Understanding Unfairness via Training Concept Influence

Yuanshun Yao [1] **Yang Liu** [1 2]

Abstract

Knowing the causes of a model's unfairness helps practitioners better understand their data and algorithms. This is an important yet relatively unexplored task. We look into this problem through the lens of the training data – one of the major sources of unfairness. We ask the following questions: how would a model's fairness performance change if, in its training data, some samples (1) were collected from a different (*e.g.* demographic) group, (2) were labeled differently, or (3) some features were changed? In other words, we quantify the fairness influence of training samples by counterfactually intervening and changing samples based on predefined concepts, *i.e.* data attributes such as features (X), labels (Y), or sensitive attributes (A). To calculate a training sample's influence on the model's unfairness w.r.t a concept, we first generate *counterfactual samples* based on the concept, *i.e.* the counterfactual versions of the sample if the concept were changed. We then calculate the resulting impact on the unfairness, via *influence function* (Koh & Liang, 2017; Rousseeuw et al., 2011), if the counterfactual samples were used in training. Our framework not only helps practitioners understand the observed unfairness and repair their training data, but also leads to many other applications, *e.g.* detecting mislabeling, fixing imbalanced representations, and detecting fairness-targeted poisoning attacks.

1. Introduction

A fundamental question in machine learning fairness is: what causes unfairness? Without knowing the answer, it is hard to understand and fix the unfairness problem. In practice, this is also one of the first questions the practitioners would ask after calculating the fairness measures and finding the model unfair. Although the question sounds simple, it is difficult to identify the exact source of unfairness in the machine learning pipeline, as admitted by many leading fairness practitioners, *e.g.* Meta (met, 2021) describes: "Unfairness in an AI model could have many possible causes, including not enough training data, a lack of features, a misspecified target of prediction, or a measurement error in the input features. Even for the most sophisticated AI researchers and engineers, these problems are not straightforward to fix."

The sources of unfairness are many, including data sampling bias or under-representation (Chai & Wang, 2022; Zhu et al., 2022; Celis et al., 2021; Bagdasaryan et al., 2019), data labeling bias (Wang et al., 2021; Wu et al., 2022b; Fogliato et al., 2020), model architecture (or feature representation) (Adel et al., 2019; Madras et al., 2018; Zemel et al., 2013; Song et al., 2019; Xing et al., 2021; Li et al., 2021a; Song et al., 2021; Li et al., 2020), distribution shift (Ding et al., 2021; Chen et al., 2022; Rezaei et al., 2021; Giguere et al., 2022) *etc.* In this work, we tackle this problem by looking at the most important and obvious source of bias – the training samples. It is because if the model's training samples are biased, then it would be unlikely the model can still remain fair without paying heavy costs later on. Specifically, we ask the following questions regarding how training samples would impact the model's unfairness: how a model's fairness measure would change if its training samples (1) were collected from a different (*e.g.* demographic) group, (2) were labeled differently, or (3) some of the features were changed? Answering those questions can help practitioners (1) *explain* the cause of the model's unfairness in terms of training data, (2) *repair* the training data to improve fairness, and (3) *detect* biased or noisy training labels, under-represented group, and corrupted features that hurt fairness.

In this work, we measure the training sample's impact on fairness using *influence function* (Cook & Weisberg, 1982; Koh & Liang, 2017), and we define the influence on fairness measure w.r.t a training *concept* – a categorical variable that describes data property. For example, we can choose the concept to be the sensitive group attribute and counterfactually intervene on it to answer the question "What is the impact on fairness if training data were sampled from a different group?" Or we can choose the concept to be the training labels, and then our method measures the impact on

[1]ByteDance Research [2]University of California, Santa Cruz. Correspondence to: Yuanshun Yao <kevin.yao@bytedance.com>.

fairness when the label is changed. We can also apply the concept to the training features or to the existence of training samples. Our flexible framework generalizes the prior works that only consider removing or reweighing training samples (Wang et al., 2022a; Li & Liu, 2022), and we can provide a broader set of explanations and give more insights to practitioners in a wider scope (*e.g.* what if a data pattern is drawn from another demographic group?). We name our influence framework as *Concept Influence for Fairness* (CIF).

In addition to explaining the unfairness, CIF can also recommend practitioners ways to fix the training data to improve fairness by counterfactually intervening in the concepts. Furthermore, our framework leads to a number of other applications including (1) detecting mislabeling, (2) detecting poisoning attacks, and (3) fixing imbalanced representation. Through experiments on 4 datasets – including synthetic, tabular, and image – we show that our method achieves satisfactory performance in a wide range of tasks.

2. Influence of Training Concepts

We start with introducing the influence function for fairness, the concept in training data, and define our *Concept Influence for Fairness* (CIF).

2.1. Fairness Influence Function

Influence Function on Group Fairness. Denote the training data by $D_{train} = \{z_i^{tr} = (x_i^{tr}, y_i^{tr})\}_{i=1}^n$ and the validation data by $D_{val} = \{z_i^{val} = (x_i^{val}, y_i^{val})\}_{i=1}^n$. Suppose the model is parameterized by $\theta \in \Theta$, and there exist a subset of training data with sample indices $\mathcal{K} = \{K_1, ..., K_k\}$. If we perturb a group $\mathcal{K}$ by assigning each sample $i \in \mathcal{K}$ with weight $w_i \in [0, 1]$, denote the resulting counterfactual model's weights by $\hat{\theta}_{\mathcal{K}}$.

Definition 1. *The fairness influence of reweighing group $\mathcal{K}$ in the training data is defined as the difference of fairness measure between the original model $\hat{\theta}$ (trained on the full training data) and the counterfactual model $\hat{\theta}_{\mathcal{K}}$:*

$$\textit{infl}(D_{val}, \mathcal{K}, \hat{\theta}) := \ell_{\textit{fair}}(\hat{\theta}) - \ell_{\textit{fair}}(\hat{\theta}_{\mathcal{K}}) \tag{1}$$

where ℓ_{fair} is the fairness measure (will be specified shortly after).

Similar to (Koh & Liang, 2017; Koh et al., 2019; Li & Liu, 2022), we can derive the closed-form solution of fairness influence function:

Proposition 1. *The first-order approximation of $\textit{infl}(D_{val}, \mathcal{K}, \hat{\theta})$ takes the following form:*

$$\textit{infl}(D_{val}, \mathcal{K}, \hat{\theta}) \approx -\nabla_\theta \ell_{\textit{fair}}(\hat{\theta})^\intercal H_{\hat{\theta}}^{-1} \left(\sum_{i \in \mathcal{K}} w_i \nabla \ell(z_i^{tr}; \hat{\theta}) \right) \tag{2}$$

where $H_{\hat{\theta}}$ is the hessian matrix i.e. $H_{\hat{\theta}} := \frac{1}{n} \nabla^2 \sum_{i=1}^n \ell(z_i^{tr}; \hat{\theta})$*, and ℓ is the original loss function (e.g. cross-entropy loss in classification).*

See Appendix A for the derivation.

Approximated Fairness Loss. The loss $\ell_{fair}(\hat{\theta})$ quantifies the fairness of a trained model $\hat{\theta}$. Similarly to prior work (Wang et al., 2022a; Sattigeri et al., 2022), we can approximate it with a surrogate loss on the validation data. Denote the corresponding classifier for θ as h_θ, we can approximate the widely used group fairness Demographic Parity (Calders et al., 2009; Chouldechova, 2017) (DP) violation as the following (assume both A and the classification task are binary):

$$\begin{aligned} \ell_{DP}(\hat{\theta}) &:= \left| \mathbb{P}(h_\theta(X) = 1 | A = 0) - \mathbb{P}(h_\theta(X) = 1 | A = 1) \right| \\ &\approx \left| \frac{\sum_{i \in D_{val}: a_i = 0} g(z_i^{val}; \theta)}{\sum_{i \in D_{val}} \mathbb{I}[a_i = 0]} - \frac{\sum_{i \in D_{val}: a_i = 1} g(z_i^{val}; \theta)}{\sum_{i \in D_{val}} \mathbb{I}[a_i = 1]} \right| \end{aligned} \tag{3}$$

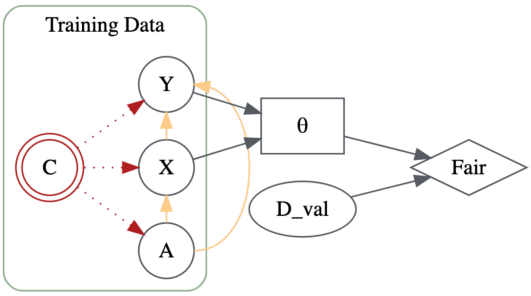
where g is the logit of the predicted probability for class 1. See Appendix B for the approximated violation of Equality of Opportunity (Hardt et al., 2016) (EOP), and Equality of Odds (Woodworth et al., 2017) (EO).

