FairGAN$^+$: Achieving Fair Data Generation and Classification through Generative Adversarial Nets

Depeng Xu, Shuhan Yuan, Lu Zhang, Xintao Wu
University of Arkansas, Fayetteville, AR, USA
Email: {depengxu, sy005, lz006, xintaowu}@uark.edu

***Abstract*—How to achieve fairness is important for next generation machine learning. Two tasks that are equally important in fair machine learning are how to obtain fair datasets and how to build fair classifiers. In this work, we propose a new generative adversarial network (GAN) model for fair machine learning, named FairGAN$^+$. FairGAN$^+$ contains a generator to generate close-to-real samples, a classifier to predict class labels and three discriminators to assist adversarial learning. FairGAN$^+$ simultaneously achieves fair data generation and classification by co-training the generative model and the classifier through joint adversarial games with the discriminators. Evaluations on real world data show the effectiveness of FairGAN$^+$ on both fair data generation and fair classification.**

***Index Terms*—fairness-aware learning, generative adversarial networks, fair data generation, fair classification**

I. Introduction

Discrimination indicates unfair treatment towards individuals based on the group to which they are perceived to belong. In machine learning, discrimination may be unintentional but have powerful effect on vulnerable groups. There are two major tasks in fairness-aware machine learning: (1) releasing fair datasets and (2) building models that make fair predictions. The first task is important for data owners to release data for various purposes, including scientific data analysis and training machine learning models. Previous researches propose to mitigate discriminative bias in the data by either modifying the labels [1]–[3] or the whole records [4], [5]. Some recent studies incorporate generative adversarial networks (GAN) to generate close-to-real fair data [6], [7]. Generative adversarial networks are deep neural net architectures comprised of a generator and a discriminator [8]. After playing the adversarial game with the discriminator, the generator can generate high quality synthetic data that are indistinguishable from real data. Models trained from the generated high quality data instead of real data have impressive predictive power, especially when the real data are very limited [9]. In [6], the authors proposed a model called FairGAN, which adds an additional discriminator into the original GAN to train the generator to generate fair data through adversarial learning. The generated data by FairGAN have good utility and are fair.

For the second task, when building machine learning models, if there exist historical biased decisions against the protected group in the training data, models learned from such data may also make discriminative predictions against the protected group [1], [4], [10]. In addition, the learning process can also introduce biases into predicted decisions [11]. Many works assume that the models trained from fair data can make fair predictions. However, it has been shown that there is no theoretical guarantee on this assumption [12]. Thus, some researches propose to mitigate discriminative bias in model predictions by adjusting the learning process [13] or changing the predicted labels [11]. Recent studies [14], [15] also use adversarial learning techniques to achieve fairness in classification and representation learning. In FairGAN [6], the authors suggest that classifiers trained on fair synthetic data can make fair predictions with comparison to a naive approach that randomizes the protected attribute. The naive approach generates data with the same unprotected attributes and labels as the real data, so the classifiers learned from the generated data by the naive approach have the same bias in predictions as those learned from the real data. Hence, the naive approach is unlikely to achieve fair classification. FairGAN removes disparate treatment and disparate impact when generating synthetic data, so it is more likely to achieve fair classification. However, like all the methods that modify training data, generating fair data by FairGAN cannot guarantee fair classification for models trained from them. Another disadvantage of FairGAN is that there are classification-based fairness notions such as equality of odds or equality of opportunity which cannot be achieved via generating fair data alone. Equality of opportunity is an important fairness notion in classification models [11]. It emphasizes on that individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome.

In this paper, we propose an improved version of FairGAN, called FairGAN$^+$, to address both tasks simultaneously. Our work is motivated by the Auxiliary classifier generative adversarial network (ACGAN), which is an extension to the GAN structure that allows to learn a generative model and a classifier at the same time [16]. It improves the generator on generating data with diverse representations of class labels and the classifier on performance on limited training data. Inspired by ACGAN, we incorporate a classifier into the FairGAN$^+$ structure. In addition to releasing fair data like original FairGAN, FairGAN$^+$ can also produce a classifier that makes guaranteed fair predictions. In order to train the classifier to be fair, we adopt another discriminator to train the classifier through adversarial learning. As a result, FairGAN$^+$

978-1-7281-0858-2/19/$31.00 © 2019 IEEE

contains one generator, one classifier and three discriminators.

We conduct experiments using real world census data to show that the generator of FairGAN$^+$ can achieve fair data generation with good data utility and free from disparate treatment and disparate impact. The classifier of FairGAN$^+$ can achieve guaranteed classification fairness in notions of demographic parity or equality of odds with good classification utility. The co-training of the generative model and the classifier improves the performances of each other.

II. Preliminary

A. *Fairness and Discrimination*

In fairness-aware learning, the literature has studied notions of group fairness on data and classification [5], [11].

