A. Appendix

Appendix A.1 proves Theorem 1. Appendix A.2 extends the theoretical results of the fairness discriminator to other measures. Appendix A.3 shows additional experiments. Appendix A.4 provides more details of the model training setup.

A.1. Proof for Theorem 1

Before we present the proof of the main theorem, we first recall our notation. Let $P_Z(z)$ be the distribution of Z where $z \in \mathcal{Z}$ and $\mathcal{Z}$ is the set of possible sensitive attribute values. Let $\hat{Y}|Z = z \sim P_{\hat{Y}|z}(\cdot)$ and $\hat{Y} \sim P_{\hat{Y}}(\cdot)$. Then $P_{\hat{Y}}(\cdot) = \sum_{z\in\mathcal{Z}} P_Z(z)P_{\hat{Y}|z}(\cdot)$. Also, let $Y \sim P_Y(\cdot)$.

For convenience, let us repeat the statement of Theorem 1 here:

$$\begin{aligned} &I(Z;\hat{Y}) = \\ &\max_{D_z(\hat{y}):\sum_z D_z(\hat{y})=1,\ \forall \hat{y}} \sum_{z\in\mathcal{Z}} P_Z(z)\mathbb{E}_{P_{\hat{Y}|z}}\left[\log D_z(\hat{Y})\right] \\ &+ H(Z). \end{aligned}$$

We now prove the theorem.

Proof. Denote by $\boldsymbol{D}$ the collection of $D_z(\hat{y})$ for all possible values of z and $\hat{y}$, and by $\boldsymbol{\nu}$ the collection of $\nu_{\hat{y}}$ for all values of $\hat{y}$. We can construct the Lagrangian function as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{D},\boldsymbol{\nu}) = &\sum_{z\in\mathcal{Z}} P_Z(z)\mathbb{E}_{P_{\hat{Y}|z}}\left[\log D_z(\hat{Y})\right] + H(Z) \\ &+ \sum_{\hat{y}\in\mathcal{Y}} \nu_{\hat{y}} \left(1 - \sum_{z\in\mathcal{Z}} D_z(\hat{y})\right). \end{aligned}$$

We use the following KKT conditions:

$$\begin{aligned} &\frac{\partial\mathcal{L}(\boldsymbol{D},\boldsymbol{\nu})}{\partial D_z(\hat{y})} = P_Z(z)\frac{P_{\hat{Y}|z}(\hat{y})}{D_z^\star(\hat{y})} - \nu_{\hat{y}}^\star = 0, && \forall(\hat{y},z)\in\mathcal{Y}\times\mathcal{Z}, \\ &1 - \sum_{z\in\mathcal{Z}} D_z^\star(\hat{y}) = 0, && \forall\hat{y}\in\mathcal{Y}. \end{aligned}$$

Solving the two equations, we obtain $\nu_{\hat{y}}^\star = P_{\hat{Y}}(\hat{y})$ for all $\hat{y}$. Thus,

$$D_z^\star(\hat{y}) = \frac{P_Z(z)P_{\hat{Y}|z}(\hat{y})}{P_{\hat{Y}}(\hat{y})}.$$

Putting this to the above optimization,

$$\begin{aligned} &\sum_{z\in\mathcal{Z}} P_Z(z)\mathbb{E}_{P_{\hat{Y}|z}}\left[\log\frac{P_Z(z)P_{\hat{Y}|z}(\hat{Y})}{P_{\hat{Y}}(\hat{Y})}\right] + H(Z) \\ &= \sum_{z\in\mathcal{Z}} P_Z(z)\mathbb{E}_{P_{\hat{Y}|z}}\left[\log\frac{P_Z(z)P_{\hat{Y}|z}(\hat{Y})}{P_{\hat{Y}}(\hat{Y})}\right] \\ &\quad + \sum_{z\in\mathcal{Z}} P_Z(z)\log\frac{1}{P_Z(z)} \\ &= \sum_{z\in\mathcal{Z}} P_Z(z)\mathbb{E}_{P_{\hat{Y}|z}}\left[\log\frac{P_{\hat{Y}|z}(\hat{Y})}{P_{\hat{Y}}(\hat{Y})}\right] \\ &= \sum_{z\in\mathcal{Z}} P_Z(z)D_{\mathrm{KL}}(P_{\hat{Y}|z}\|P_{\hat{Y}}) \\ &\triangleq \mathrm{JS}_{P_Z}(P_{\hat{Y}|z_1},\ldots,P_{\hat{Y}|z_{|\mathcal{Z}|}}) = I(Z;\hat{Y}). \end{aligned}$$

