-bound:

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

□

G Proof of Theorem 2

Proof. Let e_{ya} be the error rate in the subgroup $\mathbb{S}_{ya}$, let b_a be the base rate in group a, denote as mea^{Fair} the fairness measure mea after the fairness attack, we have the following expression regarding DI after the fairness attack:

$$\begin{aligned}
&\text{DI}^{\text{Fair}} \\
&= \left| \int_{0.5}^{1} p_0^{\text{Fair}} df - \int_{0.5}^{1} p_1^{\text{Fair}} df \right| \\
&= \left| \int_{0.5}^{1} \left(b_0(1 - e_{10}^{\text{Fair}}) p_{\text{TP},0}^{\text{Fair}} + (1 - b_0) e_{00}^{\text{Fair}} p_{\text{FP},0}^{\text{Fair}} \right) df - \int_{0.5}^{1} \left(b_1(1 - e_{11}^{\text{Fair}}) p_{\text{TP},1}^{\text{Fair}} + (1 - b_1) e_{01}^{\text{Fair}} p_{\text{FP},1}^{\text{Fair}} \right) df \right| \\
&\leq \left| \int_{0.5+\Delta^{\text{Fair}}_{\text{TP},0}}^{1} b_0(1 - e_{10}) p_{\text{TP},0} df + \int_{0.5+\Delta^{\text{Fair}}_{\text{FP},0}}^{1} (1 - b_0) e_{00} p_{\text{FP},0} df - \int_{0.5}^{1} b_1(1 - e_{11}) p_{\text{TP},1} df \right. \\
&\left. - \int_{0.5}^{1} (1 - b_1) e_{01} p_{\text{FP},1} df - \int_{0.5-\Delta^{\text{Fair}}_{\text{TN},1}}^{0.5} (1 - b_1)(1 - e_{01}) p_{\text{TN},1} df - \int_{0.5-\Delta^{\text{Fair}}_{\text{FN},1}}^{0.5} b_1 e_{11} p_{\text{FN},1} df \right| \\
&= \left| b_1(1 - e_{11}) P_{\text{TP},1}(0.5) + (1 - b_1) e_{01} P_{\text{FP},1}(0.5) + (1 - b_1)(1 - e_{01}) P_{\text{TN},1}(0.5 - \Delta^{\text{Fair}}_{\text{TN},1}) \right. \\
&\left. + b_1 e_{11} P_{\text{FN},1}(0.5 - \Delta^{\text{Fair}}_{\text{FN},1}) - b_0(1 - e_{10}) P_{\text{TP},0}(0.5 + \Delta^{\text{Fair}}_{\text{TP},0}) - (1 - b_0) e_{00} P_{\text{FP},0}(0.5 + \Delta^{\text{Fair}}_{\text{FP},0}) \right|,
\end{aligned} \tag{13}$$

where $\Delta^{\text{Fair}}_{\text{sub},a} := \max_{i\in\{\text{sub},a\}} \delta^{\text{Fair}}_i$ is the maximum prediction shift within the subgroup, $P_{\text{sub},a}$ is the CDF of $p_{\text{sub},a}$, and the inequality is due to that the worst-case prediction shift upper-bounds the overall shift in the distribution of soft prediction. Since $P_{\text{sub},a}$ is Lipschitz continuous with constant $M_{\text{sub},a}$ ($p_{\text{sub},a}$ is uniformly bounded by $M_{\text{sub},a}$), we can further simplify equation 13 as