1) Fairness in Data: Consider a labeled dataset $\mathcal{D}$, which contains a set of unprotected attributes $\mathbf{X} \in \mathbb{R}^n$, a class label $Y \in \{0,1\}$ and a protected attribute $S \in \{0,1\}$. Note that we consider S and Y as binary variables for ease of discussion.

Statistical fairness is a notion of data fairness which measures the potential discrimination caused by the correlation between the class label Y and the protected attribute S. The property of statistical fairness is defined as $P(Y=1|S=1) = P(Y=1|S=0)$.

Research in [4] proposed the concept of **ϵ-fairness** to examine the potential discrimination caused by the correlation between the unprotected attributes $\mathbf{X}$ and the protected attribute S. A labeled dataset $\mathcal{D}$ is said to be ϵ-fair if for any classification algorithm $f : \mathbf{X} \to S$, $BER(f(\mathbf{X}), S) > \epsilon$ with empirical probabilities estimated from $\mathcal{D}$, where BER (balanced error rate) is defined as

$$BER(f(\mathbf{X}), S) = \tfrac{P(f(\mathbf{X})=0|S=1)+P(f(\mathbf{X})=1|S=0)}{2}. \quad (1)$$

BER indicates the average class-conditioned error of f on distribution $\mathcal{D}$ over the pair $(\mathbf{X}, S)$.

2) Fairness in Classification: Consider the classifier $\eta : \mathbf{X} \to Y$ which predicts the class label Y given the unprotected attributes $\mathbf{X}$. Classification fairness requires that the predicted label $\eta(\mathbf{X})$ is unbiased with respect to the protected variable S. The following notions of fairness in classification was defined by [11] and refined by [17].

Demographic parity Given a labeled dataset $\mathcal{D}$ and a classifier $\eta : \mathbf{X} \to Y$, the property of demographic parity is defined as $P(\eta(\mathbf{X}) = 1|S = 1) = P(\eta(\mathbf{X}) = 1|S = 0)$. This means that the predicted labels are independent of the protected attribute.

Equality of odds Given a labeled dataset $\mathcal{D}$ and a classifier η, the property of equality of odds is defined as $P(\eta(\mathbf{X}) = 1|Y = y, S = 1) = P(\eta(\mathbf{X}) = 1|Y = y, S = 0)$, where $y \in \{0,1\}$. Hence, for $Y = 1$, equality of odds requires the classifier η has equal true positive rates (TPR) between two subgroups $S = 1$ and $S = 0$; for $Y = 0$, the classifier η has equal false positive rates (FPR) between two subgroups.

In many binary classification cases, $Y = 1$ is a more important outcome. With only requiring non-discrimination on the specific outcome group, the equality of odds can be relaxed to the equality of opportunity.

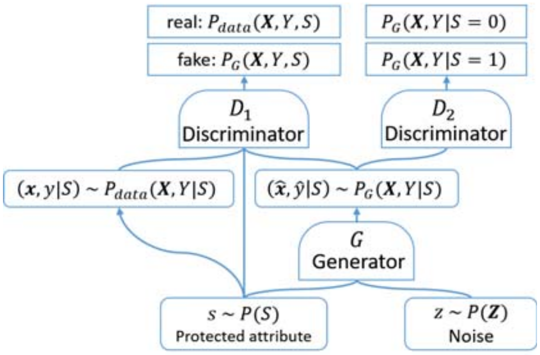
real: $P_{data}(\boldsymbol{X}, Y, S)$
fake: $P_G(\boldsymbol{X}, Y, S)$
$P_G(\boldsymbol{X}, Y|S=0)$
$P_G(\boldsymbol{X}, Y|S=1)$
D_1
Discriminator
D_2
Discriminator
$(\boldsymbol{x}, y|S) \sim P_{data}(\boldsymbol{X}, Y|S)$
$(\hat{\boldsymbol{x}}, \hat{y}|S) \sim P_G(\boldsymbol{X}, Y|S)$
G
Generator
$s \sim P(S)$
Protected attribute
$z \sim P(\boldsymbol{Z})$
Noise

Fig. 1: The structure of FairGAN

Equality of opportunity Given a labeled dataset $\mathcal{D}$ and a classifier η, the property of equality of opportunity in a classifier is defined as $P(\eta(\mathbf{X}) = 1|Y = 1, S = 1) = P(\eta(\mathbf{X}) = 1|Y = 1, S = 0)$. The equality of opportunity only focuses on the true positive rates.

