A Fair Classifier Using Mutual Information

Jaewoong Cho
EE, KAIST
Email: cjw2525@kaist.ac.kr

Gyeongjo Hwang
EE, KAIST
Email: hkj4276@kaist.ac.kr

Changho Suh
EE, KAIST
Email: chsuh@kaist.ac.kr

***Abstract*—As machine learning becomes prevalent in our daily lives involving a widening array of applications such as medicine, finance, job hiring and criminal justice, one morally & legally motivated need for machine learning algorithms is to ensure fairness for disadvantageous against advantageous groups. Fairness in machine learning aims at guaranteeing the irrelevancy of a prediction output to sensitive attributes like race, sex and religion. To this end, we take an information-theoretic approach using mutual information (MI) which can fully capture such independence. Inspired by the fact that MI between prediction and the sensitive attribute being zero is the "sufficient and necessary condition" for independence, we develop an MI-based algorithm that well trades off prediction accuracy for fairness performance often quantified as Disparate Impact (DI) or Equalized Odds (EO). Our experiments both on synthetic and benchmark real datasets demonstrate that our algorithm outperforms prior fair classifiers in tradeoff performance both w.r.t. DI and EO.**

I. Introduction

The last decade has seen an unprecedented explosion of academic and popular interests in machine learning. Machine learning is no longer just an engine behind image classifiers and spam filters. It is now employed to make critical decisions that affect human lives, cultures, and rights, e.g., filtering job applicants, and informing bail & parole decisions. With a surge of such sensitive applications, one major criterion in the design of machine learning algorithms is to ensure *fairness*.

Fairness research has developed metrics that capture various notions of discrimination. There are largely two types of measures: (i) group fairness [1], [2], [3], [4], [5], which ensures similar statistics between different demographics; (ii) individual fairness [6], [7], [8], [9], which guarantees similar prediction results across nearby examples. In this work, we focus on the former that has been widely studied in various applications. Major measures for group fairness include disparate impact [1], [4], equal opportunity [3] and equalized odds [3]. All of these capture a degree of fairness by quantifying how prediction outputs are different depending on sensitive attributes such as race, sex, age and religion. For instance, the ratio of prediction outputs (e.g., criminal reoffending vs. non-reoffending probabilities) w.r.t. a sensitive attribute (e.g., blacks vs. whites) is one prominent fairness metric.

There has been a proliferation of fairness algorithms [3], [4], [6], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] which intend to prevent discrimination. It turns out it is non-trivial to express a fairness measure in terms of model parameters (often neural-net parameters) which serve as optimization variables in an associated optimization. Hence, one prominent approach is to introduce an expressible *proxy* for fairness measure and then incorporate the proxy into formulating a regularization term in the optimization problem. For example, Zafar et.al. [4] employ as a fairness proxy a covariance function between a sensitive attribute and a prediction output. However, such proxy-based approach comes with a challenge in enforcing fairness constraints. This is because the proxy serves only as a *weak* constraint. For instance, a small covariance between a sensitive attribute and a prediction output guarantees a small correlation, but the small correlation does not ensure *independence* that fairness aims at. Hence, such method does not provide any theoretical guarantee for many practical scenarios in which real data may not admit the one-to-one relationship between uncorrelatedness and independence.

In order to overcome the above challenge, we take a different approach which enables us to fully respect fairness constraints and therefore to achieve the optimal fairness performance given a certain accuracy and vice versa. As an initial effort, we focus on two prominent group fairness measures: disparate impact (DI) [4] and equalized odds (EO) [3][1]. These constraints target the complete independence between prediction and a sensitive attribute unconditionally (disparate impact) or conditionally (equalized odds). Our approach employs one key information-theoretic notion, mutual information (MI), to fully capture such independence. The approach is inspired by the fact that MI between prediction and the sensitive attribute being zero is the "sufficient and necessary condition" for independence. We then establish mathematical relationships which enable tractability of the associated MI computation. With this theory, we propose an efficient fair classifier architecture that is easily trainable with well-known optimizers such as stochastic gradient descent (SGD) and Adam optimizer [22], [23]. We also find an intimate connection with generative adversarial networks (GANs) [16] as well as with a GAN-based fair classifier [24]. Our experiments conducted both on synthetic and benchmark real datasets (COMPAS [25] and Adult Census [26]) corroborate our theory, demonstrating that our MI-based fair classifier obtains the state-of-the-art tradeoff performances both w.r.t. DI and EO.