$$
\begin{aligned}
&\text{DI}^{\text{Fair}} \\
&\leq \big|b_1(1-e_{11})P_{\text{TP},1}(0.5) + (1-b_1)e_{01}P_{\text{FP},1}(0.5) + (1-b_1)(1-e_{01})P_{\text{TN},1}(0.5-\Delta^{\text{Fair}}_{\text{TN},1}) \\
&+b_1e_{11}P_{\text{FN},1}(0.5-\Delta^{\text{Fair}}_{\text{FN},1}) - b_0(1-e_{10})P_{\text{TP},0}(0.5+\Delta^{\text{Fair}}_{\text{TP},0}) - (1-b_0)e_{00}P_{\text{FP},0}(0.5+\Delta^{\text{Fair}}_{\text{FP},0})\big| \\
&\leq \text{DI} + b_0(1-e_{10})M_{\text{TP},0}\Delta^{\text{Fair}}_{\text{TP},0} + (1-b_0)e_{00}M_{\text{FP},0}\Delta^{\text{Fair}}_{\text{FP},0} + (1-b_1)(1-e_{01})M_{\text{TN},1}\Delta^{\text{Fair}}_{\text{TN},1} + b_1e_{11}M_{\text{FN},1}\Delta^{\text{Fair}}_{\text{FN},1} \\
&\leq \text{DI} + M(\Delta^{\text{Acc}}_{\text{TP},0} + \min_{j\in\mathbb{S}_{\text{FP},1}}(D^{\text{Acc}}_j + H_j) + \Delta^{\text{Acc}}_{\text{TN},1} + \min_{j\in\mathbb{S}_{\text{FN},0}}(D^{\text{Acc}}_j + G_j)),
\end{aligned}
$$

where $M = \max\{M_{\text{TP},0}, M_{\text{FP},0}, M_{\text{TN},1}, M_{\text{FN},1}\}$, and the two minimization terms in the last inequality correspond to the upper-bounds in Theorem 1 and Remark 1. Since the fairness robustness and accuracy robustness are equivalent regarding $x_{\text{TP},0}$ and $_{\text{TN},1}$, and D_j, H_j and G_j are determined by the intrinsic distance between samples and the accuracy robustness of $x_{\text{FP},1}$ and $x_{\text{FN},0}$, we can conclude that DI^{Fair} is upper-bounded by static fairness, i.e., the DI term, and the accuracy robustness $\delta^{\text{Acc}}_{\text{TP},0}$, $\min_{j\in\mathbb{S}_{\text{FP},1}}(D^{\text{Acc}}_j + H_j)$, $\Delta^{\text{Acc}}_{\text{TN},1}$ and $\min_{j\in\mathbb{S}_{\text{FN},0}}(D^{\text{Acc}}_j + G_j)$, which validates our fair adversarial training framework.

Similarly, we have the following upper-bound regarding EOd^{Fair}:

$$
\begin{aligned}
\text{EOd}^{\text{Fair}} &= \left|\int_{0.5}^{1} p^{\text{Fair}}_{00}df - \int_{0.5}^{1} p^{\text{Fair}}_{01}df\right| + \left|\int_{0.5}^{1} p^{\text{Fair}}_{10}df - \int_{0.5}^{1} p^{\text{Fair}}_{11}df\right| \\
&= \left|\int_{0.5}^{1}\left((1-e^{\text{Fair}}_{00})p^{\text{Fair}}_{\text{TN},0} + e^{\text{Fair}}_{00}p^{\text{Fair}}_{\text{FP},0}\right)df - \int_{0.5}^{1}\left((1-e'_{01})p^{\text{Fair}}_{\text{TN},1} + e^{\text{Fair}}_{01}p^{\text{Fair}}_{\text{FP},1}\right)df\right| \\
&+ \left|\int_{0.5}^{1}\left((1-e^{\text{Fair}}_{10})p^{\text{Fair}}_{\text{TP},0} + e^{\text{Fair}}_{10}p^{\text{Fair}}_{\text{FN},0}\right)df - \int_{0.5}^{1}\left((1-e^{\text{Fair}}_{11})p^{\text{Fair}}_{\text{TP},1} + e^{\text{Fair}}_{11}p^{\text{Fair}}_{\text{FN},1}\right)df\right| \\
&= \left|\int_{0.5}^{1}(1-e_{00})p_{\text{TN},0}df + \int_{0.5+\Delta^{\text{Fair}}_{\text{FP},0}}^{1} e_{00}p_{\text{FP},0}df - \int_{0.5-\Delta^{\text{Fair}}_{\text{TN},1}}^{1}(1-e_{01})p_{\text{TN},1}df - \int_{0.5}^{1} e_{01}p_{\text{FP},1}df\right| \\
&+ \left|\int_{0.5+\Delta^{\text{Fair}}_{\text{TP},0}}^{1}(1-e_{10})p_{\text{TP},0}df + \int_{0.5}^{1} e_{10}p_{\text{FN},0}df - \int_{0.5}^{1}(1-e_{11})p_{\text{TP},1}df - \int_{0.5-\Delta^{\text{DI}}_{\text{FN},1}}^{1} e_{11}p_{\text{FN},1}df\right| \\
&\leq \text{EOd} + e_{00}M_{\text{FP},0}\Delta^{\text{Fair}}_{\text{FP},0} + (1-e_{01})M_{\text{TN},1}\Delta^{\text{Fair}}_{\text{TN},1} + (1-e_{10})M_{\text{TP},0}\Delta^{\text{Fair}}_{\text{TP},0} + e_{11}M_{\text{FN},1}\Delta^{\text{Fair}}_{\text{FN},1} \\
&\leq \text{EOd} + M((\Delta^{\text{Acc}}_{\text{TP},0} + \min_{j\in\mathbb{S}_{\text{FP},1}}(D^{\text{Acc}}_j + H_j) + \Delta^{\text{Acc}}_{\text{TN},1} + \min_{j\in\mathbb{S}_{\text{FN},0}}(D^{\text{Acc}}_j + G_j))),
\end{aligned}
$$

