

Counterfactual Fairness in Synthetic Data Generation

A Proofs

Definition. The total variation distance for two discrete probability distributions P and Q , defined on a countable space Ω , is defined as

$$\mathbb{TV}(P, Q) = \frac{1}{2} \sum_{x \in \Omega} |P(x) - Q(x)|. \quad (1)$$

This is equivalent with this more general definition: $\mathbb{TV}(P, Q) = \sup_{A \in \Omega} |P(A) - Q(A)|$.

Proof of Proposition 1: Since \mathbb{TV} satisfies triangle inequality, we have

$$\begin{aligned} \mathbb{TV}(P'(f(X)|A=0), P'(f(X)|A=1)) &\leq \mathbb{TV}(P'(f(X)|A=0), P'(Y|A=0)) \\ &\quad + \mathbb{TV}(P'(Y|A=0), P'(Y|A=1)) \\ &\quad + \mathbb{TV}(P'(Y|A=1), P'(f(X)|A=1)). \end{aligned} \quad (2)$$

The second term above is bounded by δ since P' approximately satisfies SP. Now for the first term of RHS we have:

$$\begin{aligned} &\mathbb{TV}(P'(f(X)|A=0), P'(Y|A=0)) \\ &= \frac{1}{2} \sum_{y \in \{0,1\}} |P'(f(X)=y|A=0) - P'(Y=y|A=0)| \\ &= \frac{1}{2} \sum_{y \in \{0,1\}} |P'(f(X)=y, Y=y|A=0) + P'(f(X)=y, Y=1-y|A=0) - P'(Y=y|A=0)| \\ &= \frac{1}{2} \sum_{y \in \{0,1\}} |P'(f(X)=y, Y=y|A=0) + P'(f(X)=1-y, Y=y|A=0) \\ &\quad - P'(f(X)=1-y, Y=y|A=0) + P'(f(X)=y, Y=1-y|A=0) - P'(Y=y|A=0)| \\ &= \frac{1}{2} \sum_{y \in \{0,1\}} |P'(Y=y|A=0) - P'(f(X)=1-y, Y=y|A=0) \\ &\quad + P'(f(X)=y, Y=1-y|A=0) - P'(Y=y|A=0)| \\ &= \frac{1}{2} \sum_{y \in \{0,1\}} |P'(f(X)=y, Y=1-y|A=0) - P'(f(X)=1-y, Y=y|A=0)| \\ &\leq \frac{1}{2} \sum_{y \in \{0,1\}} |P'(f(X)=y, Y=1-y|A=0)| \\ &\leq \frac{1}{2} \sum_{y \in \{0,1\}} \frac{P'(f(X)=y, Y=1-y)}{P'(A=0)} \\ &\leq \epsilon/2P'(A=0) \end{aligned}$$

Similarly for the third term of RHS of (1) we have:

$$\mathbb{TV}(P'(Y|A=1), P'(f(X)|A=1)) \leq \epsilon/2P'(A=1).$$

This completes the proof.

Proof of Proposition 2: The error probability of f on P' is less than ϵ , that is

$$\Pr\{f(X) \neq Y\} = E_{X,Y \sim P'(X,Y)} \mathbb{1}[f(X) \neq Y] \leq \epsilon. \quad (3)$$

Now, the error probability of f on distribution P can be upper bounded as follows:

$$\begin{aligned} E_{X,Y \sim P(X,Y)} \mathbb{1}[f(X) \neq Y] &= \sum_{x,y} p(x,y) \mathbb{1}[f(x) \neq y] \\ &\leq \sum_{x,y} p'(x,y) \mathbb{1}[f(x) \neq y] + \sum_{x,y} |p(x,y) - p'(x,y)| \mathbb{1}[f(x) \neq y] \\ &\leq \epsilon + \sum_{x,y} |p(x,y) - p'(x,y)| \\ &\leq \epsilon + 2\delta \end{aligned}$$