B. *Generative Adversarial Networks*

Generative adversarial networks (GAN) consist of two components: a generator G and a discriminator D, both of which are multilayer neural networks. Given random noise variable $\mathbf{z} \sim P(\mathbf{Z})$ as input, the generator $G(\mathbf{z})$ attempts to learn a generative distribution P_G to match the real data distribution P_{data}. Meanwhile, the discriminator D is a binary classifier to predict whether a data sample is from real data distribution or the fake data generated by the generator. By playing the adversarial game, GAN is formalized as a minimax game $\min_G \max_D V(G, D)$ with the value function:

$$V(G, D) = \mathbb{E}_{\mathbf{x} \sim P_{data}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim P_{\mathbf{Z}}}[\log(1 - D(G(\mathbf{z})))].$$

Research in [16] proposed a variant of the GAN structure, called auxiliary classifier generative adversarial networks (AC-GAN). Each generated sample has a corresponding class label y in addition to the noise $\mathbf{z}$. A classifier is incorporated into the model to reconstruct the class label from the data samples. The generator $G(y, \mathbf{z})$ can produce class conditional samples that match both the real data distribution and the class conditions.

Research in [6] shows that GAN can be modified to generate synthetic data that are both close to real data and fair. Their proposed **FairGAN** contains one generator G and two discriminators D_1 and D_2, as shown in Figure 1. Each generated sample $G(\mathbf{z})$ has a corresponding protected attribute s. The generator learns the conditional distribution $P_G(\mathbf{X}, Y|S)$. The discriminator D_1 is the same as the regular GAN model. The discriminator D_2 is trained to distinguish the protected attribute S of generated samples. By playing adversarial games with both discriminators, the generator can generate high quality fake samples and does not encode any information supporting to predict the value of the protected attribute.

III. FairGAN$^+$

FairGAN$^+$ aims to simultaneously achieve fair data generation and fair classification through the GAN architecture.

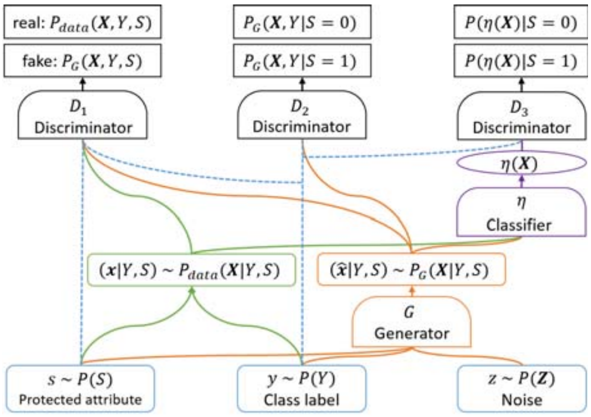
real: $P_{data}(\mathbf{X}, Y, S)$
fake: $P_G(\mathbf{X}, Y, S)$
$P_G(\mathbf{X}, Y|S = 0)$
$P_G(\mathbf{X}, Y|S = 1)$
$P(\eta(\mathbf{X})|S = 0)$
$P(\eta(\mathbf{X})|S = 1)$
D_1 Discriminator
D_2 Discriminator
D_3 Discriminator
$\eta(\mathbf{X})$
η Classifier
$(\mathbf{x}|Y, S) \sim P_{data}(\mathbf{X}|Y, S)$
$(\hat{\mathbf{x}}|Y, S) \sim P_G(\mathbf{X}|Y, S)$
G Generator
$s \sim P(S)$ Protected attribute
$y \sim P(Y)$ Class label
$z \sim P(\mathbf{Z})$ Noise

Fig. 2: The structure of FairGAN$^+$

For fair data generation, we consider statistical fairness and/or ϵ-fairness. For fair classification, we consider demographic parity, equality of odds and/or equality of opportunity.

A. Model Framework

FairGAN$^+$ consists of one generator G, one classifier η and three discriminators D_1, D_2 and D_3. Figure 2 shows the structure of FairGAN$^+$. In FairGAN$^+$, each generated sample $G(\mathbf{z})$ has a corresponding value pair of the protected attribute $s \sim P_{data}(S)$ and the class label $y \sim P_{data}(Y)$. The generator $G(\mathbf{z})$ generates fake $\hat{\mathbf{x}}$ following the conditional distribution $P_G(\mathbf{X}|Y, S)$ given a noise variable $\mathbf{z}$. We adopt the revised generator from medGAN [9] so the generator $G(\mathbf{z})$ can generate both discrete and continuous data. The classifier $\eta : \mathbf{X} \rightarrow Y$ outputs the prediction $\eta(\mathbf{X})$ from $\mathbf{X}$. The prediction $\eta(\mathbf{X})$ is trained to both accurately predict the label Y and be free from discrimination. The discriminator D_1 is trained to distinguish between the real data from $P_{data}(\mathbf{X}, Y, S)$ and the generated fake data from $P_G(\mathbf{X}, Y, S)$. The generator plays adversarial game with D_1 to generate close-to-real fake data. The discriminator D_2 is trained to distinguish values of the protected attribute of each sample, $P_G(\mathbf{X}, Y|S = 1)$ and $P_G(\mathbf{X}, Y|S = 0)$. D_2 works like a fairness constraint to data generation. The generator plays another adversarial game with D_2 so the generated data satisfy fairness notions in data. The discriminator D_3 is trained to distinguish values of the protected attribute from the predictions made by η, $P(\eta(\mathbf{X}) = 1|S = 1)$ and $P(\eta(\mathbf{X}) = 1|S = 0)$. D_3 works like a fairness constraint to classification. The classifier plays adversarial game with D_3 so its predictions satisfy fairness notions in classification.

