Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)

Achieving Causal Fairness through Generative Adversarial Networks

Depeng Xu, **Yongkai Wu**, **Shuhan Yuan**, **Lu Zhang** and **Xintao Wu**
University of Arkansas
{depengxu,yw009,sy005,lz006,xintaowu}@uark.edu

Abstract

Achieving fairness in learning models is currently an imperative task in machine learning. Meanwhile, recent research showed that fairness should be studied from the causal perspective, and proposed a number of fairness criteria based on Pearl's causal modeling framework. In this paper, we investigate the problem of building causal fairness-aware generative adversarial networks (CFGAN), which can learn a close distribution from a given dataset, while also ensuring various causal fairness criteria based on a given causal graph. CFGAN adopts two generators, whose structures are purposefully designed to reflect the structures of causal graph and interventional graph. Therefore, the two generators can respectively simulate the underlying causal model that generates the real data, as well as the causal model after the intervention. On the other hand, two discriminators are used for producing a close-to-real distribution, as well as for achieving various fairness criteria based on causal quantities simulated by generators. Experiments on a real-world dataset show that CFGAN can generate high quality fair data.

1 Introduction

Fairness-aware learning is receiving an increasing attention in machine learning fields. How to obtain the training data that satisfy fairness is an important research problem, as machine learning models learned from biased training data may also have biased performance against sensitive attributes, such as gender, race, age [Pedreshi *et al.*, 2008; Zliobaite *et al.*, 2011; Hardt *et al.*, 2016; Zhang *et al.*, 2017; Zhang *et al.*, 2018b]. In the literature, many methods have been proposed to modify the training data for mitigating biases and achieving fairness. These methods include: Massaging [Kamiran and Calders, 2009], Reweighting [Calders *et al.*, 2009], Sampling [Kamiran and Calders, 2012], Disparate Impact Removal [Feldman *et al.*, 2015], Causal-based Removal [Zhang *et al.*, 2017; Zhang *et al.*, 2018c] and Fair Representation Learning [Edwards and Storkey, 2016; Xie *et al.*, 2017; Madras *et al.*, 2018; Zhang *et al.*, 2018a].

As the general requirement of modifying datasets is to preserve the data utility as much as possible, a recent study leverages Generative Adversarial Networks (GAN) for generating high quality fair data [Xu *et al.*, 2018]. GAN is a generative model that has demonstrated impressive performance on generating synthetic data that are indistinguishable from real data [Goodfellow *et al.*, 2014]. The idea of GAN is to let a generator and a discriminator play the adversarial game with each other. In [Xu *et al.*, 2018], the authors modify the architecture of GAN to consist of one generator and two discriminators, where one discriminator aims to ensure close-to-real generation and the other discriminator aims to ensure fairness. Their method, entitled FairGAN, can meet both requirements of high data utility and fairness.

The critical limitation of FairGAN is that it can only achieve fairness in terms of a simple statistical-based fairness criterion called demographic parity. However, as paid increasing attentions recently by researchers, fairness is a causal notion that concerns the causal connection between the sensitive attributes and the challenged decisions or outputs [Zhang *et al.*, 2017; Zhang and Bareinboim, 2018; Nabi and Shpitser, 2018; Kusner *et al.*, 2017; Chiappa, 2019; Wu *et al.*, 2019; Salimi *et al.*, 2019]. Based on Pearl's causal modeling framework [Pearl, 2009], a number of causal-based fairness notions and criteria have been proposed, including total effect [Zhang and Bareinboim, 2018], direct discrimination [Zhang *et al.*, 2017], indirect discrimination [Zhang *et al.*, 2017], and counterfactual fairness [Kusner *et al.*, 2017]. Each notion captures fairness in one particular situation from the causal perspective. Total effect treats all causal effects from the sensitive attribute to the decision as unfair. Direct and indirect discrimination, on the other hand, consider the situation where discrimination is transmitted through certain paths in the causal graph. Counterfactual fairness, again considers a different situation where we focus on the fairness with respect to a particular individual or a subgroup of individuals instead of the whole population. The causal blindness of FairGAN makes it unable to handle some causal-based notions such as counterfactual fairness, and may also affect its utility as it may remove both causal and spurious effects.

In this paper, we propose a causal fairness-aware generative adversarial network (CFGAN) for generating data that achieve various causal-based fairness criteria. Motivated by CausalGAN [Kocaoglu *et al.*, 2018], we preserve the causal

structure in the generator by arranging the neural network structure of the generator following a given causal graph. As a result, the generator can be considered as to simulate the underlying causal model of generating the observational data.[1] Then, in order to handle different fairness criteria, we adopt two generators for explicitly modeling the real world and the world after we perform some hypothetical interventions. The two generators differ in some aspects to reflect the effect of interventions, but are also synchronized in terms of sharing parameters to reflect the connections between the two worlds. Then we adopt two discriminators for achieving both the high data utility and causal-based fairness. Experiments using the real world dataset show that CFGAN can generate high quality fair data based on different criteria.