2.2. Concepts in Training Data

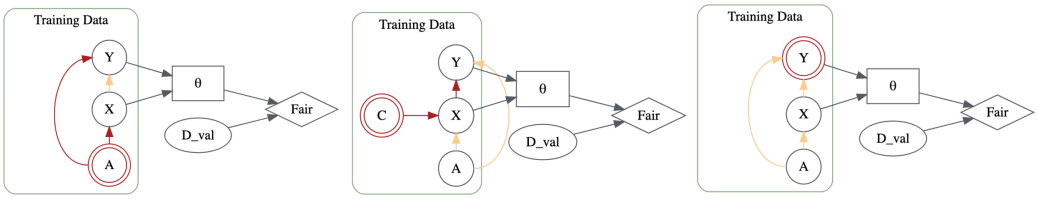
A concept is a *sample-level* categorical attribute associated with the training data. Formally, denote a concept by $C \in \mathcal{C} := \{1, 2, ..., c\}$ where C is a discrete concept that is encoded in the data (X, Y, A). We do not exclude the possibility that C can simply be either Y or A or any feature in X, but C can be broader. See Figure 1 for an illustration. Our core idea is to quantify the influence when each training sample is replaced by its "counterfactual sample" (*i.e.* the counterfactual version of the sample if its concept were changed) when we intervene on a certain concept.

Examples. We provide detailed examples of concepts and motivate why intervening on those concepts can be intuitively helpful for fairness.

- **Concept as Sensitive Attribute ($C = A$).** Intuitively speaking, the sensitive/group attribute relates closely to fairness measures due to its importance in controlling the sampled distribution of each group. Intervening on A corresponds to asking counterfactually what if a similar or counterfactual sample were from a different sensitive group.
- **Concept as Label ($C = Y$).** In many situations, there are uncertainties in the label $Y|X$. Some other times, the observed Y can either encode noise, mislabeling or subjective biases. They can all contribute to unfairness. Intervening on Y implies the counterfactual effect if we



(a) The overall causal graph of training concepts.



(b) Intervening sensitive attribute A. (c) Intervening feature X. (d) Intervening label Y.

Figure 1. (a) shows the overall causal graph assumed in our work. In training data, the concept variable C can intervene feature X, label Y, or sensitive attribute A. The model θ is trained on X and Y, and together with the validation dataset D_{val}, the validation fairness metric `Fair` is computed. Figure (b), (c), and (d) show the individual case when the concept variable intervenes A, X, and Y separately.

were to change the label (*e.g.* a sampling, a historical decision, or a human judgment) of a sample.

- **Concept as Predefined Feature Attribute ($C = attr(X)$).** Our framework allows us to predefine a relevant concept based on feature X. C can be either an externally labeled concept (*e.g.* sample-level label in image data) or a part of X (*e.g.* a categorical[1] feature in tabular data). For instance, if we want to understand how skin color would affect the model's fairness, and if so which data samples would impact the observed fairness the most w.r.t skin color, we can specify $C = attr(image) \in \{\text{dark}, \text{light}\}$. Then intervening on this concept corresponds to identifying samples from different skin colors that, if were included in the training data, would lead to a fairer model.
- **Concept as Removal.** Our setting is also flexible enough to consider the effect of removing a training sample, as commonly considered in the literature on influence function (Li & Liu, 2022). Consider there is a selection variable $S \in \{1, 0\}$ for each instance z_i^{tr} – for each of these samples that appear in the training data we have $s_i = 1$. Changing to $s_i = 0$ means the sample is counterfactually not included, *i.e.* $\hat{z}_i^{tr}(c') = \varnothing$. Allowing the concept to be removed, we can incorporate the prior works on the influence of removing samples into our framework.

[1]All concepts in X, Y, or A that we consider are assumed to be categorical because the continuous concept is not well-defined in the literature of concept.

2.3. Concept Influence for Fairness (CIF)

Our goal is to quantify the counterfactual effect of changing c for each data sample (x, y, a). Mathematically, denote by $(\hat{x}, \hat{y}, \hat{a})$ the counterfactual sample by intervening on c. Consider a training sample $z_i^{tr} := (x_i, y_i, a_i, c_i)$, and define a counterfactual sample for z_i^{tr} when intervening on $C = c'$ as follows:

$$\hat{x}(c'), \hat{y}(c'), \hat{a}(c') \sim \mathbb{P}\left(\hat{X}, \hat{Y}, \hat{A} | X = x, Y = y, A = a, do(C = c')\right), \; c' \neq c. \tag{4}$$

In the above, $do(\cdot)$ denotes the celebrated do-operation in causal models (Pearl, 2010). The definition is slightly abused – when C overlaps with any of (X, Y, A), the $do(\cdot)$ operation has a higher priority and is assumed to automatically override the other dependencies. For example, when $C = A$, we have:

$$\mathbb{P}\left(\hat{X}, \hat{Y}, \hat{A} | X = x, Y = y, A = a, do(C = c')\right) = \mathbb{P}\left(\hat{X}, \hat{Y}, \hat{A} | X = x, Y = y, \cancel{A = a}, do(A = \hat{a})\right) \tag{5}$$

Denote a counterfactual sample as $\hat{z}_i^{tr}(c') = (\hat{x}_i(c'), \hat{y}_i(c'), \hat{a}_i(c'), \hat{c}_i = c')$. Then we define the counterfactual model when replacing $z_i^{tr} = (x_i, y_i, a_i, c_i)$ with $\hat{z}_i^{tr}(c')$ as:

$$\hat{\theta}_{i,c'} := \text{argmin}_\theta \{R(\theta) - \epsilon \cdot \ell(\theta, z_i^{tr}) + \epsilon \cdot \ell(\theta, \hat{z}_i^{tr}(c'))\} \tag{6}$$

Alternatively, we can identify multiple counterfactual examples $\hat{z}_i^{tr}(c', k), k = 1, 2, ..., E$ and compute the average

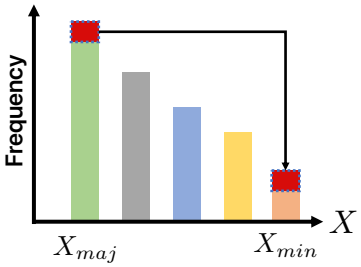


Figure 2. Illustration of the effect of intervening sensitive attribute A as rebalancing data distribution.

effects by defining: $\hat{\theta}_{i,c'} := \text{argmin}_\theta \Big\{ R(\theta) - \epsilon \cdot \ell(\theta, z_i^{tr}) + \epsilon \cdot \frac{\sum_{k=1}^{E} \ell(\theta, \hat{z}_i^{tr}(c', k))}{E} \Big\}$.

Definition 2 (Concept Influence for Fairness (CIF))**.** *The concept influence for fairness (CIF) of intervening on a concept C to c' in sample i on the fairness loss ℓ_{fair} is defined as:*

$$\textit{infl}(D_{val}, \hat{\theta}_{i,c'}) := \ell_{fair}(\hat{\theta}) - \ell_{fair}(\hat{\theta}_{i,c'}) \tag{7}$$

Invoking Proposition 1, we can easily prove:

Proposition 2. *The concept influence for fairness (CIF) of a training sample z_i^{tr} when counterfactually intervened to $\hat{z}_i^{tr}(c')$ based on the target concept c' can be computed as:*

$$\textit{infl}(D_{val}, \hat{\theta}_{i,c'}) \approx -\nabla_\theta \ell_{fair}(\hat{\theta})^\intercal H_{\hat{\theta}}^{-1} \Big(\nabla \ell(z_i^{tr}; \hat{\theta}) - \nabla \ell(\hat{z}_i^{tr}(c'); \hat{\theta}) \Big) \tag{8}$$

2.4. Why Can CIF Improve Fairness?

We provide insights into why intervening training data attributes using CIF framework can improve fairness. For simplicity, we focus on accuracy disparity as the fairness measure. The complete analysis is shown in Appendix C, and we give a brief summary here. We base the analysis on the data generation model adopted in (Feldman, 2020; Liu, 2021) to capture the impact of data patterns generated with different frequencies and the impact of label errors. This setup is a good fit for understanding how counterfactual data interventions can change the data frequency of different groups (majority group with higher frequency vs. minority group with lower frequency) and provides insights for CIF.