Proof of Proposition 3: Similar to Proposition 1, using triangle inequality for \mathbb{TV} , we have

$$\begin{aligned} \mathbb{TV}(P(f(X)|A=0), P(f(X)|A=1)) &\leq \mathbb{TV}(P(f(X)|A=0), P'(f(X)|A=0)) \\ &\quad + \mathbb{TV}(P'(f(X)|A=0), P'(f(X)|A=1)) \\ &\quad + \mathbb{TV}(P'(f(X)|A=1), P(f(X)|A=1)). \end{aligned} \quad (4)$$

The second term above is bounded by δ_1 (since f approximately satisfies SP for P'), thus the RHS is upper bounded by

$$\mathbb{TV}\left(\frac{P(f(X), A=0)}{P(A=0)}, \frac{P'(f(X), A=0)}{P'(A=0)}\right) + \mathbb{TV}\left(\frac{P'(f(X), A=1)}{P(A=1)}, \frac{P(f(X), A=1)}{P(A=1)}\right) + \delta_1. \quad (5)$$

For the first term we have:

$$\begin{aligned} &\mathbb{TV}\left(\frac{P(f(X), A=0)}{P(A=0)}, \frac{P'(f(X), A=0)}{P'(A=0)}\right) \\ &= \sum_{y \in \{0,1\}} \left| \frac{1}{P(A=0)} P(f(X)=y, A=0) - \frac{1}{P'(A=0)} P'(f(X)=y, A=0) \right| \end{aligned} \quad (6)$$

Now for $y=0$ if the first term in (6) is larger than the second term then we have:

$$\begin{aligned} &\frac{1}{P(A=0)} P(f(X)=0, A=0) - \frac{1}{P'(A=0)} P'(f(X)=0, A=0) \\ &\leq \frac{1}{P(A=0)} P(f(X)=0, A=0) - \frac{1}{P(A=0) + \delta_2} P'(f(X)=0, A=0) \end{aligned} \quad (7)$$

$$\leq \frac{z + \delta_2}{P(A=0)} - \frac{z}{P(A=0) + \delta_2} \quad (8)$$

$$\leq \frac{\delta_2(1 + P(A=0)) + \delta_2^2}{P(A=0)^2} \quad (9)$$

In above equations we used the definition of total variation ($\mathbb{TV}(P, Q) = \sup_A |P(A) - Q(A)|$). Now if we consider the other case that the second term in (6) is larger than the first term for $y=0$ we get the following upper bound:

$$\frac{\delta_2(1 + P'(A=0)) + \delta_2^2}{P'(A=0)^2} \quad (10)$$

Since $g(p) = \frac{\delta_2(1+p) + \delta_2^2}{p^2}$ is a decreasing function in p (for $p > 0$) if we define $p_0 = \min\{P(A=0), P'(A=0)\}$ we can bound the (6) with

$$\frac{2\delta_2(1 + p_0) + 2\delta_2^2}{p_0^2}. \quad (11)$$

Similarly if we define $p_1 = \min\{P(A = 1), P'(A = 1)\}$, then the second term in (5) can be upper bounded with

$$\frac{2\delta_2(1 + p_1) + 2\delta_2^2}{p_1^2}. \quad (12)$$

This completes the proof.

Proof of Proposition 4: We want to prove the following equation, assuming that P' is the distribution induced from G which satisfies $G_Y(u, a) = G_Y(u, a')$:

$$P'(Y_{A \leftarrow a} = y | X = x, A = a) = P'(Y_{A \leftarrow a'} = y | X = x, A = a). \quad (13)$$

Let us denote with Q the posterior distribution of U conditioned on A and X . Then the LHS can be written as follows (we assume that \mathcal{U} is a countable set):

$$P'(Y_{A \leftarrow a} = y | X = x, A = a) = \sum_u P'(Y_{A \leftarrow a} = y | X = x, A = a, U = u) Q(U = u | X = x, A = a) \quad (14)$$

$$= \sum_u P'(Y = y | A = a, U = u) Q(U = u | X = x, A = a) \quad (15)$$

Similarly we have:

$$P'(Y_{A \leftarrow a'} = y | X = x, A = a) = \sum_u P'(Y_{A \leftarrow a'} = y | X = x, A = a, U = u) Q(U = u | X = x, A = a) \quad (16)$$

$$= \sum_u P'(Y = y | A = a', U = u) Q(U = u | X = x, A = a) \quad (17)$$

Now comparing (15) and (17) and noting that $G_Y(u, a) = G_Y(u, a')$ completes the proof.