The objective function of FairGAN$^+$ is $J = V + L$, where V is the value function of the overall minimax games between generator, classifier and discriminators; L is the objective function of the classifier.

The value function V of the overall minimax game is described as:

$$\begin{aligned}\min_{G,\eta} \max_{D_1,D_2,D_3} & V(G, \eta, D_1, D_2, D_3) \\ &= V_1(G, D_1) + \lambda V_2(G, D_2) + \mu V_3(\eta, D_3),\end{aligned}$$

where

$$\begin{aligned}&V_1(G, D_1) \\ &= \mathbb{E}_{s\sim P(S),y\sim P(Y),\mathbf{x}\sim P_{data}(\mathbf{X}|Y,S)}[\log D_1(\mathbf{x}, y, s)] \\ &+ \mathbb{E}_{s\sim P(S),y\sim P(Y),\hat{\mathbf{x}}\sim P_G(\mathbf{X}|Y,S)}[\log(1 - D_1(\hat{\mathbf{x}}, y, s))],\end{aligned} \tag{2}$$

$$\begin{aligned}V_2(G,D_2) = &\mathbb{E}_{y\sim P(Y),\hat{\mathbf{x}}\sim P_G(\mathbf{X}|Y,S=1)}[\log D_2(\hat{\mathbf{x}}, y)] \\ &+ \mathbb{E}_{y\sim P(Y),\hat{\mathbf{x}}\sim P_G(\mathbf{X}|Y,S=0)}[\log(1 - D_2(\hat{\mathbf{x}}, y))],\end{aligned} \tag{3}$$

$$\begin{aligned}V_3(\eta, D_3) = &\mathbb{E}_{\mathbf{x}\sim P(\mathbf{X}|Y,S=1)}[\log D_3(\eta(\mathbf{x}))] \\ &+ \mathbb{E}_{\mathbf{x}\sim P(\mathbf{X}|Y,S=0)}[\log(1 - D_3(\eta(\mathbf{x}))],\end{aligned} \tag{4}$$

λ is a hyperparameter that specifies a trade-off between data utility and fairness of data generation, μ is a hyperparameter that specifies a trade-off between classification accuracy and classification fairness of η.

The first value function V_1 aims to make the generated data match the real data distribution, where the generator G seeks to learn the joint distribution $P_G(\mathbf{X}, Y, S)$ over real data $P_{data}(\mathbf{X}, Y, S)$. The second value function V_2 aims to make the generated samples not encode any information supporting to predict the value of the protected attribute S. Therefore, D_2 is trained to correctly predict S given a generated sample while the generator G aims to fool the discriminator D_2. Once the generated sample $\{\hat{\mathbf{x}}, y\}$ cannot be used to predict the protected attribute S, the correlation between $\{\hat{\mathbf{x}}, y\}$ and S is removed, i.e., $\{\hat{\mathbf{x}}, y\} \perp S$. Hence, FairGAN$^+$ can achieve fairness in data generation. The third value function V_3 aims to make the prediction $\eta(\mathbf{X})$ of samples not encode any information supporting to predict the value of the protected attribute S. Therefore, D_3 is trained to correctly predict S given a sample while the classifier η aims to fool the discriminator D_3. Once the prediction of η cannot be used to predict the protected attribute S, the correlation between $\eta(\mathbf{X})$ and S is removed, i.e., $\eta(\mathbf{X}) \perp S$. Then FairGAN$^+$ can release η, which achieves the desired fairness notion.

The objective function L of the classifier is described as:

$$\begin{aligned}\max_{G,\eta} L(G, \eta) = &\mathbb{E}_{y\sim P(Y),\mathbf{x}\sim P_{data}(\mathbf{X}|Y,S)}[y \log \eta(\mathbf{x})] \\ &+ \mathbb{E}_{y\sim P(Y),\hat{\mathbf{x}}\sim P_G(\mathbf{X}|Y,S)}[y \log \eta(\hat{\mathbf{x}})].\end{aligned}$$

The classifier η maximizes the log-likelihood of the correct class labels as it makes more accurate predictions during training. The generator G also maximizes log-likelihood of the correct class labels as it generates samples that match each class accordingly. We take advantage of the feedback loop of D_1, G, η. Improving the utility of G can improve the utility of η on making correct predictions. Improving η can improve G on generating more realistic samples conditioned on each class. We also take advantage of the feedback loop of D_2, G, D_3, η. Improving the fairness of G can improve η on making fair predictions. Improving the fairness of η can improve G on generating fair data. Hence, by simultaneously learning a generative model and a classifier, FairGAN$^+$ can perform better than a standalone generative model or a standalone classifier.