where the first term in the last inequality corresponds to static fairness, i.e., EOd without fairness perturbation, and the second term corresponds to accuracy robustness. □

H Proof of Theorem 3

Proof. Let f be the function of classifier, consider $x_{\text{TP},0}$, we have the predicted soft label for sample $x_{\text{TP},0}$ under accuracy attack at t-th iteration as follows:

$$
\begin{aligned}
&f(x^{t,\text{Acc}}_{\text{TP},0}) \\
=&f(x^{t-1,\text{Acc}}_{\text{TP},0} + \alpha\text{sign}(\nabla_{x^{t-1,\text{Acc}}_{\text{TP},0}} L_{\text{CE}})) \\
\approx&f(x^{t-1,\text{Acc}}_{\text{TP},0}) + \alpha(\nabla_{x^{t-1,\text{Acc}}_{\text{TP},0}} f(x^{t-1,\text{Acc}}_{\text{TP},0}))^T\text{sign}(-\frac{1}{f(x^{t-1,\text{Acc}}_{\text{TP},0})}\nabla_{x^{t-1,\text{Acc}}_{\text{TP},0}} f(x^{t-1,\text{Acc}}_{\text{TP},0})) \\
=&f(x^{t-1,\text{Acc}}_{\text{TP},0}) + \alpha(\nabla_x f(x^{t-1,\text{Acc}}_{\text{TP},0}))^T\text{sign}(\nabla_{x^{t-1,\text{Acc}}_{\text{TP},0}} L_{\text{CE}}) \\
=&f(x^{t-1,\text{Acc}}_{\text{TP},0}) - \alpha\|\nabla_{x^{t-1,\text{Acc}}_{\text{TP},0}} f(x^{t-1,\text{Acc}}_{\text{TP},0})\|_1 \\
=&f(x^{t-1,\text{Acc}}_{\text{TP},0}) - \delta^{t-1,\text{Fair}}_{\text{TP},0},
\end{aligned}
$$

fwhere $\delta_{\text{TP},0}^{t,\text{Fair}} := \alpha \|\nabla_{x_{\text{TP},0}^{t-1,\text{Acc}}} f(x_{\text{TP},0}^{t-1,\text{Acc}})\|_1$ is the change of $x_{\text{TP},0}$'s predicted label under ϵ-level fairness attack at t-th iteration since both are equivalent regarding $x_{\text{TP},0}$. This shows that disadvantaged TP samples that attains δ-level robustness under ϵ-level fairness attack also attains similar robustness w.r.t. accuracy attack.