Related work: Among the fairness algorithms [3], [4], [6], [10], [11], [12], [13], [14], [15], [16], [19], [20], [21], we point out two works [15], [24] that are relevant to ours

[1]We also cover another fairness measure, called equal opportunity [3], as it is a special case of equalized odds that emphasizes positive examples.

978-1-7281-6432-8/20/$31.00 ©2020 IEEE

ISIT 2020

in architecture. Xu et.al. [15] adopt an adversarial training idea [16] to generate fairness-ensured fake data. While their optimization bears a structural similarity with ours, it belongs to a distinct framework, as it targets generating fairness-ensured fake data, instead of developing a fair classifier that we aim at here. The second related work is Zhang et.al. [24]. This is also inspired by adversarial training [16]: the classifier is designed such that an adversary cannot discriminate a sensitive attribute from a prediction output. It takes an insightful yet heuristic procedure in training, thereby preventing it from being interpreted precisely as an MI-based one; see Remark 2 for details. This way, no theoretical guarantee can be made for many interested scenarios. Further it suffers from a severer stability issue in training compared to ours; also see Fig. 4.

II. Problem Formulation

Fig. 1 illustrates a supervised learning fair-classifier setting that we focus on herein. There are two types of data employed:

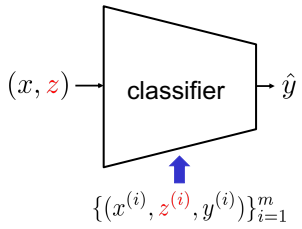

Fig. 1. A fair classifier that attempts to yield a prediction output $\hat{y}$ from normal data $x \in \mathbf{R}^d$ and sensitive attribute z so that $\hat{y}$ is independent of sensitive attribute z as much as possible. Here $\{(x^{(i)}, z^{(i)}, y^{(i)})\}_{i=1}^m$ indicate the ith data, sensitive attribute and label, respectively and m denotes the number of examples.

(i) normal (possibly objective) data; (ii) *sensitive* data (or called *sensitive attributes*). We denote the normal data by $x \in \mathbf{R}^d$. In the case of recidivism score prediction [25], such x may refer to a collection of the number of prior criminal records and a criminal type, e.g., misdemeanour or felony. For *sensitive* data, we employ a different notation, say z. In the above example, z may indicate a race type among white ($z = 1$) and black ($z = 0$). In general, the alphabet size of z is arbitrary. For instance, there are many race types such as Black, White, Asian, Hispanic, to name a few. Also there could be multiple sensitive attributes like gender and religion. In order to reflect such practical scenarios, we consider $z \in \mathcal{Z}$ with an arbitrary alphabet size that can represent a collection of possibly many sensitive attributes. Let $\hat{y}$ be the classifier output which aims to represent the ground-truth conditional distribution $p(y|x, z)$. Here $y \in \mathcal{Y}$ denotes the ground-truth label. In the recidivism score prediction case, $y = 1$ means reoffending in the near future, say within two years ($y = 0$ otherwise), while $\hat{y}$ indicates the probability of such event being occurred. We consider a supervised learning setup, so we are given m example triplets: $\{(x^{(i)}, z^{(i)}, y^{(i)})\}_{i=1}^m$. We assume that both (x, z) serve as the input, although z may not be part of the input in an effort to automatically respect the constraint w.r.t. disparate treatment [4] (another group fairness measure capturing an unequal treatment that occurs directly because of sensitive attributes). This is because our algorithm (to be presented) does care about the fairness issue while being designed.

We consider two major group fairness measures: disparate impact (DI) and equalized odds (EO). Mathematical definitions of DI and EO rely on a few notations. Let $Z \in \mathcal{Z}$ be a random variable that indicates a sensitive attribute. Let $\tilde{Y} \in \mathcal{Y}$ be a hard-decision value of the predictor output $\hat{Y}$ at certain thresholds. For instance, in the binary classifier case, $\tilde{Y} := \mathbf{1}\{\hat{Y} \geq 0.5\}$. The DI is then defined as a ratio of likelihoods of positive example events $\tilde{Y} = 1$ for different $Z = z$:

$$\mathsf{DI} := \min_{z \in \mathcal{Z}} \min_{\hat{z} \neq z} \frac{\Pr(\tilde{Y} = 1 | Z = \hat{z})}{\Pr(\tilde{Y} = 1 | Z = z)}. \tag{1}$$