B. Application to Different Classification-based Fairness

Models that modify data or generate fair data apply fairness constraints to enforce independence between the class label Y and the protected attribute S. However, such preprocessing methods cannot guarantee the independence between the predicted class label $\eta(\mathbf{X})$ and the protected attribute S. Constraints for some classification-based fairness notions require to constrain not only the association between the predicted class label $\eta(\mathbf{X})$ and the protected attribute S, but also the conditioned association when conditioning on the real class label Y. In our model, we connect the real class label Y to the constraints on the predicted class label $\eta(\mathbf{X})$ through the discriminator D_3. The inputs of D_3 are both the predicted class label $\eta(\mathbf{X})$ from η and the real class label Y from the prior distribution. This is not necessary for demographic parity but essential for equality of odds and equality of opportunity.

Demographic parity is defined as $P(\eta(\mathbf{X}) = 1|S = 1) = P(\eta(\mathbf{X}) = 1|S = 0)$. It simply requires that the predictions of η is independent of the protected attribute S. In the above general framework, we mostly discuss the model in terms of demographic parity. The value function V_3 is exactly as Equation 4. Once the prediction of η cannot be used to predict the protected attribute S, the correlation between $\eta(\mathbf{X})$ and S is removed, i.e., $\eta(\mathbf{X}) \perp S$. Then FairGAN$^+$ can release η, which achieves demographic parity.

Equality of odds is defined as $P(\eta(\mathbf{X}) = 1|Y = y, S = 1) = P(\eta(\mathbf{X}) = 1|Y = y, S = 0)$, where $y \in \{0, 1\}$. As a classification-based fairness notion, equality of odds considers that individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome. It evaluates correlation between the prediction of classifier $\eta(\mathbf{X})$ and the protected attribute S based on the value of real label Y. To achieve equality of odds for a classifier in the FairGAN$^+$ framework, we need to provide the corresponding real label y of the prediction $\eta(\mathbf{X})$ to the discriminator D_3. As a constraint of equality of odds to classification. D_3 aims to keep $P(\eta(\mathbf{X}) = 1|Y = y, S = 1) = P(\eta(\mathbf{X}) = 1|Y = y, S = 0)$. D_3 is trained to distinguish the two categories of the predictions made by η given the real label $Y = y$, $P(\eta(\mathbf{X}) = 1|Y = y, S = 1)$ and $P(\eta(\mathbf{X}) = 1|Y = y, S = 0)$. In this case, the third value function V_3 of the minimax game becomes

$$\begin{aligned}V_3(\eta, D_3) &= \mathbb{E}_{y\sim P(Y),\mathbf{x}\sim P(\mathbf{X}|Y=y,S=1)}[\log D_3(\eta(\mathbf{x})|Y = y)]\\ &+ \mathbb{E}_{y\sim P(Y),\mathbf{x}\sim P(\mathbf{X}|Y=y,S=0)}[\log(1 - D_3(\eta(\mathbf{x})|Y = y))].\end{aligned}$$

V_3 aims to make the prediction $\eta(\mathbf{X})$ of samples not encode any information supporting to predict the value of the protected attribute S given the real label $Y = y$. Therefore, D_3 is trained to correctly predict S given a sample and the corresponding real label Y while the classifier η aims to fool the discriminator D_3. Once the prediction of η cannot be used to predict the protected attribute S given the real label $Y = y$, the conditional correlation between $\eta(\mathbf{X})$ and S is removed, i.e., $\eta(\mathbf{X}) \perp S|Y = y$. Then FairGAN$^+$ can release η, which achieves equality of odds.

Equality of opportunity is a special case of equality of odds. For FairGAN$^+$ based on equality of opportunity, the idea is straightforward. It only needs to consider the case $Y = 1$, so the discriminator D_3 only aims to constrain a subgroup instead of the whole population. In this case, the third value function V_3 of the minimax game becomes

$$\begin{aligned}V_3(\eta, D_3) &= \mathbb{E}_{y\sim P(Y),\mathbf{x}\sim P(\mathbf{X}|Y=1,S=1)}[\log D_3(\eta(\mathbf{x})|Y = 1)]\\ &+ \mathbb{E}_{y\sim P(Y),\mathbf{x}\sim P(\mathbf{X}|Y=1,S=0)}[\log(1 - D_3(\eta(\mathbf{x})|Y = 1))].\end{aligned}$$

Once the prediction of η cannot be used to predict the protected attribute S given the real label $Y = 1$, the conditional correlation between $\eta(\mathbf{X})$ and S is removed, i.e., $\eta(\mathbf{X}) \perp S|Y = 1$. Then FairGAN$^+$ can release η, which achieves equality of opportunity.