2 Preliminary

2.1 Causal Model and Intervention

Definition 1. *A causal model [Pearl, 2009] is a triple* $\mathcal{M} = \{\mathbf{U}, \mathbf{V}, \mathbf{F}\}$ *where*

1) $\mathbf{U}$ *is a set of hidden random variables that are determined by factors outside the model. A joint probability distribution* $P(\mathbf{U})$ *is defined over the variables in* $\mathbf{U}$.

2) $\mathbf{V}$ *is a set of observed random variables that are determined by variables in* $\mathbf{U} \cup \mathbf{V}$.

3) $\mathbf{F}$ *is a set of deterministic functions; for each* $V_i \in \mathbf{V}$*, a corresponding function* f_{V_i} *is a mapping from* $\mathbf{U} \cup (\mathbf{V} \setminus \{V_i\})$ *to* V_i*, i.e.,* $V_i = f_{V_i}(Pa_{V_i}, \mathbf{U}_{V_i})$*, where* $Pa_{V_i} \subseteq \mathbf{V} \setminus \{V_i\}$ *is called the parents of* V_i*, and* $\mathbf{U}_{V_i} \subseteq \mathbf{U}$.

A causal model is often illustrated by a causal graph $\mathcal{G}$ [Pearl, 2009], where each observed variable is represented by a node, and the causal relationships are represented by directed edges $\rightarrow$. In this graphical representation, the definition of parents is consistent with that in the causal model. In addition, each node V_i is associated with a conditional distribution given all its parents, i.e., $P(V_i|Pa_{V_i})$.

Inferring causal effects in the causal model is facilitated by do-operator [Pearl, 2009], which simulates the physical intervention that forces some variable $X \in \mathbf{V}$ to take certain value x. For a causal model $\mathcal{M}$, intervention $do(X = x)$ is performed by replacing original function $X = f_X(Pa_X, \mathbf{U}_X)$ with $X = x$. After replacing, the distributions of all variables that are the descendants of X may be changed. We call the causal model after the intervention the interventional model, denoted by $\mathcal{M}_x$. Correspondingly, $\mathcal{M}_x$ can be illustrated by the interventional graph $\mathcal{G}_x$ where all incoming edges to X are deleted and node X is replaced with constant x. The interventional distribution for any $\mathbf{Y} \subseteq \mathbf{V} \setminus \{X\}$ is denoted by $P(\mathbf{Y}|do(X = x))$ or $P(\mathbf{Y}_x)$. Symbolically, $P(\mathbf{Y}_x)$ can be expressed as a truncated factorization formula [Pearl, 2009] and computed from the observed distribution.

[1] It is worth noting that causal effects may not be estimated from observational data in certain situations, referred to as unidentifiable situations. The generator can be treated as simulating the true causal model only in identifiable situations.

2.2 Causal Effects

With the help of do-operator, we can infer the causal effect of X on Y by comparing the difference in interventional distributions under different interventions. Based on how the intervention is transferred in the causal model (graph), there are mainly three types of causal effects: total effect, path-specific effect, counterfactual effect [Pearl, 2009].

The total effect measures the causal effect of X on Y where the intervention is transferred along all causal paths (i.e., directed paths) from X to Y.

Definition 2. *The total effect of the value change of* X *from* x_1 *to* x_2 *on* Y *is given by* $TE(x_2, x_1) = P(Y_{x_2}) - P(Y_{x_1})$.

The path-specific effect measures the causal effect of X on Y where the intervention is transferred only along a subset of causal paths from X to Y, which is also referred to as the π-specific effect denoting the subset of causal paths as π.

Definition 3. *Given a path set* π*, the* π*-specific effect of the value change of* X *from* x_1 *to* x_2 *on* Y *(with reference* x_1*) is given by* $SE_\pi(x_2, x_1) = P(Y_{x_2|_\pi}) - P(Y_{x_1|_\pi})$*, where* $P(Y_{x|_\pi})$ *represents the interventional distribution where the intervention is transferred only along* π.

In the total effect and path-specific effect, the intervention is performed on the whole population. The counterfactual effect measures the causal effect while the intervention is performed conditioning on only certain individuals or groups specified by a subset of observed variables $\mathbf{O} = \mathbf{o}$.

Definition 4. *Given a context* $\mathbf{O} = \mathbf{o}$*, the counterfactual effect of the value change of* X *from* x_1 *to* x_2 *on* Y *is given by* $CE(x_2, x_1|\mathbf{o}) = P(Y_{x_2}|\mathbf{o}) - P(Y_{x_1}|\mathbf{o})$.