Intervening labels Y is relatively straightforward. If we are able to intervene on a training label of a disadvantaged group from a wrong label to a correct one, we can effectively improve the performance of the model for this group. Therefore the label intervention can reduce the accuracy disparities. Our analysis also hints that the influence function is more likely to identify samples from the disadvantaged group with a lower presence in the data and mislabeled samples. This is because, for a minority group, a single label change would incur a relatively larger change in the influence value.

Intervening sensitive attributes A improves fairness by "balancing" the data distribution. Later in the experiments (Figure 9), we show that the influence function often identifies the data from the majority group and recommends them to be intervened to the minority group, as shown in Figure 2. In the analysis, we also show that this intervention incurs positive changes in the accuracy disparities between the two groups and therefore improves fairness.

3. Algorithmic Details

We present our algorithms for generating counterfactual samples and for computing CIF.

3.1. Generating Counterfactual Samples

To compute the fairness influence based on Eqn. 8, we need to first generate the corresponding counterfactual sample $\hat{z}_i^{tr}(c') = (\hat{x}_i(c'), \hat{y}_i(c'), \hat{a}_i(c'), \hat{c}_i = c')$ when intervening concept C to c'. Theoretically, generating the counterfactual examples requires knowing the causal graphs but we use a set of practical algorithms to approximate.

Intervening Label Y. Since there is no variable in training data dependent on Y (Figure 1(c)), we can simply change the sample's label to the target label $\hat{y}_i$ and keep other attributes unchanged, *i.e.* $\hat{z}_i^{tr}(\hat{y}_i) = (x_i, \hat{y}_i, a_i, \hat{c}_i = \hat{y}_i)$.

Intervening Sensitive Attribute A. When we intervene a sample's A, both its X and Y need to change (Figure 1(a)). This is the same as asking, *e.g.* in a loan application, "How a female applicant's profile (*i.e.* x_i) and the loan decision (*i.e.* y_i) would change, had she been a male (*i.e.* $a_i = \hat{a}_i$)?" Inspired by (Black et al., 2020), we train a W-GAN (Arjovsky et al., 2017) with *optimal transport mapping* (Villani et al., 2009) to generate *in-distributional*[2] counterfactual samples for x_i as if x_i belongs to a different a_i. To do so, we need to map the distribution of X from $A = a$ to $A = a'$. We first partition the training samples' feature into two groups: $X|A = a$ and $X|A = a'$. Then we train a W-GAN with the generator $G_{a \to a'}$ as the approximated optimal transport mapping from $X|A = a$ to $X|A = a'$ and the discriminator $D_{a \to a'}$ ensures the mapped samples $G_{a \to a'}(X)$ and the

[2]We need the counterfactual samples to be in-distributional rather than out-of-distributional because we need the change between the counterfactual sample and the original sample to be large enough to impact the fairness measure. We tried counterfactual examples (Wachter et al., 2017) that impose minimum change to the original sample, and it does not work well in mitigation because the fairness influence value they induce is too small. Other approaches like data generation via causal graph only work on synthetic data.

real samples $X|A=a'$ are indistinguishable. The training objectives are the following:

$$\begin{aligned}\ell_{G_{a\to a'}} &= \frac{1}{n}\Big(\sum_{x\in X|A=a} D(G(x)) + \lambda\cdot\sum_{x\in X|A=a} c(x,G(x))\Big)\\ \ell_{D_{a\to a'}} &= \frac{1}{n}\Big(\sum_{x'\in X|A=a'} D(x') - \sum_{x\in X|A=a} D(G(x))\Big)\end{aligned} \tag{9}$$

where n is the number of training samples, λ is the weight balancing the conventional W-GAN generator loss (*i.e.* the first term in $\ell_{G_{a\to a'}}$) and the distance cost function $c(.)$ (*i.e.* ℓ_2 norm in our case) that makes sure the mapped samples are not too far from the original distribution.

After we train the W-GAN on the training data, we can use the trained generator $G_{a\to a'}$ to map a sample x_i to its counterfactual version $\hat{x}_i = G_{a_i\to\hat{a}_i}(x_i)$. In addition, once we have the counterfactual features, we can use the original model to predict the corresponding counterfactual label (*i.e.* following the causal link $X\to Y$ in Figure 1). The resulting counterfactual sample is $\hat{z}_i^{tr}(\hat{a}_i) = (\hat{x}_i, h_{\hat{\theta}}(\hat{x}_i), \hat{a}_i, \hat{c}_i = \hat{a}_i)$.

Intervening Feature X. In image data, assume there exists an image-label attribute $C = attr(X)$, *e.g.* young or old in facial images, and intervening X means transforming the image (*i.e.* all pixel values in X) as if it belongs to a different C. In tabular data, C is one of the features in X, and when C is changed, all other features in X need to change accordingly. In both cases, similar to intervening A, we train a W-GAN to learn the mapping from the group $X|C=c$ to $X|C=c'$; the resulting generator is $G_{c\to c'}$ and the generated counterfactual feature is $\hat{x}_i = G_{c_i\to\hat{c}_i}(x_i)$. Similarly, since causal path $X\to Y$ exists in Figure 1(b), we also use the original model's predicted label as the counterfactual label. The resulting counterfactual sample is $\hat{z}_i^{tr}(\hat{c}_i) = (\hat{x}_i, h_{\hat{\theta}}(\hat{x}_i), a_i, \hat{c}_i = \hat{x}_i)$.

Removal. Removing is simply setting the counterfactual sample to be null, *i.e.* $\hat{z}_i^{tr}(c') = \varnothing$.

3.2. Computing Influence

Following (Koh & Liang, 2017), we use the Hessian vector product (HVP) to compute the product of the second and the third term in Eqn. 8 together. Let $v := \Big(\nabla\ell(z_i^{tr};\hat{\theta}) - \nabla\ell(\hat{z}_i^{tr}(c');\hat{\theta})\Big)$, we can compute $H^{-1}v$ recursively (Agarwal et al., 2017):

$$\hat{H}_r^{-1}v = v + (I - \hat{H}_0)\hat{H}_{r-1}^{-1}v \tag{10}$$

where $\hat{H}_0$ is the Hessian matrix approximated on random batches. Let t be the final recursive iteration, then the final CIF is $\text{infl}(D_{val}, \hat{\theta}_{i,c'}) \approx -\nabla_\theta\ell_{\text{fair}}(\hat{\theta})^\intercal\hat{H}_t^{-1}v$, where $\ell_{\text{fair}}(\hat{\theta})$ is the surrogate loss of fairness measure (*e.g.* Eqn. 3, 17 or 21).

4. Experiments

We present a series of experiments to validate the effectiveness of CIF in explaining and mitigating model unfairness, detecting biased/poisoned samples, and recommending resampling to balance representation.

4.1. Setup

We test CIF on 4 datasets: synthetic, COMPAS (Angwin et al., 2016), Adult (Kohavi et al., 1996), and CelebA (Liu et al., 2015). We report results on three group fairness metrics (DP, EOP, and EO, see Table 1 in Appendix B for the definition). The detailed settings are the following:

- **Synthetic**: We generate synthetic data with the assumed causal graphs in Figure 1, and therefore we have the ground-truth counterfactual samples. See Appendix D.1 for the dataset generation process. Model: logistic regression.
- **COMPAS**: Recidivism prediction data (we use the preprocessed tabular data from IBM's AIF360 toolkit (Bellamy et al., 2019)). Feature X: tabular data. Label Y: recidivism within two years (binary). Sensitive attribute A (removed from feature X): race (white or non-white). Model: logistic regression. When intervening X, we choose to flip the binary feature (age >45 or not) in X.
- **Adult**: Income prediction data (we use the preprocessed tabular data from IBM's AIF360 toolkit (Bellamy et al., 2019)). Feature X: tabular data. Label Y: if income $>50K$ or not. Sensitive attribute A (removed from feature X): sex (male or female). Model: logistic regression. When intervening X, we choose to flip the binary feature race (white or non-white) in X.
- **CelebA**: Facial image dataset. Feature X: facial images. Label Y: attractive or not (binary). Sensitive attribute A: gender (male and female). Model: ResNet18 (He et al., 2016). When intervening X, we choose to flip the binary image-level label "Young."