IV. Evaluation of FairGAN$^+$

We evaluate the performance of FairGAN$^+$ based on demographic parity and FairGAN$^+$ based on equality of odds on fair data generation and fair classification

A. Experimental Setup

1) Baselines: To evaluate the performance of FairGAN$^+$, we compare with the ACGAN model [16] on effectiveness, with the FairGAN model [6] on fair data generation, and with adversarial debiasing [14] on fair classification.

ACGAN aims to generate the synthetic samples that have the same distribution as the real data given the values of labels, i.e., $P_G(\mathbf{X}|Y, S) = P_{data}(\mathbf{X}|Y, S)$. ACGAN also contains a classifier, but does not enforce fairness in either data generation or classification. Note that, in order to match our scenario, we independently generate both the label Y and the protected attribute S in the ACGAN. Thus, ACGAN achieves statistical fairness in data generation similar to random shuffle the protected attribute. The classifiers built in both FairGAN$^+$ and ACGAN are logistic regression model.

FairGAN generates fair data and trains a logistic regression model (not built-in) on the generated data with the assumption that fair data generation automatically achieves fair classification for any classification-based fairness.

Adversarial debiasing (AD) also applies adversarial learning to achieve fairness in classification. AD cannot generate fair data as it is not a generative model.

2) Datasets: We evaluate FairGAN$^+$ and baselines on the UCI Adult dataset. It contains 48,842 samples. The class label Y is "Income", and the protected attribute S is "Gender". There are 12 unprotected attributes $\mathbf{X}$. We convert each attribute to a one-hot vector and combine all of them to a feature vector with 57 dimensions.

Besides adopting the original Adult dataset (**D1-Real**), we also generate four types of synthetic data, **D2-ACGAN** that is generated by ACGAN, **D3-FairGAN** that is generated by FairGAN with $\lambda = 1$, **D4-FairGAN$^+$(DP)** that is generated by FairGAN$^+$ based on demographic parity with $\lambda = \mu = 1$, and **D5-FairGAN$^+$(EO)** that is generated by FairGAN$^+$ based on equality of odds with $\lambda = \mu = 1$. We set the sample sizes of the synthetic datasets the same as the real dataset.

TABLE I: Data fairness and utility of real and synthetic datasets

	Metric	D1-Real	D2-ACGAN	D3-FairGAN	D4-FairGAN$^+$(DP)	D5-FairGAN$^+$(EO)
Fairness	$RD(\mathcal{D})$	0.1989	0.0120±0.0088	0.0411±0.0295	**0.0106±0.0081**	0.0116±0.0087
	BER	0.1538	0.1964±0.0033	0.3862±0.0036	**0.3867±0.0049**	0.3207±0.0121
Utility	$dist(\mathbf{X}, Y)$	NA	0.0245±0.0003	0.0233±0.0004	**0.0232±0.0003**	0.0239±0.0003
	$dist(\mathbf{X}, Y, S)$		0.0204±0.0004	**0.0208±0.0005**	0.0212±0.0005	0.0210±0.0004

3) Classifiers: After training, we get five logistic regression classifiers: **C1-Real** that is a regular logistic regression model trained on the real Adult data, **C2-ACGAN** that is trained by ACGAN, **C3-FairGAN** that is trained on data generated by FairGAN, **C4-AD** that is a debiased classifier trained on the real Adult data through adversarial learning, **C5-FairGAN$^+$(DP)** that is trained by FairGAN$^+$ based on demographic parity, and **C6-FairGAN$^+$(EO)** that is trained by FairGAN$^+$ based on equality of odds. The evaluation of the classifiers is on the real dataset.

For each model, we train five times and report the means and stand deviations of evaluation results.

B. Fair Data Generation

We evaluate whether FairGAN$^+$ can generate fair data that satisfy statistical fairness and ϵ-fairness while learning the distribution of real data precisely.

1) Data Fairness: We adopt the risk difference $RD(\mathcal{D}) = P(Y = 1|S = 1) - P(Y = 1|S = 0)$ to compare the performance of different GAN models on statistical fairness. Table I shows the risk differences in the real and synthetic datasets. The risk difference in the Adult dataset (D1-Real) is 0.1989, which indicates discrimination against female. D2-ACGAN has low risk difference due to independent priors. D3-FairGAN as expected also has low risk difference but slightly higher. D4-FairGAN$^+$(DP) and D5-FairGAN$^+$(EO) both have low risk difference as result of the adversarial game between the generator and the discriminator D_2.