2.3 Generative Adversarial Network

Generative Adversarial Networks (GAN) [Goodfellow *et al.*, 2014] are generative models that consist of two components: a generator and a discriminator. Typically, both the generator and discriminator are multilayer neural networks. Generator $G(\mathbf{Z})$ takes random noises $\mathbf{Z}$ as input and attempts to learn a generative distribution P_G to match the real data distribution P_{data}. On the contrary, the discriminative model D is a binary classifier that predicts whether an input is a real data $\mathbf{x}$ or a generated fake data from $G(\mathbf{Z})$. By playing the adversarial game, GAN is formalized as a minimax problem $\min_G \max_D V(G, D)$ with: $V(G, D) = \mathbb{E}_{\mathbf{x} \sim P_{data}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim P(\mathbf{Z})}[\log(1 - D(G(\mathbf{z})))]$, where $D(\cdot)$ outputs the probability that $\cdot$ is from real data rather than generated fake data.

2.4 CausalGAN

Research in [Kocaoglu *et al.*, 2018] shows that GAN can be modified to generate both observational and interventional distributions while preserving the causal structure among all attributes, referred to as the CausalGAN. Given a causal graph, generator $G(\mathbf{Z})$ attempts to play the role of a causal model that agrees with this causal graph in terms of both the graph structure and conditional distributions. To this end, noises $\mathbf{Z}$ are partitioned into $|\mathbf{V}|$ subsets $\{\mathbf{Z}_{V_1}, \mathbf{Z}_{V_2}, \ldots\}$, each of which $\mathbf{Z}_{V_i}$ plays the role of hidden variables $\mathbf{U}_{V_i}$. Similarly, generator $G(\mathbf{Z})$ is partitioned into $|\mathbf{V}|$ sub-neural

networks $\{G_{V_1}, G_{V_2}, \ldots\}$, each of which G_{V_i} plays the role of function f_{V_i} for generating the values of V_i. Then, if node V_j is a parent of V_i in the causal graph, the output of G_{V_j} is designed as an input of G_{V_i} to reflect this connection. Meanwhile, the adversarial game is played to ensure $P_G(G(\mathbf{Z}) = \mathbf{v}) = P(\mathbf{V} = \mathbf{v}), \forall \mathbf{v}$. The authors have proved that $G(\mathbf{Z})$ is consistent with any causal model that agrees with the same causal graph in terms of any identifiable interventional distributions, if: (1) $P(\mathbf{V})$ is strictly positive; (2) the connections of sub-neural networks G_{V_i} are arranged to reflect the causal graph structure; and (3) the generated observational distribution matches the real observational distribution, i.e., $P_G(G(\mathbf{Z}) = \mathbf{v}) = P(\mathbf{V} = \mathbf{v}), \forall \mathbf{v}$. Therefore, CausalGAN can be used to simulate the real causal model that agrees with the causal graph in identifiable situations.

3 CFGAN

To discuss the design of CFGAN, we first formulate our problem (Section 3.1), and then discuss the overall framework (Section 3.2). The CFGAN based on different fairness criteria will be discussed in Sections 3.3, 3.4 and 3.5. For all types of causal effects, we simply assume they are identifiable.

3.1 Problem Statement

In this paper, we follow the conventional notations in fairness-aware learning. We consider $\mathbf{V} = \{\mathbf{X}, Y, S\}$, where S denotes the sensitive variable, Y denotes the decision variable, and $\mathbf{X}$ denotes the set of all other variables (profile attributes). Given a causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and a dataset with m samples $(\mathbf{x}, y, s) \sim P_{data} = P(\mathbf{V})$, the goal of CFGAN is to (1) generate new data $(\hat{\mathbf{x}}, \hat{y}, \hat{s})$ which preserves the distribution of all attributes in the real data and (2) ensure that in the generated data $\hat{S}$ has no discriminatory effect on $\hat{Y}$ in terms of various causal-based criteria. Note that we use the hatted variables to denote the fake data generated by the generator. For ease of discussion, we consider both S and Y as binary variables, where s^+ denotes $S = 1$ and s^- denotes $S = 0$. It's straightforward to extend this to the multi-categorical or numerical cases. In this paper we mostly discuss the causal effect of a single variable S on another single variable Y. However, the model is capable to handle causal effects between multiple variables as well.

We consider causal fairness criteria based on total effect [Zhang and Bareinboim, 2018], direct discrimination [Zhang *et al.*, 2017], indirect discrimination [Zhang *et al.*, 2017], and counterfactual fairness [Kusner *et al.*, 2017], defined below.

Definition 5. *There is no total effect in the data if* $TE(s^+, s^-) = 0$.

Definition 6. *There is no direct discrimination in the data if* $SE_{\pi_d}(s^+, s^-) = 0$, *where* π_d *is the path set that only contains the direct edge from* S *to* Y, *i.e.,* $S \to Y$.

Definition 7. *Given a subset of attributes* $\mathbf{R} \subseteq \mathbf{X}$ *that cannot be objectively justified in decision making, there is no indirect discrimination in the data if* $SE_{\pi_i}(s^+, s^-) = 0$, *where* π_i *is the set of causal paths from* S *to* Y *that pass through* $\mathbf{R}$.

Definition 8. *Given a subset of attributes* $\mathbf{O} \subseteq \mathbf{X}$, *counterfactual fairness is achieved in the data if* $CE(s^+, s^-|\mathbf{o}) = 0$ *under any context* $\mathbf{O} = \mathbf{o}$.