Here, the second last equality is due to the definition of the generalized Jensen-Shannon divergence, and the last equality is due to its equivalence to the mutual information (Lin, 1991). □

A.2. Extensions to other fairness measures

We now extend FR-Train to the case of *equalized odds*, which is another important fairness metric, defined as follows:

Definition 2. *(Equalized Odds)*
$P(\hat{Y} = y|Y = y, Z = z_1) = P(\hat{Y} = y|Y = y, Z = z_2)$, $\forall y \in \mathcal{Y},\ \forall z_1, z_2 \in \mathcal{Z}$.

The following theorem relates the conditional mutual information $I(Z;\hat{Y}|Y)$ to the solution of an optimization problem.

Theorem 3. $I(Z;\hat{Y}|Y) =$
$\max_{D_{z|y}(\hat{y}):\sum_{z\in\mathcal{Z}} D_{z|y}(\hat{y})=1,\ \forall\hat{y}}$
$\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} P_{Y,Z}(y,z)\mathbb{E}_{P_{\hat{Y}|y,z}}\left[\log D_{z|y}(\hat{Y})\right] + H(Z|Y).$

This conditional mutual information term can be used to capture *equalized odds*. We also note that the following theorem can be modified in a straightforward manner so that it can handle $I(Z;\hat{Y}|Y=1)$, which can be used to capture *equal opportunity*.

We now prove the theorem.

Proof. Denote by $\boldsymbol{D}$ the collection of $D_{z|y}(\hat{y})$ for all possible values of (z, $\hat{y}$, and y) and by $\boldsymbol{\nu}$ the collection of $\nu_{y,\hat{y}}$ for all values of y and $\hat{y}$. We can construct the Lagrangian function as follows:

$$\mathcal{L}(\boldsymbol{D}, \boldsymbol{\nu}) = \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P_{Y,Z}(y,z) \mathbb{E}_{P_{\hat{Y}|y,z}} \left[\log D_{z|y}(\hat{Y}) \right] + H(Z|Y) + \sum_{\hat{y} \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \nu_{y,\hat{y}} \left(1 - \sum_{z \in \mathcal{Z}} D_{z|y}(\hat{y}) \right).$$

We use the following KKT conditions:

$$\frac{\partial \mathcal{L}(\boldsymbol{D}, \boldsymbol{\nu})}{\partial D_{z|y}(\hat{y})} = P_{Y,Z}(y,z) \frac{P_{\hat{Y}|y,z}(\hat{y})}{D^{\star}_{z|y}(\hat{y})} - \nu_{y,\hat{y}} = 0, \quad \forall (\hat{y}, y, z) \in \mathcal{Y} \times \mathcal{Y} \times \mathcal{Z}$$

$$1 - \sum_{z \in \mathcal{Z}} D^{\star}_{z|y}(\hat{y}) = 0, \ \forall (\hat{y}, y) \in \mathcal{Y} \times \mathcal{Y}.$$

Solving the two equations, we obtain $\nu^{\star}_{y,\hat{y}} = P_{Y,\hat{Y}}(y, \hat{y})$ for all $(y, \hat{y}) \in \mathcal{Y} \times \mathcal{Y}$. Thus,

$$D^{\star}_{z|y}(\hat{y}) = \frac{P_{Z|y}(z) P_{\hat{Y}|y,z}(\hat{y})}{P_{\hat{Y}|y}(\hat{y})}, \ \forall y, \hat{y} \in \mathcal{Y} \times \mathcal{Y}.$$