4.2. Mitigation Performance

We test the CIF-based mitigation by first computing CIF values on all training samples, and then replacing samples with the highest CIF values by their corresponding generated counterfactual samples, and retraining the model. Figure 3-5 show the fairness performance after the model training. We observe that all three fairness measures improve significantly after following CIF's mitigation recommendations. See Figure 10-12 in Appendix D.3 for the reported model accuracy.

We summarize observations: (1) Intervening on Y proves to be highly effective on real-world data but not on synthetic. We conjecture that this is because we control the

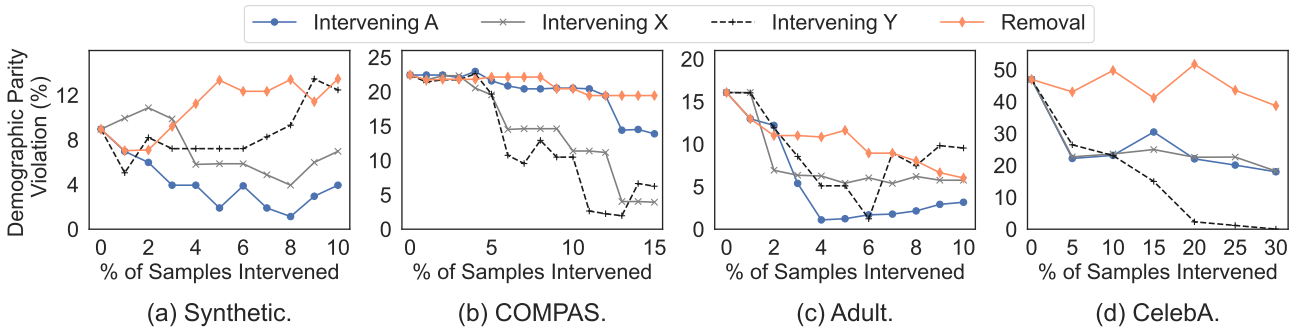


Figure 3. CIF-based mitigation performance with fairness measure Demographic Parity (DP).

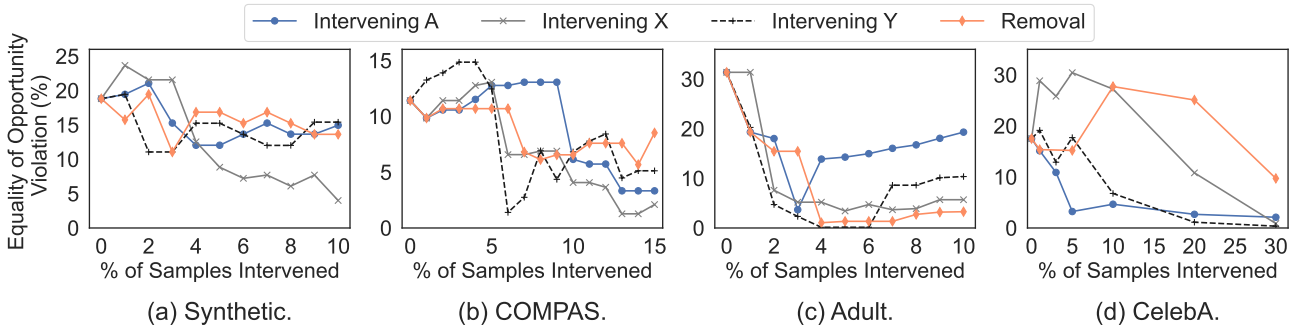


Figure 4. CIF-based mitigation performance with fairness measure Equality of Opportunity (EOP).

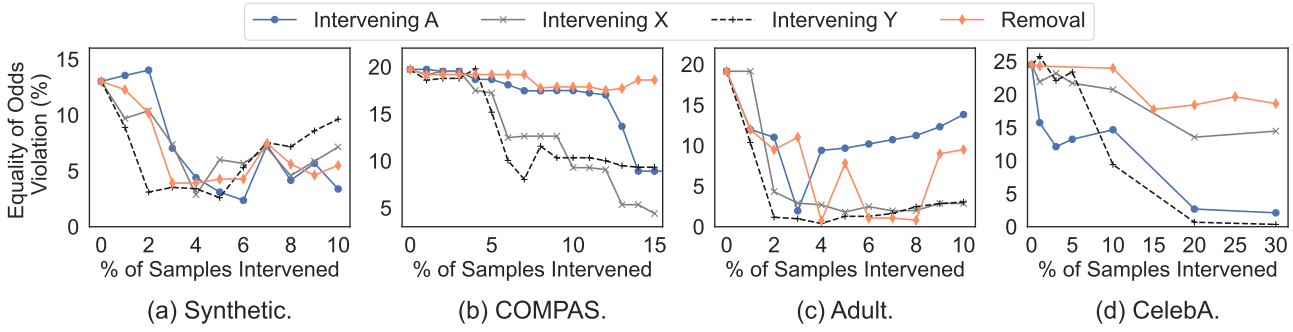


Figure 5. CIF-based mitigation performance with fairness measure Equality of Odds (EO).

synthetic data to be cleanly labeled, which is not the case for other real-world data. Later we confirm this observation by showing the effectiveness of our approach in detecting noisy labels. (2) Intervening on A proves to be helpful for most cases, especially for DP, which highly relates to the demographic variable A. (3) We set the size of synthetic data to be small (1,000) to show that simply removing training samples might not always be a good strategy, particularly on a small dataset which the model would suffer significantly from losing training samples.

Fairness-utility Tradeoff. We report the fairness-utility tradeoffs of our mitigation on COMPAS, together with the in-processing mitigation (Agarwal et al., 2018) in Figure 6. Our mitigation is comparable to (Agarwal et al., 2018); sometimes we can achieve better fairness given a similar level of accuracy (*e.g.* when accuracy is $\sim 60\%$).

Distribution of CIF Value. We show the distribution of influence values computed on COMPAS corresponding to three fairness metrics in Appendix D.5, Figure 15. Intervening Y has the highest influence value compared to other types of intervention. This is because we change the value of Y directly in this operation, which is more "unnaturally" compared to generating more "natural" counterfactual examples with W-GAN (intervening X and A) or model-predicted value of Y (intervening X). The implication is mislabelling would have a larger impact on fairness than corrupted features or incorrect group membership, which is consistent with our theoretical analysis in Section 2.4. Therefore practitioners should be particularly cautious about mislabelling that can hurt fairness significantly, *e.g.* if any

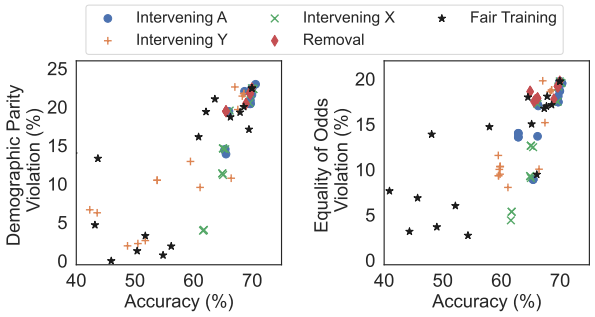


Figure 6. Fairness-accuracy tradeoff of CIF-based mitigation on COMPAS. CIF-based mitigation is comparable to in-processing mitigation method, and sometimes achieves better fairness given a similar level of accuracy.

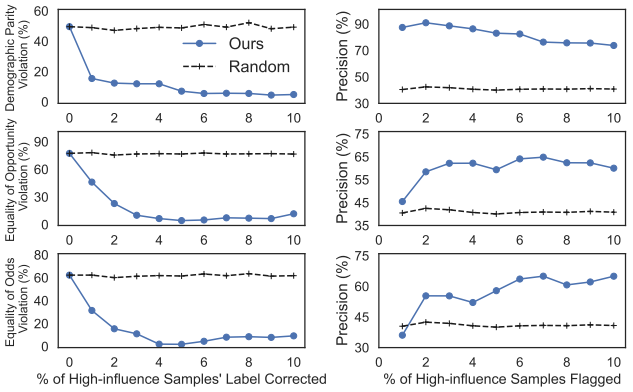


Figure 7. Precision and mitigation performance of using Y-intervened CIF to detect and correct training mislabeling that causes bias on Adult.

unprivileged group should be labeled favorable but ended up getting labeled unfavorable.