B Comparison to previous definitions of fairness in SDG

Comparison with the definition in FairGAN [1]: In section 4.1 of [1], authors define their goal to be: Given a dataset $\{X, Y, S\} \sim P_{\text{data}}$, FairGAN aims to generate a fair dataset $\{\hat{X}, \hat{Y}, \hat{S}\} \sim P_G$ which achieves the statistical parity w.r.t the protected attribute \hat{S} , i.e., $P(\hat{Y} = 1 | \hat{S} = 1) = P(\hat{Y} = 1 | \hat{S} = 0)$. Meanwhile, our goal is to ensure that given a generated dataset $\{\hat{X}, \hat{Y}\}$ as training samples, a classification model seeks an accurate function $\eta : \hat{X} \rightarrow \hat{Y}$ while satisfying fair classification with respect to the protected attribute on the real dataset, i.e., $P(\eta(X) = 1 | \hat{S} = 1) = P(\eta(X) = 1 | \hat{S} = 0)$.

In their proposed method they attempt to achieve $I(\hat{X}, \hat{Y}; \hat{S}) = 0$. This is because the second discriminator is given both \hat{X} and \hat{Y} and is expected to estimate \hat{S} . The goal is to create samples such that discriminator 2 cannot find \hat{S} , thus the mechanism attempts to make $I(\hat{X}, \hat{Y}; \hat{S}) = 0$. Now note that for statistical parity we only need $I(\hat{Y}; \hat{S}) = 0$, the addition of \hat{X} was required to ensure that $I(\eta(X); Y)$ is also zero, thus the predictor will also satisfy statistical parity constraint. Notice that this is a very strong condition, and with this definition, *any* predictor will be fair. This is in contrast with our definition that only expects accurate predictors to be fair.

Comparison with the definition in CFGAN [2]: Similar to our work, CFGAN attempts to produce a distribution P' such that P' satisfies the fairness notion of choice. However, there is no formal definition or discussion on how such a data will work on the real data (with distribution P).

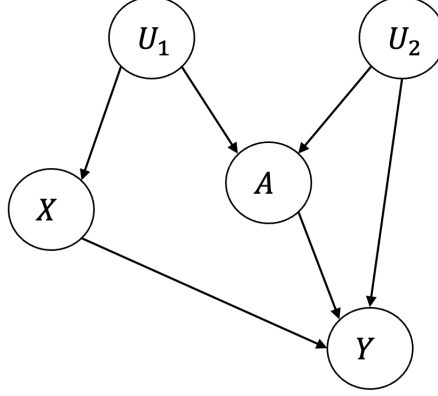


Figure 1: Example of a more complicated causal graph

B.1 Comparison with the definition in DECAF [3]

First let us review the definition of fair synthetic data given in [3]:

Definition 1. A probability distribution $P'(X)$ is $(\mathcal{I}(A, Y), P)$ -fair, iff the optimal predictor $\hat{Y} = f^*(X)$ of Y trained on $P'(X)$ satisfies $\mathcal{I}(A, Y)$ when evaluated on $P(X)$.

Closeness of P' and P : This definition is ideal from the fairness point of view, and gives us exactly what we need. That is having a distribution that when used for training a predictor will give us a predictor that is fair not on the training data but on the unseen real data. However, this definition (by itself) does not guarantee any resemblance of the synthetic data with the actual data. In fact, it is possible to consider a scenario where the support of the distribution P' is disjoint from the support of P and this definition still holds. This by itself is not a problem, but it is unclear how in practice one can impose this definition and also ensure that P' and P are close (closeness of P and P' is required since we want the synthetic data to look like actual data). For example, in DECAF work, there is no guarantee that P' and P are even remotely close. In fact, it is not hard to create two distributions using their method such that P' and P have an arbitrarily large KL distance. For example, consider a dataset where we have $A \rightarrow X \rightarrow Y$. Assume that A is a Bernoulli binary random variable ($P(A = 0) = P(A = 1)$). When $A = 0$, then $X = 0$ and when $A = 1$ then $X = m$. Also assume that $Y = \mathcal{N}(X, 1)$. Now, considering DECAF method, (e.g., for satisfying SP) we need to remove both edges from A to X and then from X to Y , then Y will be either constant or a distribution independent from X and so m (depending on which strategy one chooses as explained in Section 5.2). Thus, it is clear that by increasing m we can have arbitrarily large KL-distance between P and P' .