Putting this to the above optimization,

$$\begin{aligned}
&\sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P_{Y,Z}(y,z) \mathbb{E}_{P_{\hat{Y}|y,z}} \left[\log \frac{P_{Z|y}(z) P_{\hat{Y}|y,z}(\hat{Y})}{P_{\hat{Y}|y}(\hat{Y})} \right] \\
&+ H(Z|Y) \\
=& \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P_{Y,Z}(y,z) \mathbb{E}_{P_{\hat{Y}|y,z}} \left[\log \frac{P_{Z|y}(z) P_{\hat{Y}|y,z}(\hat{Y})}{P_{\hat{Y}|y}(\hat{Y})} \right] \\
&+ \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P_{Y,Z}(y,z) \log \frac{1}{P_{Z|y}(z)} \\
=& \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P_{Y,Z}(y,z) \mathbb{E}_{P_{\hat{Y}|y,z}} \left[\log \frac{P_{\hat{Y}|y,z}(\hat{Y})}{P_{\hat{Y}|y}(\hat{Y})} \right] \\
=& \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P_{Y}(y) P_{Z|y}(z) \mathbb{E}_{P_{\hat{Y}|y,z}} \left[\log \frac{P_{\hat{Y}|y,z}(\hat{Y})}{P_{\hat{Y}|y}(\hat{Y})} \right] \\
=& \sum_{y \in \mathcal{Y}} P_{Y}(y) \sum_{z \in \mathcal{Z}} P_{Z|y}(z) \mathbb{E}_{P_{\hat{Y}|y,z}} \left[\log \frac{P_{\hat{Y}|y,z}(\hat{Y})}{P_{\hat{Y}|y}(\hat{Y})} \right] \\
=& \sum_{y \in \mathcal{Y}} P_{Y}(y) \sum_{z \in \mathcal{Z}} P_{Z|y}(z) D_{\mathrm{KL}}(P_{\hat{Y}|y,z} \| P_{\hat{Y}|y}) \\
\triangleq& \sum_{y \in \mathcal{Y}} P_{Y}(y) \cdot \mathrm{JS}_{P_{Z|y}}(P_{\hat{Y}|z_1,y}, \ldots, P_{\hat{Y}|z_{|\mathcal{Z}|},y}) \\
=& \sum_{y \in \mathcal{Y}} P_{Y}(y) I(Z; \hat{Y} | Y = y) = I(Z; \hat{Y} | Y).
\end{aligned}$$

The third last equality is due to the definition of the generalized Jensen-Shannon divergence; the second last equality is due to its equivalence to the mutual information (Lin, 1991); and the last equality is due to the definition of conditional mutual information. □

We now discuss how to actually compute the mutual information. We compute the following empirical version using the examples $\{(x^{(i)}, z^{(i)}, y^{(i)})\}_{i=1}^{m}$.

$$\max_{D_{z|y}(\hat{y}): \sum_{z \in \mathcal{Z}} D_{z|y}(\hat{y}) = 1; \forall \hat{y}} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P_{Y,Z}(y,z) \sum_{i:(y^{(i)}, z^{(i)}) = (y,z)} \frac{1}{m_{y,z}} \log D_{z|y}(\hat{y}^{(i)}) + H(Z|Y).$$

Now for a sufficiently large value of m, $m_{y,z} \approx P_{Y,Z}(y,z)m$. Therefore, the above expression is approximated as:

$$\max_{D_{z|y}(\hat{y}): \sum_{z \in \mathcal{Z}} D_{z|y}(\hat{y}) = 1; \forall \hat{y}} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} \sum_{i:(y^{(i)}, z^{(i)}) = (y,z)} \frac{1}{m} \log D_{z|y}(\hat{y}^{(i)}) + H(Z|Y).$$

Hence, we can set L_2 (i.e., the loss w.r.t. the fairness discriminator) to the above expression. The rest of the objective function is the same. Figure 5 shows the resulting FR-Train architecture.

A.3. Additional experiments

A.3.1. Synthetic data

We continue our experiments from Section 4.1. In particular, we perform FR-Train with different amounts of poisoning, and evaluate robust training with meta learning using smaller validation sets.