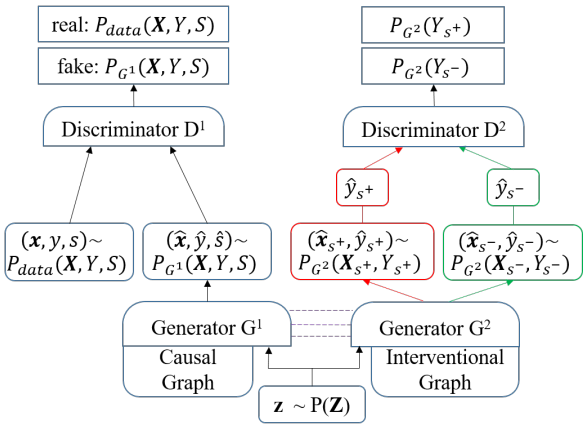


Figure 1: The framework of CFGAN

3.2 Model Framework

We propose the CFGAN model which consists of two generators (G^1, G^2) and two discriminators (D^1, D^2). Figure 1 shows the framework of CFGAN.

As shown in Sections 2.2 and 3.1, in general, causal-based fairness criteria compare the intervention distributions of Y under two different interventions $do(S = s^+)$ and $do(S = s^-)$. To implement these criteria, CFGAN adopts two generators. One generator G^1 plays the role of original causal model $\mathcal{M}$ similar to CausalGAN, while the other generator G^2 explicitly plays the roles of different interventional models $\mathcal{M}_s$ based on the type of causal effects. Generator G^1 aims to generate observational data whose distribution is close to the real observational distribution, and generator G^2 aims to generate interventional data that satisfy the criterion defined in Section 3.1. The two generators share the input noises and parameters to reflect the connections between the two causal models, and differ in connections of sub-neural networks to reflect the intervention. Then, CFGAN adopts two discriminators, where one discriminator D^1 tries to distinguish the generated data from the real data, and the other discriminator D^2 tries to distinguish the two intervention distributions under $do(S = s^+)$ and $do(S = s^-)$. Finally, generators and discriminators play the adversarial game to produce high quality fair data.

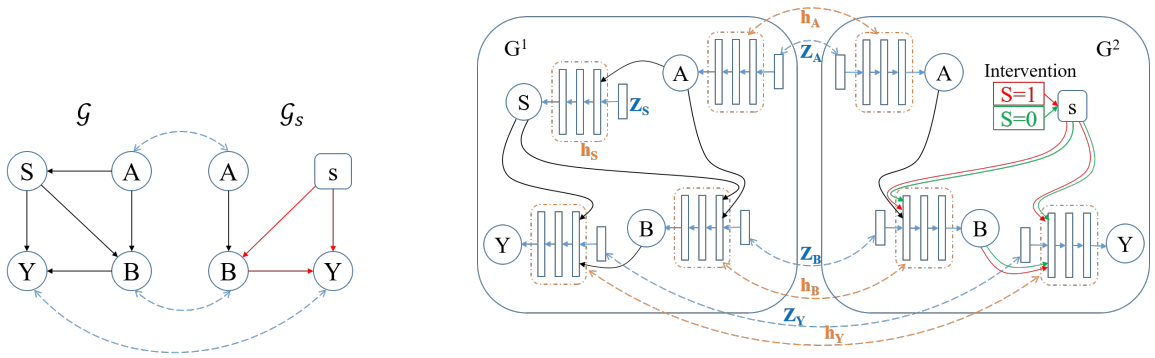
Next, we give the details in designing the generators and discriminators for different fairness criteria.

3.3 CFGAN based on Total Effect

We first show the CFGAN with no total effect (Definition 5).

Generators. Generator G^1 is designed to agree with the causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. It consists of $|\mathbf{V}|$ sub-neural networks, where each of them corresponds to a node in $\mathbf{V}$. All sub-neural networks are connected following the connections in $\mathcal{G}$. To be specific, each sub-neural network $G^1_{V_i}$ takes as input an independent noise vector $\mathbf{Z}_{V_i}$ as well as the output of any other sub-neural network $G^1_{V_j}$ if V_j is a parent of V_i in $\mathcal{G}$. Then, it outputs sample values of V_i, i.e., $\hat{v}_i$.