4.3. Additional Applications of CIF

We provide three examples of additional applications that can be derived from our CIF framework.

Fixing Mislabeling. We flip training labels Y in the Adult dataset to artificially increase the model's unfairness. Following (Wang et al., 2021), we add group-dependent label noise, *i.e.* the probability of flipping a sample's Y is based on its A, to enlarge the fairness gap. See Appendix D.6 for the experimental details. We then compute Y-intervened CIF on each sample, and flag samples with the top CIF value. In Figure 7, we report the precision of our CIF-based detection and mitigation performance if we flip the detected samples' labels and retrain the model. Our detection can flag the incorrect labels that are known to be the source of the unfairness with high precision (compared to randomly flagging the same percentage) and improves the model fairness if the detected labels are corrected.

Defending against Poisoning Attacks. We demonstrate another application of defending models against fairness poisoning attacks. To generate poisoned training samples that cause the model's unfairness, we choose poisoned training samples with the same probability based on the group- and label-dependent probability in the previous application. In addition to flipping the samples' labels, we also set the target feature (*i.e.* race in Adult) to be a fixed value (*i.e.* white) regardless of the original feature value. The attack that modifies a sample's feature to be a fixed value and changes its label is known as backdoor attack (Gu et al., 2019; Li et al., 2021b; Wu et al., 2022a), a special type of poisoning attack. After the poisoning, all fairness measures become worse (see Appendix D.6 for more details). For detection, we compute X-intervened CIF on the poisoned feature, and flag samples with high CIF value. For mitigation, if we flag a sample to be poisoned, we remove it from the training set and retrain the model. Figure 8 shows the precision of our detection and the mitigation performance after removal. We observe a high precision and reasonably good fairness improvement.

Resampling Imbalanced Representations. To create an extremely imbalanced representation in the training set, in Adult we upsample the positive samples in the privileged group (*i.e.* male) by 200%, further increasing the percentage of positive samples that belong to the privileged group, and therefore the training samples are overwhelmingly represented by the privileged group. The resulting fairness becomes worse (see Appendix D.6 for more details). We then compute A-intervened CIF, and replace the high-influence samples with their counterfactual samples (*i.e.* adding counterfactual samples in the unprivileged group and reducing samples from the privileged group). In Figure 9, we report the percentage of high-influence samples that belong to the privileged group (*i.e.* how much CIF recommends the data balancing) and the mitigation performance. The high-influence samples are almost all from the privileged group, which is expected, and if they were converted to the counterfactual samples as if they are from the unprivileged group, *i.e.* recollecting and resampling the training distribution, then fairness can improve.

5. Related Work

Influence Function. The goal of influence function is to quantify the impact of training data on the model's output. (Koh & Liang, 2017) popularizes the idea of training data

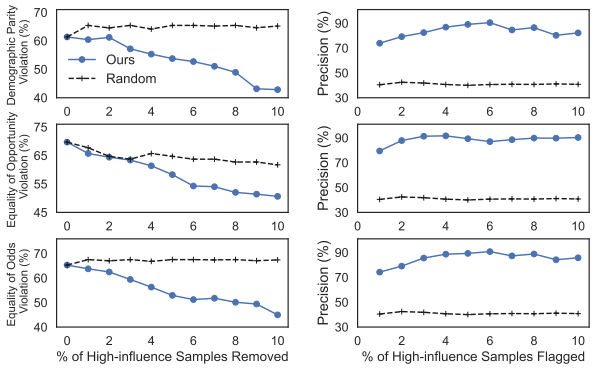


Figure 8. Precision and mitigation performance of using X-intervened CIF to detect and correct poisoned training samples that cause unfairness on Adult.

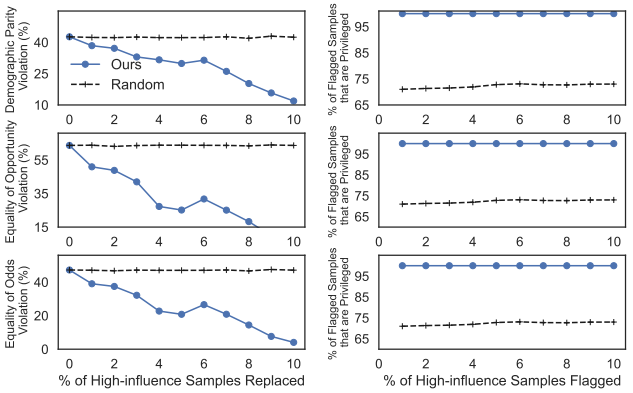


Figure 9. Performance of using A-intervened CIF to detect and correct imbalanced training representation that causes unfairness on Adult.

influence to the attention of our research community and has demonstrated its power in a variety of applications. Later works have aimed to improve the efficiency of computing influence functions. For example, Tracein (Pruthi et al., 2020) proposes a first-order solution that leverages the training gradients of the samples, and a neural tangent kernel approach for speeding up this task. Other works have explored the computation of group influence (Basu et al., 2020b), the robustness of influence function (Basu et al., 2020a), its application in explainable AI (Linardatos et al., 2020) and other tasks like graph networks (Chen et al., 2023).

Influence Function for Fairness. Our work is closely relevant to the recent discussions on quantifying training data's influence on a model's fairness properties. (Wang et al., 2022a) computes the training data influence to fairness when removing a certain set of training samples. (Li & Liu, 2022) discusses a soft version of the removal and computes also the optimal "removal weights" for each sample to improve fairness. And (Sattigeri et al., 2022) leverages the computed influence to perform a post-hoc model update to improve its fairness. Note that those works consider the fairness effect of removing or reweighing training samples. Our work targets a more flexible and powerful definition of influence that can give practitioners a wider scope of understanding by introducing the idea of concepts and generating counterfactual samples as well as result in a wider range of potential applications.

Data Repairing for Fairness. Our work is also related to the work on data repairing to improve fairness. (Krasanakis et al., 2018; Lahoti et al., 2020) discuss the possibilities of reweighing training data to improve fairness. (Zhang et al., 2022) proposes a "reprogramming" framework that modified the features of training data. (Liu & Wang, 2021) explores the possibility of resampling labels to improve the fairness of training. Other works study the robustness of model w.r.t fairness (Wang et al., 2022b; Chhabra et al., 2023; Li et al., 2022). Another line of research that repairs training data is through training data pre-proccessing (Calmon et al., 2017; Celis et al., 2020; Kamiran & Calders, 2012; du Pin Calmon et al., 2018), synthetic fair data (Sattigeri et al., 2019; Jang et al., 2021; Xu et al., 2018; van Breugel et al., 2021), and data augmentation (Sharma et al., 2020; Chuang & Mroueh, 2021).

6. Conclusions and Limitations

We propose *Concept Influence for Fairness* (CIF), which generalizes the definition of influence function for fairness from focusing only on the effects of removing or reweighing the training samples to a broader range of dimensions related to the training data's properties. The main idea is to consider the effects of intervening on a certain *concept* of training data, which is a more flexible framework to help practitioners better understand unfairness with a wider scope and leads to more potential downstream applications.

We point out two limitations: (1) CIF needs to generate counterfactual samples w.r.t different concepts, which can be computationally expensive and (2) in CIF-based mitigation, it can be non-trivial to determine the optimal number of training samples to intervene that would maximally improve fairness.

References

What AI fairness in practice looks like at Facebook . https://ai.facebook.com/blog/what-ai-fairness-in-practice-looks-like-at-facebook/, 2021.

Adel, T., Valera, I., Ghahramani, Z., and Weller, A. One-network adversarial fairness. In *Proc. of AAAI*, 2019.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *Proc. of ICML*, 2018.

Agarwal, N., Bullins, B., and Hazan, E. Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1): 4148–4187, 2017.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. In *Ethics of Data and Analytics*, pp. 254–264. Auerbach Publications, 2016.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proc. of ICML*, 2017.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

Basu, S., Pope, P., and Feizi, S. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020a.

Basu, S., You, X., and Feizi, S. On second-order group influence functions for black-box predictions. In *International Conference on Machine Learning*, pp. 715–724. PMLR, 2020b.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Black, E., Yeom, S., and Fredrikson, M. Fliptest: fairness testing via optimal transport. In *Proc. of FAccT*, 2020.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Proc. of NeurIPS*, 2017.

Celis, L. E., Keswani, V., and Vishnoi, N. Data preprocessing to mitigate bias: A maximum entropy based approach. In *Proc. of ICML*, 2020.