FR-Train with different amounts of poisoning Table 9 shows FR-Train performances with the different levels of poisoning. Even on the heavily poisoned (say 40%) data, FR-Train shows marginal performance degradations (< 6.5% decrease in DI).

Table 9. Accuracy and fairness performances of FR-Train on the poisoned synthetic test datasets for different amount of poisoning. We used the same label poisoning attack described in Section 2.

Data	Poisoning amount	DI	Accuracy
Clean	0%	0.818	0.807
Poisoned	10%	0.827	0.814
	15%	0.813	0.800
	20%	0.802	0.800
	25%	0.784	0.803
	30%	0.780	0.800
	35%	0.770	0.802
	40%	0.765	0.806

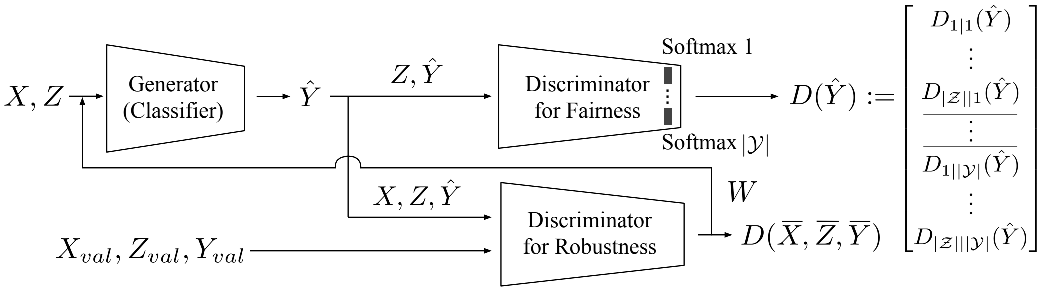

X, Z
Generator (Classifier)
$\hat{Y}$
$Z, \hat{Y}$
Discriminator for Fairness
Softmax 1
Softmax $|\mathcal{Y}|$
$D(\hat{Y}) :=$
$D_{1|1}(\hat{Y})$
$D_{|\mathcal{Z}||1}(\hat{Y})$
$D_{1||\mathcal{Y}|}(\hat{Y})$
$D_{|\mathcal{Z}|||\mathcal{Y}|}(\hat{Y})$
$X, Z, \hat{Y}$
W
$X_{val}, Z_{val}, Y_{val}$
Discriminator for Robustness
$D(\overline{X}, \overline{Z}, \overline{Y})$

Figure 5. The architecture of FR-Train for equalized odds.

Meta learning with different validation set sizes Table 10 shows the accuracy and fairness results for RML for different validation set sizes. We observe a drastic decrease of accuracy and fairness when the validation set size is 0.1% of the training data.

Table 10. Accuracy and fairness performances of the meta learning method by (Ren et al., 2018) on the clean and poisoned synthetic test datasets for different validation set sizes. We used the same label poisoning attack described in Section 2, and the amount of poisoning is 10% of $\mathcal{D}_{tr}$.

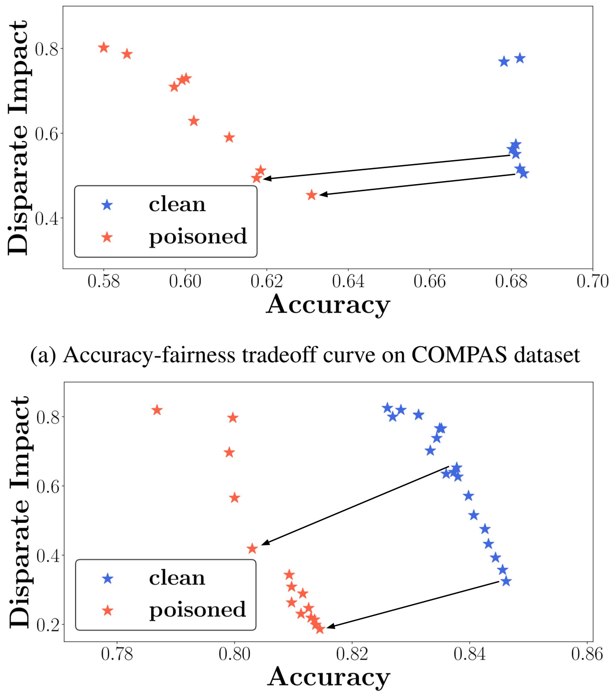
Data	Val. set size	Disparate impact	Accuracy
Clean	10%	0.429	0.883
Poisoned	10%	0.395	0.869
	5%	0.378	0.852
	0.5%	0.290	0.830
	0.1%	0.098	0.714

A.3.2. Real data

We continue our experiments from Sections 2 and 4.2.