The other generator G^2 is designed to agree with the interventional graph $\mathcal{G}_s = (\mathbf{V}, \mathbf{E} \setminus \{V_j \to S\}_{V_j \in \mathbf{Pa}_S})$, where all incoming edges to S are deleted under intervention $do(S = s)$. The structure of G^2 is similar to that of G^1, except for



(a) Causal graph $\mathcal{G}$ and interventional graph $\mathcal{G}_s$

(b) Generators G^1 and G^2

Figure 2: An example of the generators G^1 and G^2 for CFGAN based on total effect. S is set to 1 or 0 to sample from the interventional distributions $P_{G^2}(A_{s^+}, B_{s^+}, Y_{s^+})$ (red) and $P_{G^2}(A_{s^-}, B_{s^-}, Y_{s^-})$ (green) respectively.

that sub-neural network G^2_S is set as $G^2_S \equiv 1$ if $s = s^+$, and $G^2_S \equiv 0$ if $s = s^-$. The two generators G^1 and G^2 are synchronized by sharing the same set of parameters for each pair of corresponding sub-neural networks, i.e., $G^1_{V_i}$ and $G^2_{V_i}$ for each V_i except for S, as well as the same noise vectors $\mathbf{Z} = \mathbf{z}$. As a result, G^1 can generate samples from the observational distribution, and G^2 can generate samples from two interventional distributions, i.e., $(\hat{\mathbf{x}}, \hat{y}, \hat{s}) \sim P_{G^1}(\mathbf{X}, Y, S)$, $(\hat{\mathbf{x}}_{s^+}, \hat{y}_{s^+}) \sim P_{G^2}(\mathbf{X}_{s^+}, Y_{s^+})$, if $s = s^+$, $(\hat{\mathbf{x}}_{s^-}, \hat{y}_{s^-}) \sim P_{G^2}(\mathbf{X}_{s^-}, Y_{s^-})$, if $s = s^-$.

Consider an example in Figure 2 which involves 4 variables $\{A, S, B, Y\}$. Figure 2a shows the causal graph $\mathcal{G}$ and the interventional graph $\mathcal{G}_s$ under $do(S = s)$, where the double headed arrows indicate the pair of nodes that share the same hidden variables and the function. Figure 2b shows the structures of the generators where G^1 agrees with $\mathcal{G}$ and G^2 agrees with $\mathcal{G}_s$. The double headed arrows indicate the sharing of noises and parameters of sub-neural networks. As shown, the edge from A to S is deleted in $\mathcal{G}_s$, which is also reflected in G^2. In addition, for each pair of nodes in the graphs, e.g., B in $\mathcal{G}$ and B in $\mathcal{G}_s$, the corresponding sub-neural networks are also synchronized, e.g., G^1_B and G^2_B.

Discriminators. Discriminator D^1 is designed to distinguish between the real observational data $(\mathbf{x}, y, s) \sim P_{data}(\mathbf{X}, Y, S)$ and the generated fake observational data $(\hat{\mathbf{x}}, \hat{y}, \hat{s}) \sim P_{G^1}(\mathbf{X}, Y, S)$. The other discriminator D^2 is designed to distinguish between the two interventional distributions $\hat{y}_{s^+} \sim P_{G^2}(Y_{s^+})$ and $\hat{y}_{s^-} \sim P_{G^2}(Y_{s^-})$.

Putting the generators and discriminators together, generator G^1 plays the adversarial game with the discriminator D^1, and generator G^2 plays the adversarial game with the discriminator D^2. The overall minimax game is described as:

$$\min_{G^1,G^2} \max_{D^1,D^2} J(G^1,G^2,D^1,D^2) = J_1(G^1,D^1) + \lambda J_2(G^2,D^2),$$

where

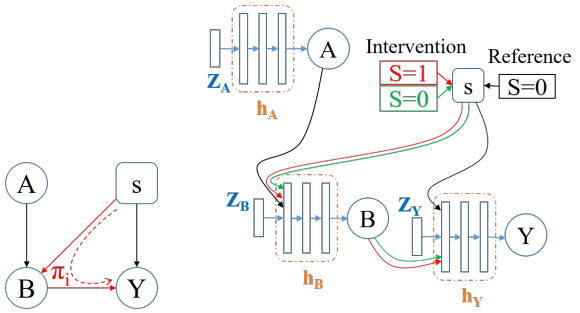
$$\begin{aligned} J_1(G^1, D^1) &= \mathbb{E}_{(\mathbf{x},y,s)\sim P_{data}(\mathbf{X},Y,S)}[\log D^1(\mathbf{x}, y, s)] \\ &\quad + \mathbb{E}_{(\hat{\mathbf{x}},\hat{y},\hat{s})\sim P_{G^1}(\mathbf{X},Y,S)}[1 - \log D^1(\hat{\mathbf{x}}, \hat{y}, \hat{s})], \\ J_2(G^2, D^2) &= \mathbb{E}_{\hat{y}_{s^+}\sim P_{G^2}(Y_{s^+})}[\log D^2(\hat{y}_{s^+})] \\ &\quad + \mathbb{E}_{\hat{y}_{s^-}\sim P_{G^2}(Y_{s^-})}[1 - \log D^2(\hat{y}_{s^-})], \end{aligned}$$

and λ is a hyperparameter which controls a trade-off between utility and fairness of data generation. The first value function J_1 aims to achieve $P_{G^1}(\mathbf{X}, Y, S) = P_{data}(\mathbf{X}, Y, S)$, i.e., to make the generated observational data indistinguishable from the real data. The second value function J_2 aims to achieve $P_{G^2}(Y_{s^+}) = P_{G^2}(Y_{s^-})$. Since Definition 5 requires $TE(s^+, s^-) = 0$, or equivalently $P(Y_{s^+}) = P(Y_{s^-})$, J_2 actually makes the generated interventional data satisfy the fairness criterion. As G^1 and G^2 share the same sets of parameters, the observational data generated by G^1 can be considered as being generated by a causal model which is close to the real causal model and also satisfies the fairness criterion. Finally, the generated fair data can be released to public.