For $x_{\text{TP},1}$, let $\delta_{\text{TP},1}^{t,\text{Acc}} := |f(x_{\text{TP},1}^{t,\text{Acc}}) - f(x_{\text{TP},1}^{t-1,\text{Acc}})|$, we have its change in predicted soft label under accuracy attack at t-th iteration as follows:

$$
\begin{aligned}
&\delta(x_{\text{TP},1}^{t,\text{Acc}}) \\
=&|f(x_{\text{TP},1}^{t,\text{Acc}}) - f(x_{\text{TP},1}^{t-1,\text{Acc}})| \\
=&|f(x_{\text{TP},1}^{t-1,\text{Acc}} + \alpha \text{sign}(\nabla_{x_{\text{TP},1}^{t-1,\text{Acc}}} L_{\text{CE}})) - f(x_{\text{TP},1}^{t-1,\text{Acc}})| \\
\approx&\alpha (\nabla_{x_{\text{TP},1}^{t-1,\text{Acc}}} f(x_{\text{TP},1}^{t-1,\text{Acc}}))^T \text{sign}(\nabla_{x_{\text{TP},1}^{t-1,\text{Acc}}} L_{\text{CE}}) \\
=&\alpha \|\nabla_{x_{\text{TP},1}^{t-1,\text{Acc}}} f(x_{\text{TP},1}^{t-1,\text{Acc}})\|_1 \\
\leq&\delta_{\text{TP},0}^{t,\text{Fair}} + \sqrt{n}\alpha K d(x_{\text{TP},0}^{t-1,\text{Acc}}, x_{\text{TP},1}^{t-1,\text{Acc}}).
\end{aligned}
\tag{14}
$$

Taking the summation over all iterations, we have

$$
\delta_{\text{TP},1}^{\text{Acc}} \leq \delta_{\text{TP},0}^{\text{Fair}} + \sum_{t=1}^{T} \sqrt{n}\alpha K d(x_{\text{TP},0}^{t-1,\text{Acc}}, x_{\text{TP},1}^{t-1,\text{Acc}}), \tag{15}
$$

where $\delta_{\text{TP},0}^{\text{Fair}}$ is the change of predicted soft label of sample $x_{\text{TP},0}$ under ϵ-level fairness attack. Since the inequality hold true for all $x_{\text{TP},0}$, we can further write equation 15 as