We further evaluate the ϵ-fairness (disparate impact) by calculating the balanced error rates (BERs) shown in Equation 1. Note that we adopt a linear SVM to predict S and then calculate BER. Table I shows the BERs in the real and synthetic datasets. The BER in D1-Real is 0.1538, which means a linear SVM can predict S given $\mathbf{x}$ with high accuracy. Hence, there is disparate impact in the real dataset. The BER in D2-ACGAN is 0.1964±0.0033, which shows that ACGAN captures the disparate impact in the real dataset due to its unawareness of ϵ-fairness. On the contrary, the BERs in D3-FairGAN, D4-FairGAN$^+$(DP) and D5-FairGAN$^+$(EO) are 0.3862±0.0036, 0.3867±0.0049 and 0.3207±0.0121, respectively, which indicates using the generated $\hat{\mathbf{x}}$ to predict the real S has much higher error rate. So the disparate impacts in FairGAN and FairGAN$^+$ are small.

2) Data Utility: We evaluate the closeness between each synthetic dataset and the real dataset by calculating the Euclidean distance of joint probabilities w/o the protected attribute S, i.e. $P(\mathbf{X}, Y)$ and $P(\mathbf{X}, Y, S)$. The Euclidean distance is calculated between the estimated probability mass functions on the sample space, where $dist(\mathbf{X}, Y) = ||P_{data}(\mathbf{X}, Y) - P_G(\mathbf{X}, Y)||_2$ and $dist(\mathbf{X}, Y, S) = ||P_{data}(\mathbf{X}, Y, S) - P_G(\mathbf{X}, Y, S)||_2$. A smaller distance indicates better closeness. In Table I, all synthetic datasets have small distances to the real dataset for joint and conditional probabilities. The distances of FairGAN$^+$ is even smaller than FairGAN. We can observe that FairGAN$^+$ still achieves good data utility after satisfying fairness constraints. As the co-training of the generative model and the classifier improves G on data generation, FairGAN$^+$ has a more efficient trade-off between utility and fairness of data generation.

C. Fair Classification

We evaluate the classifiers that are trained in ACGAN and FairGAN$^+$ on the real dataset.

1) Classification Fairness: We first compare different classifiers including C5-FairGAN$^+$(DP) on demographic parity using risk difference $RD(\eta) = P(\eta(\mathbf{X}) = 1|S = 1) - P(\eta(\mathbf{X}) = 1|S = 0)$. Table II shows the risk difference of different classifiers. C1-Real has a risk difference of 0.1834, which indicates a regular logistic regression model learns the risk difference in the real Adult data and makes unfair predictions. C2-ACGAN has a high risk difference 0.1119±0.0182, which is discriminative due to ACGAN's unawareness of demographic parity. C3-FairGAN has a lower risk difference 0.0901±0.0220 but still discriminative (above the threshold 0.05). This indicates that FairGAN's assumption on fair classification from fair data generation does not always hold. The risk difference is lower as the result of fair data but it is still not small enough to guarantee fairness. However, C4-AD also has a lower risk difference 0.0760±0.0058 but still discriminative (above the threshold 0.05). C5-FairGAN$^+$(DP) has a low risk difference 0.0141±0.0065. As C5-FairGAN$^+$(DP) learns to make predictions uncorrelated to S, FairGAN$^+$ based on demographic parity can achieve demographic parity.

Then we compare different classifiers including C6-FairGAN$^+$(EO) on equality of odds. Equality of odds is a more complex fairness notion based on classification. We use difference in true positive rates $DTPR = P(\eta(\mathbf{X}) = 1|Y = 1, S = 1) - P(\eta(\mathbf{X}) = 1|Y = 1, S = 0)$ and difference in false positive rates $DFPR = P(\eta(\mathbf{X}) = 1|Y = 0, S = 1) - P(\eta(\mathbf{X}) = 1|Y = 0, S = 0)$ to measure equality of odds. Table II shows the $DTPR$ and $DFPR$ of different classifiers. C1-Real has 0.1017 difference in TPR and 0.0746 difference in FPR, which indicates a regular logistic regression model without any fairness constraint makes predictions with inequality of odds. C2-ACGAN has 0.0854±0.0316 difference in TPR and 0.0395±0.0098 difference in FPR, which is smaller than C1-Real but still slightly discriminative. C3-FairGAN has 0.1473±0.0608 difference in TPR and 0.0361±0.0145