3.4 CFGAN based on Direct and Indirect Discrimination

Both direct and indirect discrimination are based on path-specific effects. In this section, we focus on the indirect discrimination criterion, and direct discrimination criterion can be achieved similarly. Given a path set π_i that contains the paths pass through unjustified attributes, Definition 7 requires that $SE_{\pi_i}(s^+, s^-) = 0$, or equivalently $P(Y_{s^+|_{\pi_i}}) = P(Y_{s^-|_{\pi_i}})$ with reference s^-.

The design of generator G^1 is similar to that in Section 3.3, but G^2 is different in that it needs to simulate the situation where the intervention is transferred along π_i only. To this end, we first similarly design the structure of G^2 to agree with the interventional graph $\mathcal{G}_s = (\mathbf{V}, \mathbf{E} \setminus \{V_j \to S\}_{V_j \in \mathbf{Pa}_S})$. Then, we consider two types of value settings for sub-neural network G^2_S: the reference setting and the interventional setting. For the reference setting, G^2_S is always set as $G^2_S \equiv 0$. For the interventional setting, G^2_S is set as $G^2_S \equiv 1$ if $s = s^+$ and $G^2_S \equiv 0$ if $s = s^-$. On the other hand, each of other sub-neural networks may output two types of sample values according to the value setting of G^2_S, referred to as the reference value and interventional value respectively. For a sub-neural network, if its corresponding node is not on any path in π_i, it always takes reference values as input and outputs reference values. However, for any other sub-neural network $G^2_{V_j}$ that is on at least one path in π_i, it may take both types of val-



(a) $\mathcal{G}_s$ and π_i (b) Generator G^2

Figure 3: An example of the generator G^2 for CFGAN based on indirect discrimination. S is set to 1 or 0 and the transmission is set only along $\pi = \{S \to B \to Y\}$ to sample from the interventional distributions $P_{G^2}(A_{s+|_\pi}, B_{s+|_\pi}, Y_{s+|_\pi})$ (red) and $P_{G^2}(A_{s-|_\pi}, B_{s-|_\pi}, Y_{s-|_\pi})$ (green) respectively. S is set to be 0 for the reference setting.

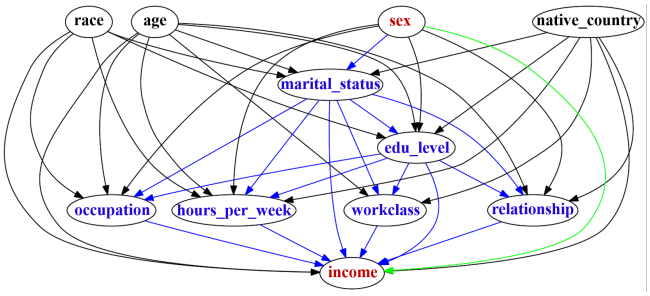


Figure 4: The causal graph for Adult dataset: the blue paths represent the indirect path set π_i.

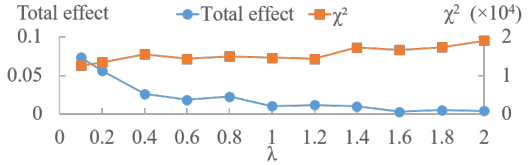


Figure 5: Total effect and χ^2 under different λ

ues as input and output both. Specifically, for any sub-neural network $G^2_{V_i}$ where V_i is a child of V_j, if edge $V_j \to V_i$ does not belong to any path in π_i, then $G^2_{V_j}$ will feed the reference output values to $G^2_{V_i}$. Otherwise, the interventional output values will be fed. As a result, the interventional distribution generated by G^2 simulates the situation of the path-specific effect, which we denote as $P_{G^2}(\mathbf{X}_{s|_\pi}, Y_{s|_\pi})$.

Consider an example (Figure 3) with the same causal graph in Figure 2. The interventional graph $\mathcal{G}_s$ and $\pi_i = \{S \to B \to Y\}$ is shown in Figure 3b, and generator G^2 is shown in Figure 3b. Since B is on the path in π_i, G^2_B takes interventional values of S as input and outputs interventional values to G^2_Y. On the other hand, G^2_Y takes interventional values from G^2_B and reference values from $G^2_S \equiv 0$ as input.

To achieve no indirect discrimination, discriminator D^2 is designed to distinguish between two interventional distributions $\hat{y}_{s+|_{\pi_i}} \sim P_{G^2}(Y_{s+|_{\pi_i}})$ and $\hat{y}_{s-|_{\pi_i}} \sim P_{G^2}(Y_{s-|_{\pi_i}})$. By playing the adversarial game with G^2, the corresponding value function J_2 aims to achieve $P_{G^2}(Y_{s+|_{\pi_i}}) = P_{G^2}(Y_{s-|_{\pi_i}})$. Similarly, since G^1 and G^2 share the parameters, the observational data generated by G^1 can also be considered as satisfying the no indirect discrimination criterion.