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

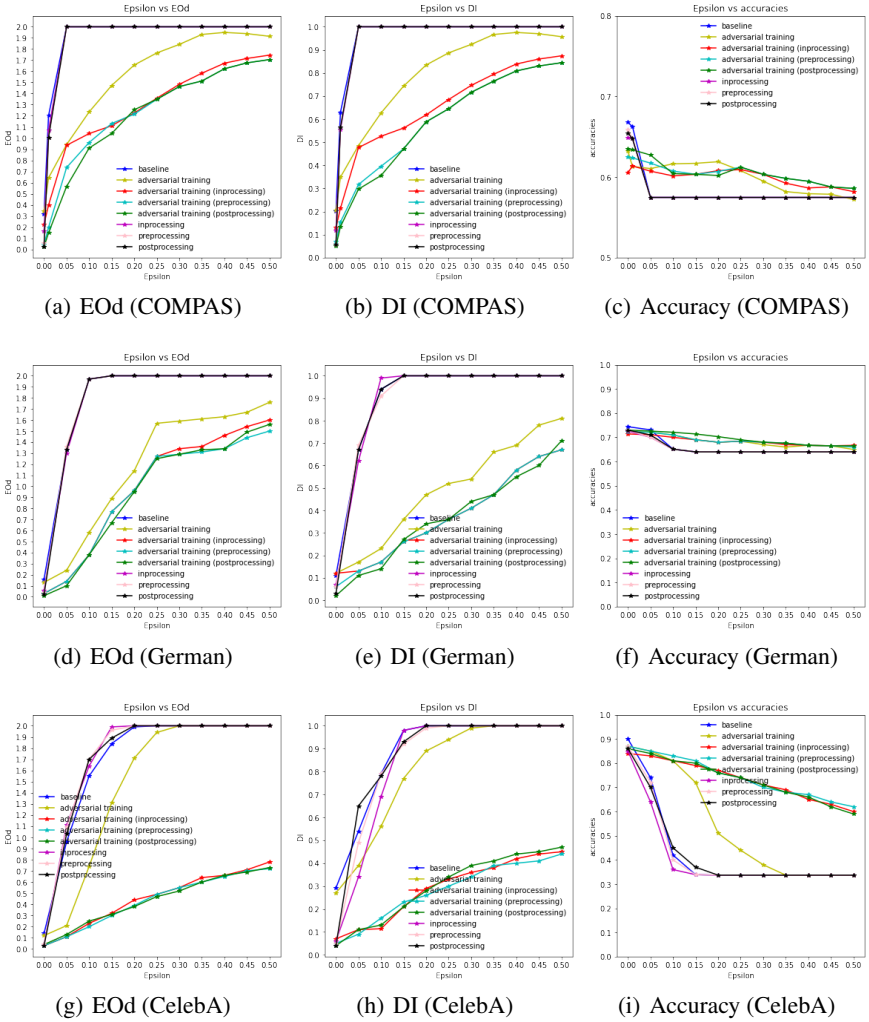
And the lower bound $\delta_{\text{TP},1}^{\text{Acc}} \geq 0$ naturally holds true for samples under accuracy attack. This shows that for samples in the advantaged group, the change of predicted soft label under accuracy attack is lower-bounded by the fairness robustness of its neighbor sample(s) in the disadvantaged group up to an additive constant. For f'' under adversarial training w.r.t. fairness and f under normal training, we have similar upper-bound except that we now have $\delta_{\text{TP},0}^{\text{Fair}} \geq \delta_{\text{TP},0}^{''\text{Fair}}$, which indicates that the adversarial classifier achieves tighter upper-bound than that of a normal classifier. □

I RESULTS OF ROBUSTNESS AGAINST DI ATTACK

We include the results of fair adversarial training in Tab. 3-30 and Fig. 9 to better distinguish between different fairness methods. Results of classifiers under DI attack on COMPAS and German dataset are shown in Fig. 2. in Fig. 8.

J MORE RESULTS ON ROBUSTNESS AGAINST ACCURACY ATTACK

We show the results on robustness against accuracy attack on COMPAS, GERMAN and CelebA datasets in Fig. 10-12.



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

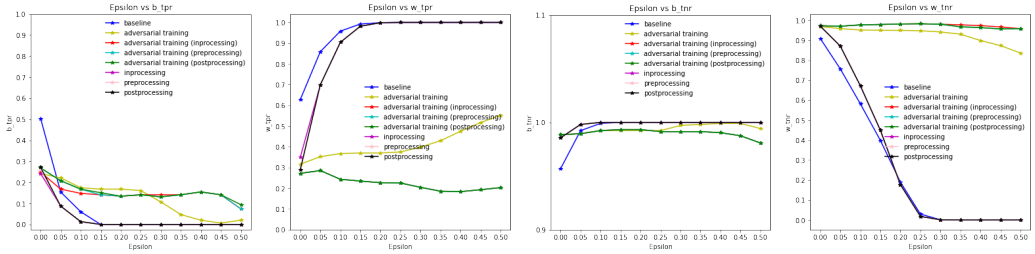
(d) EOd (German) (e) DI (German) (f) Accuracy (German)