Celis, L. E., Mehrotra, A., and Vishnoi, N. Fair classification with adversarial perturbations. In *Proc. of NeurIPS*, 2021.

Chai, J. and Wang, X. Fairness with adaptive weights. In *Proc. of ICML*, 2022.

Chen, Y., Raab, R., Wang, J., and Liu, Y. Fairness transferability subject to bounded distribution shift. *arXiv preprint arXiv:2206.00129*, 2022.

Chen, Z., Li, P., Liu, H., and Hong, P. Characterizing the influence of graph elements. In *Proc. of ICLR*, 2023.

Chhabra, A., Li, P., Mohapatra, P., and Liu, H. Robust fair clustering: A novel fairness attack and defense framework. In *Proc. of ICLR*, 2023.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Chuang, C.-Y. and Mroueh, Y. Fair mixup: Fairness via interpolation. In *Proc. of ICLR*, 2021.

Cook, R. D. and Weisberg, S. *Residuals and influence in regression*. New York: Chapman and Hall, 1982.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

du Pin Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1106–1119, 2018.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *Proc. of STOC*, 2020.

Fogliato, R., Chouldechova, A., and G'Sell, M. Fairness evaluation in presence of biased noisy labels. In *Proc. of AISTATS*, 2020.

Giguere, S., Metevier, B., Brun, Y., Castro da Silva, B., Thomas, P., and Niekum, S. Fairness guarantees under demographic shift. In *Proc. of ICLR*, 2022.

Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proc. of CVPR*, 2016.

Jang, T., Zheng, F., and Wang, X. Constructing a fair classifier with generated fair data. In *Proc. of AAAI*, 2021.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proc. of ICML*, 2017.

Koh, P. W. W., Ang, K.-S., Teo, H., and Liang, P. S. On the accuracy of influence functions for measuring group effects. In *Proc. of NeurIPS*, 2019.

Kohavi, R. et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proc. of KDD*, volume 96, pp. 202–207, 1996.

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., and Kompatsiaris, Y. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pp. 853–862, 2018.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Li, P. and Liu, H. Achieving fairness at no utility cost via data reweighing with influence. In *Proc. of ICML*, 2022.

Li, P., Zhao, H., and Liu, H. Deep fair clustering for visual learning. In *Proc. of CVPR*, 2020.

Li, P., Wang, Y., Zhao, H., Hong, P., and Liu, H. On dyadic fairness: Exploring and mitigating bias in graph connections. In *Proc. of ICLR*, 2021a.

Li, P., Xia, E., and Liu, H. Learning antidote data to individual unfairness. *arXiv preprint arXiv:2211.15897*, 2022.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. In *Proc. of NeurIPS*, 2021b.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Liu, Y. Understanding instance-level label noise: Disparate impacts and treatments. In *Proc. of ICML*, 2021.

Liu, Y. and Wang, J. Can less be more? when increasing-to-balancing label noise rates considered beneficial. In *Proc. of NeurIPS*, 2021.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proc. of ICCV*, December 2015.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *Proc. of ICML*, 2018.

Pearl, J. Causal inference. *Causality: objectives and assessment*, pp. 39–58, 2010.

Pruthi, G., Liu, F., Kale, S., and Sundararajan, M. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

Rezaei, A., Liu, A., Memarrast, O., and Ziebart, B. D. Robust fairness under covariate shift. In *Proc. of AAAI*, 2021.

Rousseeuw, P. J., Hampel, F. R., Ronchetti, E. M., and Stahel, W. A. *Robust statistics: the approach based on influence functions*. John Wiley & Sons, 2011.

Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., and Varshney, K. R. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

Sattigeri, P., Ghosh, S., Padhi, I., Dognin, P., and Varshney, K. R. Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting. In *Proc. of NeurIPS*, 2022.

Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., and Varshney, K. R. Data augmentation for discrimination prevention and bias disambiguation. In *Proc. of AIES*, 2020.

Song, H., Li, P., and Liu, H. Deep clustering based fair outlier detection. In *Proc. of KDD*, 2021.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. Learning controllable fair representations. In *Proc. of AISTATS*, 2019.

van Breugel, B., Kyono, T., Berrevoets, J., and van der Schaar, M. Decaf: Generating fair synthetic data using causally-aware generative networks. In *Proc. of NeurIPS*, 2021.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Wang, J., Liu, Y., and Levy, C. Fair classification with group-dependent label noise. In *Proc. of FAccT*, 2021.

Wang, J., Wang, X. E., and Liu, Y. Understanding instance-level impact of fairness constraints. In *Proc. of ICML*, 2022a.

Wang, Z., Dong, X., Xue, H., Zhang, Z., Chiu, W., Wei, T., and Ren, K. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proc. of CVPR*, 2022b.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pp. 1920–1953. PMLR, 2017.

Wu, B., Chen, H., Zhang, M., Zhu, Z., Wei, S., Yuan, D., and Shen, C. Backdoorbench: A comprehensive benchmark of backdoor learning. In *Proc. of NeurIPS*, 2022a.

Wu, S., Gong, M., Han, B., Liu, Y., and Liu, T. Fair classification with instance-dependent label noise. In *Conference on Causal Learning and Reasoning*, pp. 927–943. PMLR, 2022b.

Xing, X., Liu, H., Chen, C., and Li, J. Fairness-aware unsupervised feature selection. In *Proc. of CIKM*, 2021.

Xu, D., Yuan, S., Zhang, L., and Wu, X. Fairgan: Fairness-aware generative adversarial networks. In *Proc. of IEEE Big Data*, 2018.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proc. of ICML*, 2013.

Zhang, G., Zhang, Y., Zhang, Y., Fan, W., Li, Q., Liu, S., and Chang, S. Fairness reprogramming. In *Proc. of NeurIPS*, 2022.

Zhu, Z., Luo, T., and Liu, Y. The rich get richer: Disparate impact of semi-supervised learning. In *Proc. of ICLR*, 2022.

This Appendix is organized as follows:

- Section A includes the derivation of the influence function for fairness.
- Section B shows how we use the surrogate loss to approximate fairness measure.
- Section C includes the full results of our theoretical analysis on why CIF can help mitigation, including all the proofs and theorems.
- Section D includes additional details about the experiments as well as additional results.

A. Derivation of Fairness Function on Group Fairness

Assume the risk of θ is $R(\theta) := \frac{1}{n}\sum_{i=1}^{n}\ell(z_i^{tr};\theta)$, and the model trained on the entire training set is $\hat{\theta} := \text{argmin}_\theta R(\theta)$. The resulting model weights if we assign weight $w_i \in [0,1]$ to each sample $i \in \mathcal{K}$ and then upweight them by some small ϵ is the following:

$$\hat{\theta}_{\mathcal{K}} := \text{argmin}_\theta\{R(\theta) + \epsilon\sum_{i\in\mathcal{K}} w_i \cdot \ell(z_i^{tr};\theta)\}, \tag{11}$$

By the first order condition of $\hat{\theta}_{\mathcal{K}}$ we have

$$0 = \nabla R(\hat{\theta}_{\mathcal{K}}) + \epsilon\sum_{i\in\mathcal{K}} w_i \cdot \nabla\ell(z_i^{tr};\hat{\theta}_{\mathcal{K}})$$

When $\epsilon \to 0$, with the Taylor expansion (and first-order approximation) we have:

$$0 \approx \left(\nabla R(\hat{\theta}) + \epsilon\sum_{i\in\mathcal{K}} w_i \cdot \nabla\ell(z_i^{tr};\hat{\theta})\right) + \left(\nabla^2 R(\hat{\theta}) + \epsilon\sum_{i\in\mathcal{K}} w_i \cdot \nabla^2\ell(z_i^{tr};\hat{\theta})\right)\cdot(\hat{\theta}_{\mathcal{K}} - \hat{\theta})$$

By the first-order condition of $\hat{\theta}$ we have $\nabla R(\hat{\theta}) = 0$, and re-arranging terms we have

$$\frac{\hat{\theta}_{\mathcal{K}} - \hat{\theta}}{\epsilon} = -\left(H_{\hat{\theta}} + \epsilon\cdot\sum_{i\in\mathcal{K}} w_i \cdot \nabla^2\ell(z_i^{tr};\hat{\theta})\right)^{-1}\cdot\left(\sum_{i\in\mathcal{K}} w_i \cdot \nabla\ell(z_i^{tr};\hat{\theta})\right)$$