One natural interpretation is that a classifier is more fair when the ratio is closer to 1; becomes unfair if the ratio is far away from 1. So the DI quantifies such fairness degree: the larger DI, the more fair the situation is. On the other hand, the EO is defined as the same ratio except that the ground-truth label $Y = y$ is given:

$$\mathsf{EO} := \min_{y \in \mathcal{Y}} \min_{z \in \mathcal{Z}} \min_{\hat{z} \neq z} \frac{\Pr(\tilde{Y} = 1 | Z = \hat{z}, Y = y)}{\Pr(\tilde{Y} = 1 | Z = z, Y = y)}. \tag{2}$$

Now how to design a classifier so as to maximize the DI or EO? One natural approach is to incorporate a fairness-related constraint as a *regularization* term into the conventional classifier optimization which often takes the following form:

$$\min_w \frac{1}{m} \sum_{i=1}^m \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) \tag{3}$$

where $\ell_{\mathsf{CE}}(y, \hat{y}) := -\sum_j y_j \log \hat{y}_j$ indicates cross entropy loss [22], and w denotes weights (parameters) for a classifer. Here we assume that the label y is of one-hot-vector type: $y := [y_1, \ldots, y_{|\mathcal{Y}|}]^T$. Since maximizing DI is equivalent to minimizing $1 - \mathsf{DI}$ (due to $0 \leq \mathsf{DI} \leq 1$), an interested optimization reads:

$$\min_w \frac{1}{m} \sum_{i=1}^m \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) + \lambda \cdot (1 - \mathsf{DI}) \tag{4}$$

where λ denotes a regularization factor that balances predication accuracy against the DI-associated objective (minimizing $1 - \mathsf{DI}$). Similarly one can formulate an EO-associated optimization by replacing $1 - \mathsf{DI}$ with $1 - \mathsf{EO}$.

Here a challenge arises in solving the regularized optimization (4). Recalling the definition (1) of DI, we see that DI is a complicated function of w. We have no idea as to how to express DI in terms of w. One effort to address this challenge was made by Zafar et.al. [4]. They introduce an easily-expressible *proxy* for the fairness measure, reflected in $1-\mathsf{DI}$. Specifically they employ a covariance function between $\hat{Y}$ and Z. However, this proxy serves only as a *weak* constraint because a small covariance does not necessarily imply the independence although the reverse always hold. In this work,

we introduce a different regularization term that can serve as a *strong* constraint for the independence. Details will be presented in Section III for DI, and in Section IV for EO.

III. Disparate Impact

Our approach is inspired by mutual information (MI). Notice that $\mathsf{DI} = 1$ means that the sensitive attribute Z is independent of the hard decision $\tilde{Y}$ of the prediction. One key property of MI is that MI between two input random varibles being zero is the "sufficient and necessary condition" for the independence between the two inputs. This motivates us to represent the constraint of $\mathsf{DI} = 1$ as $I(Z;\tilde{Y}) = 0$. This captures the complete independence between Z and $\tilde{Y}$. Since the predictor output is $\hat{Y}$ (instead of $\tilde{Y}$), we consider another stronger condition that concerns $\hat{Y}$ directly:

$$I(Z;\hat{Y}) = 0. \tag{5}$$

Notice that the condition (5) is indeed stronger than $I(Z;\tilde{Y}) = 0$ because $I(Z;\tilde{Y}) \leq I(Z;\tilde{Y},\hat{Y}) = I(Z;\hat{Y})$ where the equality comes from the fact that $\tilde{Y}$ is a function of $\hat{Y}$. Notice that this together with (5) gives $I(Z;\tilde{Y}) = 0$. Hence, the condition (5) enforces the $\mathsf{DI} = 1$ constraint.

This motivates us to consider the following optimization:

$$\min_{w} \frac{1}{m}\sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)},\hat{y}^{(i)}) + \lambda \cdot I(Z;\hat{Y}). \tag{6}$$