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

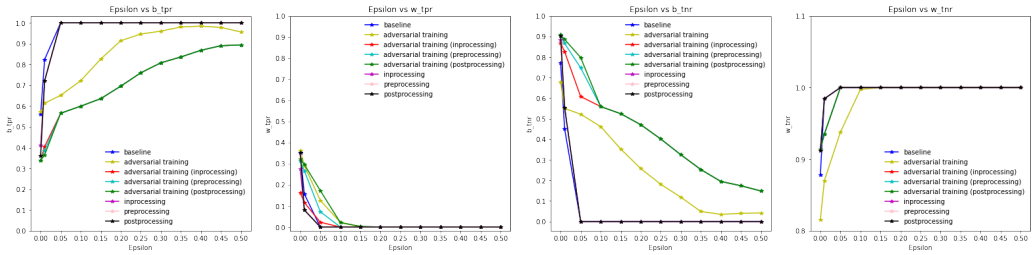
Figure 8: Change in accuracy, DI and EOd under DI attack on COMPAS, German and CelebA datasets. Our adversarial training methods (preprocessing, in-processing, post-processing) obtain improved fairness (lower EOd and DI) and higher accuracy with significant margin.

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

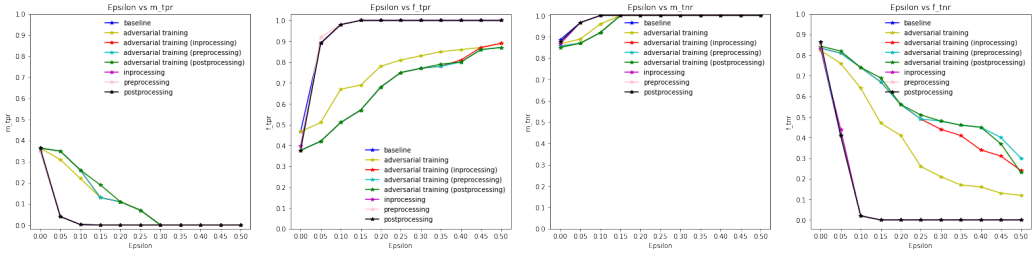
Table 3: results of accuracy for adversarial fair training on Adult dataset under DI attack.



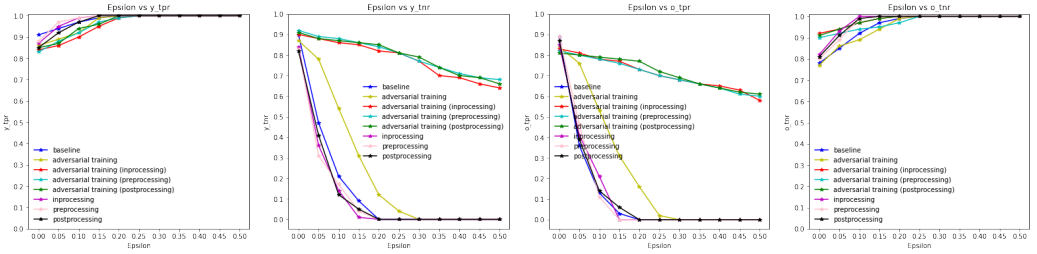
(a) Black TPR (Adult) (b) White TPR (Adult) (c) Black TNR (Adult) (d) White TNR (Adult)



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



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



(m) Young TPR (CelebA) (n) Young TNR (CelebA) (o) Elder TPR (CelebA) (p) Elder TNR (CelebA)

Figure 9: Change of true positive rate (TPR) and true negative rate (TNR) under DI attack on the four datasets.

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 4: results of EOd for adversarial fair training on Adult dataset under DI 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 5: results of DI for adversarial fair training on Adult dataset under DI 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 6: results of white TPR for adversarial fair training on Adult dataset under DI 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 7: results of white TNR for adversarial fair training on Adult dataset under DI 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 8: results of black TPR for adversarial fair training on Adult dataset under DI 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 9: results of black TNR for adversarial fair training on Adult dataset under DI 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 10: results of accuracy for adversarial fair training on COMPAS dataset under DI attack.

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 11: results of EOd for adversarial fair training on COMPAS dataset under DI 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	0.37	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 12: results of DI for adversarial fair training on COMPAS dataset under DI attack.