3.5 CFGAN for Counterfactual Fairness

In counterfactual fairness, the intervention is performed conditioning on a subset of variables $\mathbf{O} = \mathbf{o}$. Thus, different from previous fairness criteria that concern the interventional model only, counterfactual fairness concerns the connection between the original causal model and the interventional model. We reflect this connection in CFGAN by building a direct dependency between the samples generated by G^1 and the samples generated by G^2. Specifically, the structures of G^1 and G^2 are similar to those in Section 3.3. However, for each noise vector $\mathbf{z}$, we first generate the observational sample by using G^1, and observe whether in the sample we have $\mathbf{O} = \mathbf{o}$. Only for those noise vectors with $\mathbf{O} = \mathbf{o}$ in the generated samples, we use them for generating interventional samples by using G^2. Thus, the interventional distribution generated by G^2 is conditioned on $\mathbf{O} = \mathbf{o}$, denoted by $P_{G^2}(\mathbf{X}_s, Y_s|\mathbf{o})$. Finally, discriminator D^2 is designed to distinguish between $\hat{y}_{s+}|\mathbf{o} \sim P_{G^2}(Y_{s+}|\mathbf{o})$ and $\hat{y}_{s-}|\mathbf{o} \sim P_{G^2}(Y_{s-}|\mathbf{o})$, producing the value function that aims to achieve $P_{G^2}(Y_{s+}|\mathbf{o}) = P_{G^2}(Y_{s-}|\mathbf{o})$.

4 Experiments

4.1 Experiment Setup

The dataset we use for evaluation is the UCI Adult income dataset [Dheeru and Karra Taniskidou, 2017]. It contains 65,123 samples with 11 variables. Following the setting in [Zhang *et al.*, 2017], we binarize each attribute to reduce the complexity for causal graph discovery. We treat *sex* as the sensitive variable S, *income* as the decision variable Y. The estimated causal graph is shown in Figure 4.

We evaluate the performance of CFGAN in generating fair data for different types of causal fairness. The fairness threshold is 0.05, i.e., the effect should be in $[-0.05, 0.05]$ to be fair. We compare CFGAN with other data generating approaches for different fairness respectively as other approaches may only be able to achieve one or two types of fairness.

Specifically, we consider two baselines: (1) the original dataset; and (2) CausalGAN [Kocaoglu *et al.*, 2018], which preserves the causal structure of the original data but is unaware of the fairness constraint. For total effect, we compare with FairGAN [Xu *et al.*, 2018], which removes all information correlated to S in other attributes. For indirect discrimination (we skip the results for direct discrimination as the original dataset contains no direct discrimination), we further compare with PSE-DR [Zhang *et al.*, 2017], which is a direct/indirect discrimination removing algorithm by modifying the causal graph and generating new fair data based on the modified causal graph. For counterfactual fairness, we instead compare with A1 and A3 [Kusner *et al.*, 2017]. A1

	Total effect	Indirect discrimination	χ^2	Classifier accuracy			
				SVM	DT	LR	RF
Real data	0.1936	0.1754	0	0.8178	0.8177	0.8170	0.8178
CausalGAN	0.1721	0.1508	14482	0.8143	0.8136	0.8160	0.8137
FairGAN	0.0021	0.0133	41931	0.8088	0.8081	0.8136	0.8082
PSE-DR	NA	0.0243	12468	0.8073	0.8073	0.8128	0.8075
CFGAN (TE)	0.0102	NA	14566	0.8134	0.8126	0.8120	0.8127
CFGAN (SE)	NA	0.0030	19724	0.8037	0.8030	0.8103	0.8024

Table 1: The total effect and indirect discrimination of real and generated datasets

	Counterfactual effect				χ^2	Classifier accuracy			
	$\mathbf{o}_1$	$\mathbf{o}_2$	$\mathbf{o}_3$	$\mathbf{o}_4$		SVM	DT	LR	RF
Real data	0.2023	0.1293	0.1266	0.1785	0	0.8178	0.8177	0.8170	0.8178
CausalGAN	0.1824	0.1155	0.1466	0.0959	14482	0.8143	0.8136	0.8160	0.8137
A1	0.0000	0.0000	0.0000	0.0000	17757	0.7615	0.7615	0.7615	0.7615
A3	0.2159	0.1127	0.1056	0.1860	12313	0.8159	0.8159	0.8159	0.8159
CFGAN (CE)	0.0209	0.0034	-0.0030	-0.0482	13904	0.8130	0.8123	0.8130	0.8115

Table 2: The counterfactual effect of real and generated datasets ($\mathbf{O} = \{race, native\_country\}$)

generates fair decisions using a classifier that is built on non-descendants of S. A3 is similar to A1 but presupposes an additive noise model for estimating noise terms, which are then used for building the classifier. For both A1 and A3, we use SVM as the classifier for generating fair decisions.