Now the question of interest is: How to express $I(Z;\hat{Y})$ in terms of classifier parameters w? We found an interesting way to express it. To see this, let us investigate the relationship between MI and Kullback-Leibler (KL) divergence [27]:

$$\begin{aligned} I(Z;\hat{Y}) &= D_{\mathsf{KL}}\left(P_{\hat{Y},Z} \| P_{\hat{Y}} P_Z\right) \\ &= \sum_{\hat{y},z} P_{\hat{Y},Z}(\hat{y},z) \log \frac{P_{\hat{Y},Z}(\hat{y},z)}{P_{\hat{Y}}(\hat{y}) P_Z(z)} \\ &= \sum_{\hat{y},z} P_{\hat{Y},Z}(\hat{y},z) \log \underbrace{\frac{P_{\hat{Y},Z}(\hat{y},z)}{P_{\hat{Y}}(\hat{y})}}_{=:D^*(\hat{y};z)} + H(Z). \end{aligned} \tag{7}$$

where $P_{\hat{Y}}, P_Z$ and $P_{\hat{Y},Z}$ indicate probability distributions of $\hat{Y}, Z$ and $(\hat{Y}, Z)$, respectively. Defining the log-inside term in the above as $D^*(\hat{y};z)$, we get: $\sum_z D^*(\hat{y};z) = 1 \ \ \forall \hat{y}$. Here a key trick is to represent (7) in terms of a *function optimization* such that the objective function can be expressed in terms of the classifier parameter w. In this work, we establish such trick via the following theorem:

Theorem 1 (MI via function optimization for DI*):*

$$\begin{aligned} I(Z;\hat{Y}) = H(Z) + \\ \max_{D(\hat{y};z):\sum_z D(\hat{y};z)=1} \sum_{\hat{y},z} P_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z). \end{aligned} \tag{8}$$

Proof: Using the multiple equality constraints, reflected in $\sum_z D(\hat{y};z) = 1$, we define the Lagrange function w.r.t. the max optimization in (8) as:

$$\begin{aligned} \mathcal{L}(D(\hat{y};z),\nu(\hat{y})) = &\sum_{\hat{y},z} P_{\hat{Y},Z}(\hat{y},z) \log D(\hat{y};z) \\ &+ \sum_{\hat{y}} \nu(\hat{y}) \left(1 - \sum_{z} D(\hat{y};z)\right) \end{aligned}$$

where $\nu(\hat{y})$'s are Lagrange multipliers. Since the objective function is *concave* in $D(\cdot;\cdot)$, one can solve the problem via the KKT conditions [28]:

$$\begin{aligned} \frac{d\mathcal{L}(D(\hat{y};z),\nu(\hat{y}))}{dD(\hat{y};z)} &= \frac{P_{\hat{Y},Z}(\hat{y},z)}{D^\star(\hat{y};z)} - \nu^\star(\hat{y}) = 0 \ \ \forall \hat{y}, z \\ \frac{d\mathcal{L}(D(\hat{y};z),\nu(\hat{y}))}{d\nu(\hat{y})} &= 1 - \sum_{z\in\mathcal{Z}} D^\star(\hat{y};z) = 0 \ \ \forall \hat{y}. \end{aligned}$$

Notice that plugging the following,

$$D^\star(\hat{y};z) = \frac{P_{\hat{Y},Z}(\hat{y},z)}{P_{\hat{Y}}(\hat{y})}, \qquad \nu^\star(\hat{y}) = P_{\hat{Y}}(\hat{y}).$$

we satisfy the KKT conditions. This implies that $D^\star(\hat{y},z)$ is indeed the optimal solution. Since $D^\star(\hat{y};z)$ is exactly the same as $D^*(\hat{y};z)$ that we defined in (7), we complete the proof. ■

Next let us express the formula (8) in terms of the classifier parameter w. Notice that the formula (8) contains probability quantities $P_{\hat{Y},Z}(\hat{y},z)$'s which are not available. Instead we are given m examples: $\{(x^{(i)},z^{(i)},y^{(i)})\}_{i=1}^m$. So we need to worry about how to compute the probability quantities only with such examples. Once the classifier weight w is given, one can first compute $\{\hat{y}^{(i)}\}_{i=1}^m$ from the train samples. This together with the samples allows us to compute *empirical* distributions: $\mathbb{Q}_{\hat{Y},Z}(\hat{y}^{(i)},z^{(i)})$. So as an estimate, here we take the *empirical* version of the true distributions:

$$\mathbb{Q}_{\hat{Y},Z}(\hat{y}^{(i)},z^{(i)}) = \frac{1}{m} \quad \forall i. \tag{9}$$

Putting this into (8), we then get:

$$I(Z;\hat{Y}) \approx \max_{D(\hat{y};z):\sum_z D(\hat{y};z)=1} \sum_{i=1}^{m} \frac{1}{m} \log D(\hat{y}^{(i)};z^{(i)}) + H(Z).$$