Fairness Constraints on real datasets We show the accuracy-fairness tradeoffs of Fairness Constraints (Zafar et al., 2017) on the COMPAS and AdultCensus datasets. Figures 6a and 6b show that both accuracy and fairness of Fairness Constraints decrease on the poisoned data, showing a strictly-worse tradeoff.

Training with only validation set We evaluate the baseline that simply trains fairness algorithms on the clean validation set. Table 11 shows that the baseline performs worse than those in Tables 2 and 3. For example, training FC on the AdultCensus crowdsourced validation set yields (DI, Acc) = (0.756, 0.761), which is worse than the FC baseline result (DI, Acc) = (0.826, 0.825) as shown in Table 3. We thus observe that the validation set is sufficient to help discern clean and poisoned data in FR-Train, but not large enough for algorithms to obtain high performance.



(a) Accuracy-fairness tradeoff curve on COMPAS dataset

(b) Accuracy-fairness tradeoff curve on AdultCensus dataset

Figure 6. Accuracy-fairness tradeoff curves of Fairness Constraints on real datasets.

FR-Train using other fairness measures As we showed in Appendix A.2, FR-Train respects equalized odds and equal opportunity. Table 12 shows the experimental results on the synthetic and real datasets for equalized odds. We see that FR-Train significantly improves equalized odds with reasonable accuracy. The results w.r.t. equal opportunity are similar and thus not shown here.

A.4. Training methodology

The generator G is a neural network with zero or one hidden layer. The discriminator D^f is a single layer neural network, and the discriminator D^r is a neural network with one hidden layer. We used 8 or 16 nodes in the hidden layers. We set an Adam optimizer (Kingma & Ba, 2014)

Table 11. Accuracy and fairness performances of the baseline that trains with only validation set. We use the same validation sets utilized in FR-Train.

Method	COMPAS		AdultCensus	
	DI	Acc.	DI	Acc.
FC	0.796	0.647	0.761	0.756
LBC	0.796	0.647	0.795	0.799
AD	0.762	0.646	0.682	0.693

Table 12. Accuracy and fairness performances on synthetic and real test datasets w.r.t. equalized odds. Two algorithms are compared: (1) LR (non-fairness method) and (2) FR-Train.

Dataset	Method	Equalized odds		Accuracy
		$Y = 0$	$Y = 1$	
Synthetic Data	LR	0.351	0.804	0.885
	FR-Train	0.888	0.936	0.865
COMPAS	LR	0.427	0.557	0.674
	FR-Train	0.718	0.959	0.628
AdultCensus	LR	0.286	0.909	0.848
	FR-Train	0.503	0.917	0.842

for the generator, and a stochastic gradient descent (SGD) optimizer for each discriminator. We empirically observe that one can stabilize the training procedure by freezing the parameters of the fairness discriminator D^f for the initial phase of training. Thus, we choose to freeze the parameters of the fairness discriminator D^f for the first few epochs until the generator achieves a certain accuracy. We pre-train the generator for the first few epochs and use the generator/discriminator update ratio of 1:3 (or 1:5) for the rest of training.

Also, we use the following details for choosing the values of λ_1, λ_2, and C. For clean data, we set λ_2 as a small value (e.g., 0.1) and vary λ_1 from 0 to 0.85. For poisoned data, we set λ_2 as 0.2, 0.3, or 0.4, and vary λ_1 from 0 to $0.95 - \lambda_2$. Given the values of λ_1 and λ_2, we also normalize L_1 (the generator loss) by multiplying it with $(1 - \lambda_1 - \lambda_2)$. We set C to be a value between 0 to 3.