TABLE II: Classification fairness and accuracy of different classifiers

		C1-Real	C2-ACGAN	C3-FairGAN	C4-AD	C5-FairGAN$^+$(DP)	C6-FairGAN$^+$(EO)
Demographic Parity	$RD(\eta)$	0.1834	0.1119±0.0182	0.0901±0.0220	0.0760±0.0058	**0.0141±0.0065**	NA
Equality of Odds	$DTPR$	0.1017	0.0854±0.0316	0.1473±0.0608	0.0388±0.0234	NA	**0.0312±0.0316**
	$DFPR$	0.0746	0.0395±0.0098	0.0361±0.0145	**0.0184±0.0121**		0.0245±0.0124
Classification Accuracy		0.8448	0.8359±0.0017	**0.8256±0.0021**	0.7902±0.0043	0.8178±0.0035	0.8218±0.0062

difference in FPR, which is largely discriminative for subgroup with $Y = 1$. This also indicates that FairGAN's assumption on fair classification from fair data generation is wrong. Especially for advanced classification-based fairness notions, simply training on fair data has very limited effect on such fairness notions. Hence, it is more practical to achieve such classification-based fairness by directly applying constraints on the classifier. C4-AD has has 0.0388±0.0234 difference in TPR and 0.0184±0.0121 difference in FPR, which indicates AD is effective on achieving equality of odds. C6-FairGAN$^+$(EO) has 0.0312±0.0316 difference in TPR and 0.0245±0.0124 difference in FPR, which is similar to C4-AD. As C6-FairGAN$^+$(EO) learns to make predictions uncorrelated to S based on the real label Y, FairGAN$^+$ based on equality of odds can achieve equality of odds in the classifier.

2) Classification Accuracy: We evaluate the accuracy of different classifiers on the real Adult dataset. Table II shows the classification accuracy of different classifiers. The accuracy of C1-Real is 0.8448. The accuracy of C2-ACGAN is 0.8359±0.0017. The accuracy of C3-FairGAN is 0.8256±0.0021. However, these classifiers are unfair to the protected group. The accuracy of C4-AD is 0.7902±0.0043. The accuracies of C5-FairGAN$^+$(DP) and C6-FairGAN$^+$(EO) are 0.8178±0.0035 and 0.8218±0.0062, respectively. They are slightly lower than other classifiers, which indicates that AD and FairGAN$^+$ models have a trade-off between classification accuracy and fairness. In FairGAN$^+$, small utility loss is caused by playing the adversarial game with the third discriminator D_3 to achieve respective notions of classification fairness. We can observe that C5-FairGAN$^+$(DP) and C6-FairGAN$^+$(EO) have higher accuracy than C4-AD after satisfying the desired classification-based fairness. As the co-training of the generative model and the classifier improves η on classification, FairGAN$^+$ has a more efficient trade-off between the utility and fairness of classification.

V. Conclusion

In this paper, we have developed FairGAN$^+$ to (1) generate fair data, which is free from disparate treatment and disparate impact w.r.t. the real protected attribute, while retaining high data utility, and (2) release a fair classifier that satisfies classification-based fairness, such as demographic parity, equality of odds and equality of opportunity. FairGAN$^+$ consists of one generator, one classifier and three discriminators. In particular, the generator generates fake samples conditioned on the protected attribute. The classifier learns to predict labels based on the unprotected attribute. The first discriminator is trained to identify whether samples are real or fake. The second discriminator is trained to distinguish whether the generated samples are from the protected group or unprotected group. The generator can generate fair data with high utility by playing the adversarial games with these two discriminators. The third discriminator is trained to distinguish whether the predicted label is based on the protected attribute. After training, the classifier can make accurate predictions and satisfy classification-based fairness by playing the adversarial game with the third discriminator. The experimental results showed the effectiveness of FairGAN$^+$ on both fair data generation and fair classification, and the better trade-off between the utility and fairness when co-training a generative model and a classifier.

Acknowledgment

This work was supported in part by NSF 1646654, 1841119, 1920920, and 1940093.

References

[1] L. Zhang, Y. Wu, and X. Wu, "A causal framework for discovering and removing direct and indirect discrimination," in *IJCAI*, 2017.

[2] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*, 2009.

[3] I. Zliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination," in *ICDM*, 2011.

[4] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *KDD*, 2015.

[5] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, 2012.

[6] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *BigData*, 2018.

[7] D. Xu, Y. Wu, S. Yuan, L. Zhang, and X. Wu, "Achieving causal fairness through generative adversarial networks," in *IJCAI*, 2019.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *NeurIPS*, 2014.

[9] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *MLHC*, 2017.

[10] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in data release," in *KDD*, 2017.

[11] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NeurIPS*, 2016.

[12] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in prediction," in *IJCAI*, 2018.

[13] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *AISTATS*, 2017.

[14] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *AIES*, 2018.

[15] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," 2018.

[16] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *ICML*, 2017.

[17] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data decisions and theoretical implications when adversarially learning fair representations," in *FAT/ML*, 2017.