Proof of Proposition 1 in [3]: The proof of proposition 1 in [3] seems incomplete. Firstly, there seems to be a typo in the proof. $f^*(X)$ is the ideal classifier trained on P' not P , thus we cannot assume that $f^*(X) = P(Y|X)$, we may assume that $f^*(X) = P'(Y|X)$ (although it is unclear why this would be ideal classifier for the real data). Then we have $P'(Y|X) = P'(Y|\partial_{\mathcal{G}'}Y)$. The missing step is why $P'(Y|\partial_{\mathcal{G}'}Y) = P(Y|\partial_{\mathcal{G}'}Y)$ holds. Note, that if we use the method suggested in the paper this equation does not hold. For instance, assume that only one edge is removed have:

$$P'(Y|\partial_{\mathcal{G}'}Y) = P(Y|\partial_{\mathcal{G}'}Y, do(X_i = \tilde{x}_{ij})) \neq P(Y|\partial_{\mathcal{G}'}Y). \quad (18)$$

C Generalization of counterfactual fairness

In this section by an example we explain in more details how to generalize our proposed method for a given causal graph. Consider the causal graph in Figure 1. Here we have two unobserved variables U_1 and U_2 , X represents features, A is the sensitive attribute, and Y is the output. Considering this graph, U_1 and U_2 will be the noise for the generators (note that their underlying distribution is known, and also we assumed that unobserved variables are independent). Then we will have two generators G_1 and G_2 to produce X and A given U_1 and U_1, U_2 respectively. Now all variables to generate Y are available. G_3 will have A , X , and U_2 as input and will produce Y as the output. The generator architecture is represented in Figure 1.

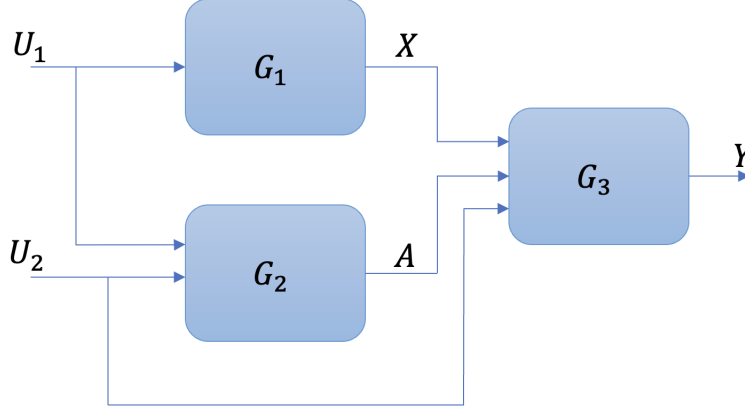


Figure 2: The GAN structure corresponding to Causal Graph in Figure 1

The generated samples X, A, Y will be fed to a discriminator. Also, for each sample we can alternate the value of A while the values of U_1 and U_2 and X is fixed to get the counterfactual output and then we can add counterfactual loss to the loss function of the generator.

D Reproducibility

The codes for generating our experiment results can be found here: <https://github.com/MahedAb/FairSyn>

References

- [1] D. Xu, S. Yuan, L. Zhang, and X. Wu, “Fairgan: Fairness-aware generative adversarial networks,” in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575, IEEE, 2018.
- [2] D. Xu, Y. Wu, S. Yuan, L. Zhang, and X. Wu, “Achieving causal fairness through generative adversarial networks,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.
- [3] B. van Breugel, T. Kyono, J. Berrevoets, and M. van der Schaar, “Decaf: Generating fair synthetic data using causally-aware generative networks,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.