Taking the limit of $\epsilon \to 0$ on both sides we have

$$\frac{\partial\hat{\theta}_{\mathcal{K}}}{\partial\epsilon}\Big|_{\epsilon=0} = -H_{\hat{\theta}}^{-1}\cdot\left(\sum_{i\in\mathcal{K}} w_i \cdot \nabla\ell(z_i^{tr};\hat{\theta})\right)$$

Finally, the fairness influence of assigning training sample i in group $\mathcal{K}$ with weight w_i is:

$$\begin{aligned}
\text{infl}(D_{val},\mathcal{K},\hat{\theta}) &:= \ell_{\text{fair}}(\hat{\theta}) - \ell_{\text{fair}}(\hat{\theta}_{\mathcal{K}}) && (12)\\
&\approx \frac{\partial\ell_{\text{fair}}(\hat{\theta}_{\mathcal{K}})}{\partial\epsilon}\Big|_{\epsilon=0} && (13)\\
&= \nabla_\theta\ell_{\text{fair}}(\hat{\theta})^\intercal\frac{\hat{\theta}_{\mathcal{K}}}{\partial\epsilon}\Big|_{\epsilon=0} && (14)\\
&= -\nabla_\theta\ell_{\text{fair}}(\hat{\theta})^\intercal H_{\hat{\theta}}^{-1}\left(\sum_{i\in\mathcal{K}} w_i\nabla\ell(z_i^{tr};\hat{\theta})\right) && (15)
\end{aligned}$$

B. Approximating Fairness Metrics

Similarly to DP, we can approximate the violation of Equality of Opportunity (EOP) with:

$$\begin{aligned}
\ell_{EOP}(\hat{\theta}) &:= \left|\mathbb{P}(h_\theta(X)=1|A=0,Y=1) - \mathbb{P}(h_\theta(X)=1|A=1,Y=1)\right| && (16)\\
&\approx \left|\frac{\sum_{i\in D_{val}:a_i=0,y_i=1} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=0,y_i=1]} - \frac{\sum_{i\in D_{val}:a_i=1,y_i=1} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=1,y_i=1]}\right| && (17)
\end{aligned}$$

And for Equality of Odds (EO), we have

$$\begin{aligned}
\ell_{EO}(\hat{\theta}) &:= \frac{1}{2}\Big(\left|\mathbb{P}(h_\theta(X)=1|A=0,Y=1) - \mathbb{P}(h_\theta(X)=1|A=1,Y=1)\right| + && (18)\\
&\qquad \left|\mathbb{P}(h_\theta(X)=1|A=0,Y=0) - \mathbb{P}(h_\theta(X)=1|A=1,Y=0)\right|\Big) && (19)\\
&\approx \frac{1}{2}\Big(\left|\frac{\sum_{i\in D_{val}:a_i=0,y_i=1} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=0,y_i=1]} - \frac{\sum_{i\in D_{val}:a_i=1,y_i=1} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=1,y_i=1]}\right| + && (20)\\
&\qquad \left|\frac{\sum_{i\in D_{val}:a_i=0,y_i=0} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=0,y_i=0]} - \frac{\sum_{i\in D_{val}:a_i=1,y_i=0} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=1,y_i=0]}\right|\Big) && (21)
\end{aligned}$$

We summarize the definition and surrogate approximation of three group fairness measures as follows:

Fairness Measure	Definition	Surrogate Approximation
Demographic Parity (DP)	$\left\|\mathbb{P}(h_\theta(X)=1\|A=0)-\mathbb{P}(h_\theta(X)=1\|A=1)\right\|$	$\left\|\frac{\sum_{i\in D_{val}:a_i=0} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=0]}-\frac{\sum_{i\in D_{val}:a_i=1} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=1]}\right\|$
Equality of Opportunity (EOP)	$\left\|\mathbb{P}(h_\theta(X)=1\|A=0,Y=1)-\mathbb{P}(h_\theta(X)=1\|A=1,Y=1)\right\|$	$\left\|\frac{\sum_{i\in D_{val}:a_i=0,y_i=1} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=0,y_i=1]}-\frac{\sum_{i\in D_{val}:a_i=1,y_i=1} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=1,y_i=1]}\right\|$
Equality of Odds (EO)	$\frac{1}{2}\left(\begin{array}{l}\left\|\mathbb{P}(h_\theta(X)=1\|A=0,Y=1)-\mathbb{P}(h_\theta(X)=1\|A=1,Y=1)\right\|+\\ \left\|\mathbb{P}(h_\theta(X)=1\|A=0,Y=0)-\mathbb{P}(h_\theta(X)=1\|A=1,Y=0)\right\|\end{array}\right)$	$\frac{1}{2}\left(\begin{array}{l}\left\|\frac{\sum_{i\in D_{val}:a_i=0,y_i=1} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=0,y_i=1]}-\frac{\sum_{i\in D_{val}:a_i=1,y_i=1} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=1,y_i=1]}\right\|+\\ \left\|\frac{\sum_{i\in D_{val}:a_i=0,y_i=0} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=0,y_i=0]}-\frac{\sum_{i\in D_{val}:a_i=1,y_i=0} g(z_i^{val};\theta)}{\sum_{i\in D_{val}}\mathbb{I}[a_i=1,y_i=0]}\right\|\end{array}\right)$

Table 1. Fairness definition and surrogate approximation.

C. Theoretical Analysis: Why Can CIF Improve Fairness?

We base our analysis on the data generation model adopted in (Feldman, 2020; Liu, 2021) to capture the impact of data patterns generated with different frequencies and the impact of label errors. This setup is a good fit for understanding how counterfactual data interventions change the frequencies of data of different groups and therefore provides insights for CIF.

In this setup, each feature X takes value from a *discretized* set $\mathcal{X}$. For each $X \in \mathcal{X}$, sample a quantity q_X independently and uniformly from a set $\lambda := \{\lambda_1, ..., \lambda_N\}$. The probability of observing an X is given by $D(X) = q_X/(\sum_{X\in\mathcal{X}} q_X)$. Each X is mapped to a true label $Y = f(X)$. But our observed training labels can be noisy, denoting as $\tilde{Y} \sim \mathbb{P}(\tilde{Y}|X, Y)$. n pairs of $(X, \tilde{Y})$ are observed and collected for the dataset. Denote by S_l the set of all samples that appear l times in the dataset, and denote by $l[X]$ the number of appearances for X. Each X is also associated with a sensitive group attribute A. Denote by h_θ as the classification model defined by $\theta \in \Theta$ (parametric space) and the generalization error over a given distribution $\mathcal{D}$ as

$$\mathsf{err}_{\mathcal{D}}(h_\theta) := \mathbb{E}_{\mathcal{D}}[\mathbf{1}(h_\theta(X) \neq Y)] \,.$$

The following expected generalization error is defined in (Feldman, 2020):

$$\mathsf{err}(\theta|D) := \mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathsf{err}_{\mathcal{D}}(h_\theta)\right] \,,$$

where $\mathbb{P}[\cdot|D]$ is the distribution for the data distribution inferred from the dataset D. It is proved that:

Theorem 1 ((Feldman, 2020)). $\mathsf{err}(\theta|D) \geq \min_{\theta'\in\Theta} \mathsf{err}(\theta'|D) + \sum_{l\in[n]} \tau_l \cdot \sum_{X\in S_l} \mathbb{P}[h_\theta(X) \neq Y]$.

In the above τ_l is a constant that depends on l. We call this the *importance* of an l-appearance sample. It is proven in (Feldman, 2020) that when l is small, for instance $l = 1$, τ_l is at the order of $O(\frac{1}{n})$, and when l is large τ_l is at the order of $O(\frac{l^2}{n^2})$ (Liu, 2021).

Consider an ideal setting where we train a parametric model θ that fully memorizes the training data that $R(\theta) = 0$, and therefore $\mathbb{P}[h_\theta(X) \neq Y] = \tilde{\mathbb{P}}[\tilde{Y} \neq Y|X]$, where $\tilde{\mathbb{P}}[\tilde{Y} \neq Y|X]$ is the empirical label distribution for sample pattern X. Theorem 1 can easily generalize to each group D_a:

Proposition 3. $\mathsf{err}(\theta|D_a) \geq \min_{\theta'\in\Theta} \mathsf{err}(\theta'|D_a) + \sum_{l\in[n]} \frac{\tau_l}{\sum_{X\in D_a} \tau_{l[X]}} \cdot \sum_{X\in D_a\cap S_l} \tilde{\mathbb{P}}[\tilde{Y} \neq Y|X]$.