M	adv+pre	adv+in	adv+post
0.000	0.31	0.336	0.3
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 13: results of white TPR for adversarial fair training on COMPAS dataset under DI 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 14: results of white TNR for adversarial fair training on COMPAS dataset under DI 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 15: results of black TPR for adversarial fair training on COMPAS dataset under DI 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	0.35	0.37	0.35
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 16: results of black TNR for adversarial fair training on COMPAS dataset under DI 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 17: results of accuracy for adversarial fair training on German dataset under DI 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 18: results of EOd for adversarial fair training on German dataset under DI 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 19: results of DI for adversarial fair training on German dataset under DI 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 20: results of male TPR for adversarial fair training on German dataset under DI 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 21: results of male TNR for adversarial fair training on German dataset under DI 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 22: results of female TPR for adversarial fair training on German dataset under DI 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	0.30	0.370
0.500	0.300	0.240	0.230

Table 23: results of female TNR for adversarial fair training on German dataset under DI attack.

M	adv+pre	adv+in	adv+post
0.000	0.870	0.840	0.860
0.050	0.850	0.830	0.840
0.100	0.830	0.810	0.810
0.150	0.810	0.790	0.800
0.200	0.760	0.770	0.760
0.250	0.740	0.740	0.740
0.300	0.700	0.710	0.710
0.350	0.680	0.690	0.680
0.400	0.670	0.650	0.660
0.450	0.640	0.630	0.620
0.500	0.620	0.600	0.590

Table 24: results of accuracy for adversarial fair training on CelebA dataset under DI attack.

M	adv+pre	adv+in	adv+post
0.000	0.030	0.030	0.040
0.050	0.110	0.110	0.130
0.100	0.200	0.230	0.250
0.150	0.300	0.30	0.30
0.200	0.390	0.440	0.380
0.250	0.490	0.490	0.470
0.300	0.550	0.550	0.520
0.350	0.600	0.640	0.600
0.400	0.650	0.660	0.660
0.450	0.700	0.710	0.690
0.500	0.720	0.780	0.730

Table 25: results of EOd for adversarial fair training on CelebA dataset under DI attack.

M	adv+pre	adv+in	adv+post
0.000	0.050	0.070	0.040
0.050	0.090	0.110	0.110
0.100	0.160	0.114	0.130
0.150	0.230	0.210	0.210
0.200	0.260	0.290	0.280
0.250	0.300	0.330	0.340
0.300	0.340	0.360	0.390
0.350	0.390	0.380	0.410
0.400	0.400	0.420	0.440
0.450	0.410	0.440	0.450
0.500	0.440	0.450	0.470

Table 26: results of DI for adversarial fair training on CelebA dataset under DI attack.

M	adv+pre	adv+in	adv+post
0.000	0.830	0.840	0.850
0.050	0.880	0.860	0.870
0.100	0.920	0.900	0.940
0.150	0.970	0.950	0.960
0.200	0.990	0.990	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 27: results of young TPR for adversarial fair training on CelebA dataset under DI attack.

M	adv+pre	adv+in	adv+post
0.000	0.920	0.900	0.910
0.050	0.890	0.880	0.880
0.100	0.880	0.860	0.870
0.150	0.860	0.850	0.860
0.200	0.840	0.820	0.850
0.250	0.810	0.810	0.810
0.300	0.770	0.770	0.790
0.350	0.740	0.700	0.740
0.400	0.710	0.690	0.700
0.450	0.690	0.660	0.690
0.500	0.680	0.640	0.660

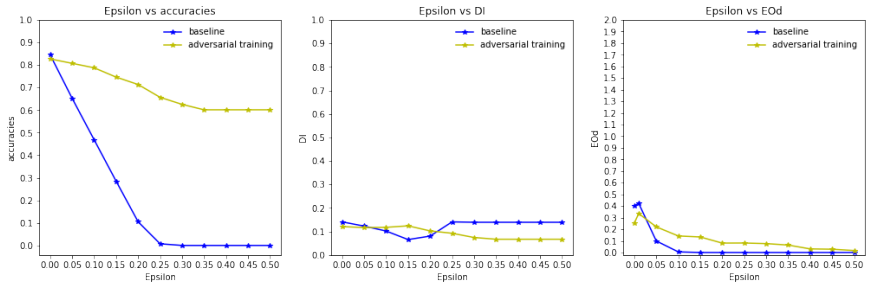
Table 28: results of young TNR for adversarial fair training on CelebA dataset under DI attack.