Lastly by parameterizing $D(\cdot;\cdot)$ with θ and excluding $H(Z)$ (irrelevant of (θ, w)), we obtain the following optimization:

$$\begin{aligned} \min_{w} \max_{\theta:\sum_z D_\theta(\hat{y};z)=1} \frac{1}{m} \Bigg\{ &\sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)},\hat{y}^{(i)}) \\ &+ \lambda \sum_{i=1}^{m} \log D_\theta(\hat{y}^{(i)};z^{(i)}) \Bigg\}. \end{aligned} \tag{10}$$

In view of this optimization, our fair classifier architecture can be illustrated as in Fig. 2. Notice that we have the number $|\mathcal{Z}|$ of outputs in $D_\theta(\hat{y};\cdot)$ that we call the "discriminator". In an attempt to ensure the equality constraint in (10), one can use the softmax activation in the output [22]. Here for parameterization, we use neural networks. For training (w,θ), we use alternating GD; and for training each model, we employ either GD, SGD or Adam optimizer [22], [23].

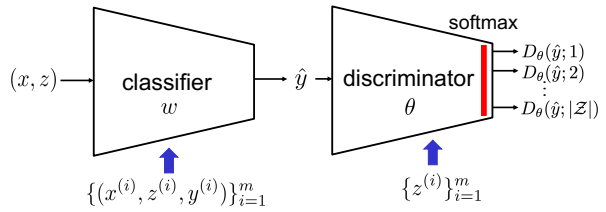

Fig. 2. A proposed fair classifier for disparate impact.

Remark 1 (Connection with GANs [16]): Specializing to the binary sensitive attribute setting, we can make an intimate connection with GANs. To see this, we rewrite the regularization term in (10) as: when $|\mathcal{Z}| = 2$ and $z^{(i)} \in \{0, 1\}$,

$$\frac{1}{m}\sum_{i=1}^{m} z^{(i)} \log D_\theta(\hat{y}^{(i)}; z^{(i)}) + (1 - z^{(i)}) \log \left(1 - D_\theta(\hat{y}^{(i)}; z^{(i)})\right).$$

Note that this is the same as the value function in GANs [16]. While it bears a strong similarity to GANs, it comes with two major distinctions. First our setting has a classifier instead of a generator; hence, it includes additional loss term that captures prediction accuracy reflected in the cross entropy loss. Second we have the number $|\mathcal{Z}|$ of outputs (possibly more than two) in the discriminator where each $D_\theta(\hat{y}; z)$ can be interpreted as the probability that $\hat{y}$ belongs to z. □

Remark 2 (Connection with Zhang et.al. [24]): An adversarial learning idea was already introduced in the context of fair classifiers [24]. So the fair classifier proposed therein is similar to ours in architecture. However it comes with a couple of distinctions. The first is that the weight update in their training procedure includes an additional "projection term" on top of the gradient of a loss function. Second, the additional term effectively makes the regularization factor λ in the objective function (10) vary as a function of iteration, thus making the tuning knob for fairness not fully controllable. Lastly their algorithm suffers from severer stability in training, compared to ours; see Fig. 4 for details. □

IV. Equalized Odds

In this section, we extend to another fairness measure EO. First we make a connection between the definition (2) of EO and $I(Z; \hat{Y}|Y)$. As before, we can easily show that the constraint of $I(Z; \hat{Y}|Y) = 0$ implies $I(Z; \tilde{Y}|Y) = 0$, thereby leading to $\mathsf{EO} = 1$. This then naturally motivates the following optimization:

$$\min_w \frac{1}{m}\sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) + \lambda \cdot I(Z; \hat{Y}|Y). \tag{11}$$

Next, as in (7), one can manipulate $I(Z; \hat{Y}|Y)$ as:

$$I(Z; \hat{Y}|Y) = \sum_{y,\hat{y},z} P_{\hat{Y},Z,Y}(\hat{y}, z, y) \log \underbrace{\frac{P_{\hat{Y},Z|y}(\hat{y}, z)}{P_{\hat{Y}|y}(\hat{y})}}_{=:D^*(\hat{y};z,y)} + H(Z|Y)$$

where $P_{\hat{Y},Z,Y}$ indicates the joint distribution of $(\hat{Y}, Z, Y)$; and $P_{\hat{Y},Z|y}$ and $P_{\hat{Y}|y}$ denote the distributions of $(\hat{Y}, Z)$ and $\hat{Y}$, respectively, conditioned on $Y = y$. Here a key trick corresponding to Theorem 1 is to represent the above in terms of another function optimization, formally stated below.