For data utility, we compute the χ^2 distance, where a smaller χ^2 indicates better utility. We also use the generated data to train classifiers and measure the accuracy. We evaluate 4 classifiers: support vector machine (SVM), decision tree (DT), logistic regression (LR) and random forest (RF).

4.2 Total Effect

We calculate the total effect for the original dataset and different generated datasets. The results are shown in Table 1. As can be seen, the original data has a total effect of 0.1936, and CausalGAN preserves similar total effect. FairGAN produces no total effect, but with the worst utility in terms of χ^2. This may be because FairGAN removes too much information due to its causal blindness. The generated data by CFGAN based on total effect (CFGAN (TE), $\lambda = 1$) produces no total effect, and also preserves good data utility.

4.3 Indirect Discrimination

For indirect discrimination, we consider all the paths passing through $marital\_status$ as π_i. The results are also shown in Table 1. Similar to total effect, CausalGAN preserves indirect discrimination close to the original data, and FairGAN removes indirect discrimination but causes the largest utility loss. On the other hand, PSE-DR and our method (CFGAN (SE), $\lambda = 1$) can remove indirect discrimination and also have good data utility. We see that the two methods achieve comparable performance based on different techniques.

4.4 Counterfactual Fairness

For counterfactual fairness, we consider the observation of two attributes, i.e., $\mathbf{O} = \{race, native\_country\}$, which has 4 value combinations. Table 2 shows the results for all 4 subgroups. As can be seen, the original data and CausalGAN contain biases in terms of counterfactual fairness in all subgroups. A1 is counterfactual fair as expected since it is proved to be so in [Kusner *et al.*, 2017]. However, the data utility is bad especially in terms of classifier accuracy, since it only uses non-descendants of sex in labeling decisions. A3 cannot achieve counterfactual fairness, probably because its linear assumption does not fit the original data well. Finally, our method (CFGAN (CE), $\lambda = 1$) achieves both counterfactual fairness and good data utility.

4.5 Parameter Sensitivity

We evaluate the trade-off between utility and fairness when changing λ in the overall minimax game. A larger λ indicates a stronger enforcement on the fairness and compromise on utility. Figure 5 shows the results for total effect, where we get a fairly good trade-off between utility and fairness at $\lambda = 1$. We observe similar results for other fairness types.

5 Conclusions

We proposed the causal fairness-aware generative adversarial networks (CFGAN) for generating high quality fair data. We considered various causal-based fairness criteria, including total effect, direct discrimination, indirect discrimination, and counterfactual fairness. CFGAN consists of two generators and two discriminators. The two generators aim to simulate the original causal model and the interventional model. This is achieved by arranging the neural network structure of the generators following the original causal graph and the interventional graph. Then, two discriminators are adopted for achieving both the high data utility and causal fairness. Experiments using the Adult dataset showed that CFGAN can achieve all types of fairness with relatively small utility loss.

Acknowledgments

This work was supported in part by NSF 1646654, 1564250, and 1841119.

References

[Calders *et al.*, 2009] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 2009.

[Chiappa, 2019] Silvia Chiappa. Path-Specific Counterfactual Fairness. *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[Dheeru and Karra Taniskidou, 2017] Dua Dheeru and Efi Karra Taniskidou. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017.

[Edwards and Storkey, 2016] Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico*, 2016.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, New York, NY, USA, 2015.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. 2014.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, None, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

[Kamiran and Calders, 2009] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, February 2009.

[Kamiran and Calders, 2012] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, October 2012.

[Kocaoglu *et al.*, 2018] Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *6th International Conference on Learning Representations*, 2018.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 2017.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm, Sweden*, 2018.

[Nabi and Shpitser, 2018] Razieh Nabi and Ilya Shpitser. Fair Inference On Outcomes. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[Pedreshi *et al.*, 2008] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, New York, New York, USA, 2008.

[Salimi *et al.*, 2019] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal database repair for algorithmic fairness. *CoRR*, abs/1902.08283, 2019.

[Wu *et al.*, 2019] Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 2019.

[Xie *et al.*, 2017] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable Invariance through Adversarial Feature Learning. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

[Xu *et al.*, 2018] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*, pages 570–575, 2018.

[Zhang and Bareinboim, 2018] Junzhe Zhang and Elias Bareinboim. Fairness in Decision-Making – The Causal Explanation Formula. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[Zhang *et al.*, 2017] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia*, 2017.

[Zhang *et al.*, 2018a] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA*, 2018.

[Zhang *et al.*, 2018b] Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving non-discrimination in prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 3097–3103, 2018.

[Zhang *et al.*, 2018c] Lu Zhang, Yongkai Wu, and Xintao Wu. Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[Zliobaite *et al.*, 2011] Indre Zliobaite, Faisal Kamiran, and Toon Calders. Handling Conditional Discrimination. In *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011.