M	adv+pre	adv+in	adv+post
0.000	0.820	0.830	0.810
0.050	0.800	0.810	0.800
0.100	0.780	0.780	0.790
0.150	0.760	0.770	0.780
0.200	0.730	0.730	0.770
0.250	0.700	0.700	0.720
0.300	0.680	0.680	0.690
0.350	0.660	0.660	0.660
0.400	0.640	0.650	0.640
0.450	0.610	0.630	0.620
0.500	0.600	0.580	0.610

Table 29: results of elder TPR for adversarial fair training on CelebA dataset under DI attack.

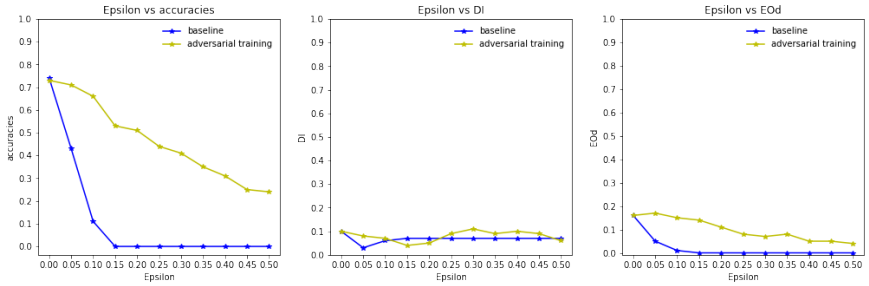
M	adv+pre	adv+in	adv+post
0.000	0.900	0.920	0.910
0.050	0.920	0.940	0.940
0.100	0.940	0.970	0.970
0.150	0.950	0.990	0.990
0.200	0.970	1.000	1.000
0.250	1.000	1.000	1.000
0.300	1.000	1.000	1.000
0.350	1.000	1.000	1.000
0.400	1.000	1.000	1.000
0.450	1.000	1.000	1.000
0.500	1.000	1.000	1.000

Table 30: results of elder TNR for adversarial fair training on CelebA dataset under DI attack.



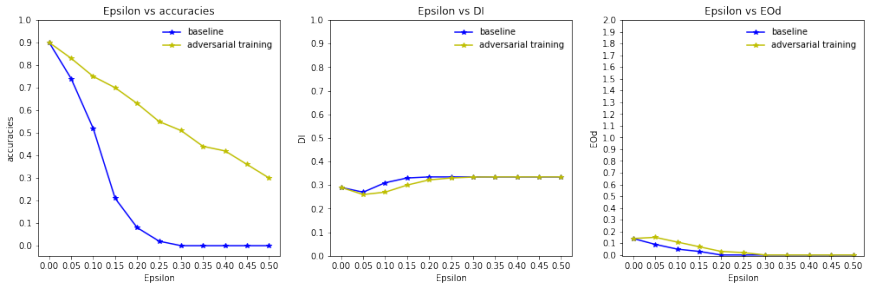
(a) Accuracy (b) DI (c) EOd

Figure 10: Results of a classifier adversarially trained w.r.t. DI. Change of accuracy, DI and EOd under accuracy attack on COMPAS dataset.



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

Figure 11: Results of a classifier adversarially trained w.r.t. DI. Change of accuracy, DI and EOd under accuracy attack on German dataset.



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

Figure 12: Results of a classifier adversarially trained w.r.t. DI. Change of accuracy, DI and EOd under accuracy attack on CelebA dataset.