Theorem 2 (MI via function optimization for EO*):*

$$I(Z; \hat{Y}|Y) = H(Z|Y) + \max_{D(\hat{y};z,y):\sum_z D(\hat{y};z,y)=1} \sum_{\hat{y},y,z} P_{\hat{Y},Z,Y}(\hat{y}, z, y) \log D(\hat{y}; z, y). \tag{12}$$

Proof: First define the Lagrange function w.r.t. the max optimization in the above as:

$$\mathcal{L}(D(\hat{y}; z, y), \nu(\hat{y}, y)) = \sum_{\hat{y},z,y} P_{\hat{Y},Z,Y}(\hat{y}, z, y) \log D(\hat{y}; z, y) + \sum_{\hat{y},y} \nu(\hat{y}, y) \left(1 - \sum_z D(\hat{y}; z, y)\right)$$

where $\nu(\hat{y}, y)$'s are Lagrange multipliers. Since the objective function is concave in $D(\cdot; \cdot, \cdot)$, we can solve the problem via the KKT conditions:

$$\frac{d\mathcal{L}(D(\hat{y}; z, y), \nu(\hat{y}, y))}{dD(\hat{y}; z, y)} = \frac{P_{\hat{Y},Z,Y}(\hat{y}, z, y)}{D^\star(\hat{y}; z, y)} - \nu^\star(\hat{y}) = 0 \ \ \forall \hat{y}, z, y$$
$$\frac{d\mathcal{L}(D(\hat{y}; z, y), \nu(\hat{y}, y))}{d\nu(\hat{y}, y)} = 1 - \sum_z D^\star(\hat{y}; z, y) = 0 \ \ \forall \hat{y}, y.$$

Using the uniqueness of optimality, one can readily find:

$$D^\star(\hat{y}; z, y) = \frac{P_{\hat{Y},Z|y}(\hat{y}, z)}{P_{\hat{Y}|y}(\hat{y})}, \qquad \nu^\star(\hat{y}, y) = P_{\hat{Y},Y}(\hat{y}, y).$$

Since $D^\star(\hat{y}; z, y)$ is the same as $D^*(\hat{y}; z, y)$, we complete the proof. ■

Again using the empirical version of the true distribution $P_{\hat{Y},Z,Y}(\hat{y}, z, y)$, one can approximate $I(Z; \hat{Y}|Y)$ as:

$$I(Z; \hat{Y}|Y) \approx H(Z|Y) + \max_{D(\hat{y};z,y):\sum_z D(\hat{y};z,y)=1} \sum_{i=1}^{m} \frac{1}{m} \log D(\hat{y}^{(i)}; z^{(i)}, y^{(i)}).$$

Finally parameterizing $D(\cdot; \cdot, \cdot)$ with θ, we obtain:

$$\min_w \max_{\theta:\sum_z D_\theta(\hat{y};z,y)=1} \frac{1}{m} \left\{ \sum_{i=1}^{m} \ell_{\mathsf{CE}}(y^{(i)}, \hat{y}^{(i)}) + \lambda \sum_{i=1}^{m} \log D_\theta(\hat{y}^{(i)}; z^{(i)}, y^{(i)}) \right\}. \tag{13}$$

Here to ensure the equality constraint in (13), we use the number $|\mathcal{Y}|$ of softmax activations in the output. Again we employ neural networks for (w, θ); and alternating GD for training.

V. Experiments

We provide experimental results on synthetic and two benchmark real datasets (COMPAS [25] and Adult Census [26]). All of our results are on a separate test set and we use the metrics DI and EO represented in (1) and (2).

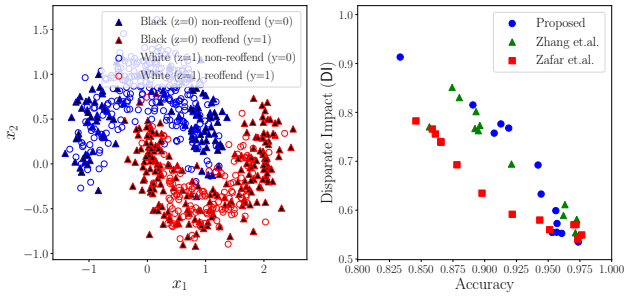

Fig. 3. (Left) Visualization of our synthetic dataset; (Right) Accuracy-fairness performance on the synthetic dataset w.r.t. disparate impact.

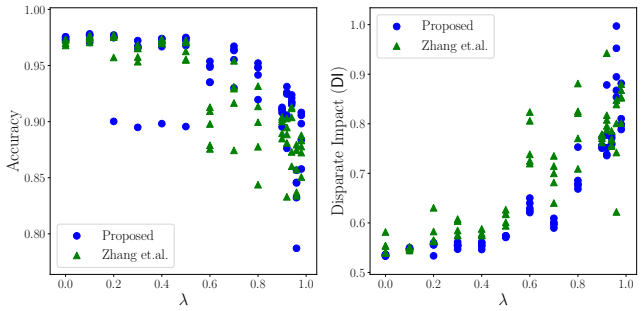

Fig. 4. Accuracy (left) and DI (right) performance as a function of a tuning knob λ that emphasizes a fairness constraint.

The conditional probabilities of the hard decision value $\tilde{Y}$ are empirically approximated.

For the synthetic data, we consider a simple yet non-trivial dataset (called the Moon dataset [29]) which is not linearly separable. See the left figure in Fig. 3. We consider a setting in which $m = 10,000$, x has two non-sensitive attributes (say x_1 and x_2), z is binary (say $z = 0$ for blacks and $z = 1$ for whites), and y is also binary (say $y = 1$ for re-offending; $y = 0$ otherwise). To generate an *unfair* dataset, we employ a simple method. We first generate m labels $y^{(i)}$'s so that they are i.i.d. each being according to $\mathsf{Bern}(0.5)$. For indices of positive examples $(y^{(i)} = 1)$, if $x_1^{(i)} \in (0.262, 1.734)$, we generate $z^{(i)}$ as per $\mathsf{Bern}(0.55)$; otherwise, $z^{(i)}$ follows $\mathsf{Bern}(0.1)$ (more biased towards $z = 0$ blacks). On the other hand, for negative examples $(y^{(i)} = 0)$, if $x_1^{(i)} \in (-0.734, 0.734)$, we generate $z^{(i)}$ as per $\mathsf{Bern}(0.9)$; otherwise, $z^{(i)}$ follows $\mathsf{Bern}(0.35)$. This way, we could generate a balanced yet unfair dataset in which $\Pr(Y = 1|Z = 0) \approx 0.658$ and $\Pr(Y = 1|Z = 1) \approx 0.352$, while respecting $Y \sim \mathsf{Bern}(0.5)$ and $Z \sim \mathsf{Bern}(0.5)$.

Fig. 3 (right) demonstrates accuracy and fairness performance for disparate impact (DI), evaluated on the synthetic test dataset ($m_{\mathsf{test}} = 5,000$) while sweeping the tuning knob λ from 0 and 1. Given the value of λ, we also normalize cross entropy loss by multiplying $(1 - \lambda)$. Here each point corresponds to a particular λ and it represents an average value over 5 trials with different seeds in training. We see that our fair classifier outperforms by respecful margins the other two baselines : (i) Zafar et.al. [4] (employing a covariance proxy for fairness measure); (ii) Zhang et.al. [24] (a GAN-based fair classifier). We also find a similar tradeoff tendency for equalized odds (EO) although it is not reported herein.

While Fig. 3 (right) shows a similar tradeoff performance between ours and Zhang et.al. [24], they exhibit different stability performances in training. See Fig. 4. Each point in the figures represents a performance evaluated only on a particular seed in training. So the spreadness of different points along the same λ captures the degree of stability in training; the more dispersed, the more unstable. We see that Zhang et.al. [24] yields more spread points, relative to ours, meaning that our algorithm is more stable. We also find that the stability issue is sensitive to a choice of optimizers for classifier and discriminator. In the reported setting, we employ Adam for classifier and SGD for discriminator.

We also evaluate on real datasets: COMPAS [25] and Adult Census [26]. For simplicity, we consider only one sensitive attribute (race for COMPAS, sex for AdultCensus), so z is binary. Table I demonstrates tradeoff performances that focus more on a fairness performance. To generate each number, we consider a range of λ in which DI (or EO) is beyond 0.9, and the reported numbers indicate the average and standard deviation over 5 trials with different seeds in training. If there is no such λ, we choose a λ that maximizes DI (or EO) at best. Notice that our algorithm outperforms the two baselines both w.r.t. DI and EO in most cases. Here Zafar et.al. [4] yields the same performance for different seeds in training; hence, the standard deviation is zero, reflected in NA in the table.

TABLE I
ACCURACY AND FAIRNESS PERFORMANCES ON REAL DATASETS (COMPAS [25] AND ADULT CENSUS [26]).

Data	Method	DI	Acc.	EO	Acc.
COMPAS	Proposed	0.959(±0.012)	0.675(±0.005)	0.908(±0.029)	0.638(±0.008)
	Zafar et.al. [4]	0.986(±NA)	0.641(±NA)	0.752(±NA)	0.641(±NA)
	Zhang et.al. [24]	0.951(±0.037)	0.654(±0.010)	0.890(±0.049)	0.626(±0.033)
AdultCensus	Proposed	0.942(±0.052)	0.837(±0.002)	0.753(±0.220)	0.844(±0.004)
	Zafar et.al. [4]	0.849(±NA)	0.822(±NA)	0.606(±NA)	0.822(±NA)
	Zhang et.al. [24]	0.883(±0.039)	0.837(±0.002)	0.791(±0.084)	0.842(±0.004)

VI. CONCLUSION

We developed an MI-based fair classifier that achieves the best tradeoff performance between prediction accuracy and a fairness performance quantified as disparate impact (DI) or equalized odds (EO). The proposed algorithm is based on our representation for MI in terms of function optimization which bears a strong similarity to the GAN formulation. One future work of interest is to address the stability issue in training that our algorithm still suffers from, as identified in Fig. 4.

VII. ACKNOWLEDGEMENTS

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01396, Development of framework for analyzing, detecting, mitigating of bias in AI model and training data)

REFERENCES

[1] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2015.

[2] S. Barocas and Selbst, "Big data's disparate impact," *California Law Review*, 2016.

[3] M. Hardt, E. Price, and N. Srebo, "Equality of opportunity in supervised learning," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2016.

[4] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," *20th International Conference on Artificial Intelligence and Statistics Conference (AISTATS)*, Apr. 2017.

[5] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," *International Conference on World Wide Web (WWW)*, Apr. 2017.

[6] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, "Fairness through awareness," *Innovations in Theoretical Computer Science Conferennce (ITCS)*, Jan. 2012.

[7] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. hsin Chi, and A. Beutel, "Counterfactual fairness in text classification through robustness," *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, Feb. 2018.

[8] M. Kearns, A. Roth, and S. Sharifi-Malvajerdi, "Average individual fairness: Algorithms, generalization and experiments," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2019.

[9] M. Yurochkin, A. Bower, and Y. Sun, "Training individually fair ml models with sensitive subspace robustness," *International Conference on Learning Representations (ICLR)*, Apr. 2020.

[10] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil, "Empirical risk minimization under fairness constraints," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2018.

[11] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," *International Conference on Machine Learning (ICML)*, July 2018.

[12] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2017.

[13] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," *International Conference on Machine Learning (ICML)*, June 2013.

[14] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," *Machine Learning and Knowledge Discovery in Databases*, Sept. 2012.

[15] D. Xu, S. Yuan, L. Zhang, and X. Wu, "FairGAN: Fairness-aware generative adversarial networks," *IEEE International Conference on Big Data*, Dec. 2018.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2014.

[17] J. Mary, C. Calauzenes, and N. El Karoui, "Fairness-aware learning for continuous attributes and treatments," *International Conference on Machine Learning (ICML)*, June 2019.

[18] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011.

[19] G. Louppe, M. Kagan, and K. Cranmer, "Learning to pivot with adversarial networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2017.

[20] J. Liao, C. Huang, P. Kairouz, and L. Sankar, "Learning generative adversarial representations (gap) under fairness and censoring constraints," *arXiv preprint arXiv:1910.00411*, 2019.

[21] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, "Learning controllable fair representations," *22nd International Conference on Artificial Intelligence and Statistics Conference (AISTATS)*, Apr. 2018.

[22] A. Géron, *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. O'Reilly, Sept. 2017.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, Dec. 2014.

[24] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, Feb. 2018.

[25] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to 272 predict future criminals. And it's biased against blacks," *https://www.propublica.org/article/machine-bias-risk-assessments-incriminal-sentencing*, May 2015.

[26] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York Wiley, 2th ed., July 2006.

[28] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, Mar. 2004.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, Nov. 2011.