Denote the following *excessive generalization error* for group a:

$$\mathsf{err}_a^+(\theta|D) := \sum_{l\in[n]} \frac{\tau_l}{\sum_{X\in D_a} \tau_{l[X]}} \cdot \sum_{X\in D_a\cap S_l} \tilde{\mathbb{P}}[\tilde{Y} \neq Y|X].$$

Importantly, the above error term captures the vital quantities that are interesting to our problem: (1) the relevant frequency $\frac{\tau_l}{\sum_{X\in D_a} \tau_{l[X]}}$ captures the importance of the pattern with different frequencies and (2) $\tilde{\mathbb{P}}[\tilde{Y} \neq Y|X]$ the label noise rate of sample pattern X.

To set up the discussion, suppose we have two groups a, a'. a is the advantaged group with a smaller $\mathsf{err}_a^+(\pi, \theta|D)$; there is an $X_a \in D_a$ with a larger l_a. On the other hand, there is an $X_{a'} \in D_{a'}$, an $l_{a'}$-appearance sample. We further assume

that $l_a > l_{a'}$ ($X_{a'}$ has a lower representation). The rest of the discussion will focus on the following generalization error disparity as the fairness metric:

$$F(\theta) := |\mathsf{err}_a^+(\theta|D) - \mathsf{err}_{a'}^+(\theta|D)| \,.$$

The excessive generalization error for each group can be viewed as the expected influence of a model θ on the test data for that particular group. So the rest of the analysis focuses on the impact of flipping a sample's label to the group's excessive generalization error and then $F(\theta)$.

Intervening Labels (Y). On the high level, intervening a wrong label from the disadvantaged group a' to the correct one will effectively reduce $\tilde{\mathbb{P}}[\tilde{Y} \neq Y|X]$ for some $X \in D_{a'}$, and therefore reduces the gap from it to the advantaged groups. The literature on influence functions (Koh & Liang, 2017) has demonstrated its power to detect mislabelled samples. But why would the influence function identify samples from the disadvantaged group and samples with wrong labels?

Consider a specific sample $X_{a'} \in D_{a'}$, and suppose its label is wrong. Intervening the wrong label to the correct label for this rare sample leads to a reduction in noise rate $\tilde{\mathbb{P}}[\tilde{Y} \neq Y|X]$ for $X_{a'}$. Therefore we know that intervening on this "rare sample" reduces $\mathsf{err}_{a'}^+(\theta|D)$, and the disparity $F(\theta)$. On the other hand, intervening the label for X_a from the privileged group reduces $\mathsf{err}_a^+(\theta|D)$ but this would further increase the gap $F(\theta)$. Therefore, flipping (*i.e.* intervening on) the wrong labels from the disadvantaged group leads to a larger drop in disparity.

Intervening Sensitive Attributes (A). Suppose $X_a \in D_a$ (from the privileged group) is identified to be intervened. After the counterfactual intervention, X_a is intervened to $X_{a'}$ (from the disadvantaged group), we show the gap in the excessive generalization errors between a and a' is reduced as follows:

(1) Increase in generalization error for the privileged group: For group a's generalization error, since we are removing one sample from it, the *importance* of X_a drops from τ_{l_a} to τ_{l_a-1} as τ_l monotonically increases w.r.t l (recall τ_l implies the importance of a l-frequency sample, the higher l is the more important it generally is). When X_a is a cleaner example that $\tilde{\mathbb{P}}[\tilde{Y} \neq Y|X_a]$ is sufficiently small, especially smaller than the average noise rate $\tilde{\mathbb{P}}[\tilde{Y} \neq Y|X \in D_a]$ of the group a, removing one sample of it results in an increase in the average generalization error (Proposition 4).

(2) Decrease in generalization error for the privileged group: For group a', because of the addition, the weight of $X_{a'}$ increases by $\tau_{l_{a'}+1} - \tau_{l_{a'}}$. Therefore, adding a cleaner sample to group a' not only reduces $X_{a'}$'s empirical label noise rate $\tilde{\mathbb{P}}[\tilde{Y} \neq Y|X_a]$, but also increases the relative weight of $\tau_{l_{a'}}$. Again using Proposition 4, we know that increasing the weight of a smaller quantity will then reduce the average of the group.

To summarize the above, intervening A effectively (1) increases $\mathsf{err}_a^+(\theta|D)$ (*i.e.* increasing the excessive generalization error for the privileged group) and (2) decreases $\mathsf{err}_{a'}^+(\theta|D)$ (*i.e.* decreasing the excessive generalization error for the disadvantaged group). Therefore the counterfactual intervention on A reduces the gaps in the excessive generalization errors between the two groups.

C.1. Proof of Proposition 3

Recall we assume a simplified case where we train a parametric model θ that fully memorizes the training data that $R(\theta) = 0$, and therefore $\mathbb{P}[h_\theta(X) \neq Y] = \tilde{\mathbb{P}}[\tilde{Y} \neq Y|X]$. Following the proof from (Feldman, 2020), it is easy to show that

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(h_\theta(X) \neq Y, X \in D_a)\right] \geq \min_{\theta'\in\Theta} \mathsf{err}(\theta', X \in D_a) + \sum_{l\in[n]} \tau_l \cdot \sum_{X\in D_a\cap S_l} \tilde{\mathbb{P}}[\tilde{Y} \neq Y|X]$$

This is done simply by restricting generalization error to focus on data coming from a particular subset D_a. Note that

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(h_\theta(X) \neq Y, X \in D_a)\right] = \mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(h_\theta(X) \neq Y|X \in D_a) \cdot \mathbb{P}_{\mathcal{D}}(X \in D_a)\right]$$

Assuming the independence of the samples drawn, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(h_\theta(X) \neq Y, X \in D_a)\right] = \mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(h_\theta(X) \neq Y|X \in D_a)\right] \\ \cdot \mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(X \in D_a)\right]\end{aligned}$$

From the above, we derive that

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(h_\theta(X) \neq Y|X \in D_a)\right] = \frac{\mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(h_\theta(X) \neq Y, X \in D_a)\right]}{\mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(X \in D_a)\right]} \,. \tag{22}$$

According to the definition of τ in (Feldman, 2020) we have

$$\begin{aligned}\mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathbb{P}_{\mathcal{D}}(X\in D_a)\right] &= \mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\sum_{X\in D_a}\mathcal{D}(X)\right]\\ &= \sum_{X\in D_a}\mathbb{E}_{\mathcal{D}\sim\mathbb{P}[\cdot|D]}\left[\mathcal{D}(X)\right]\\ &= \sum_{X\in D_a}\tau_{l[X]} \qquad \text{(Definition of } \tau\text{)}\end{aligned}$$

Plugging the above back into Eqn 22 gives

$$\mathsf{err}(\theta|D_a) \geq \min_{\theta'\in\Theta}\mathsf{err}(\theta'|D_a) + \sum_{l\in[n]}\frac{\tau_l}{\sum_{X\in D_a}\tau_{l[X]}}\cdot\sum_{X\in D_a\cap S_l}\tilde{\mathbb{P}}[\tilde{Y}\neq Y|X].$$

C.2. Basic Theorem for Proposition 4

We next prove the following:

Proposition 4. *For a set of non-negative numbers $\{b_1,...,b_N\}$ with their associated non-negative weights $\{w_1,...,w_N\}$ such that $\sum_{i=1}^N w_i = 1$. Denote the average as $\bar{b} := \sum_{i=1}^N w_i b_i$. Then*

(1) For $b_i < \bar{b}$, change its weight from w_i to $w_i' < w_i$, and every other weight stays unchanged s.t. $w_j' = w_j$. Given the following renormalization $w_j' = \frac{w_j'}{\sum_i w_i'}$, we have $\bar{b}' := \sum_i w_i' b_i > \bar{b}$.

(2) For any particular $b_i < \bar{b}$, change its b_i to $b_i' < b_i$ and keep other $b_j, j\neq i$ unchanged that $b_j' = b_j$. Furthermore, change its weight from w_i to $w_i' > w_i$, and every other weight stays unchanged s.t. $w_j' = w_j$. Given the following renormalization $w_j' = \frac{w_j'}{\sum_i w_i'}$, we have $\bar{b}' := \sum_i w_i' b_i < \bar{b}$.

Proof. To prove (1), we have

$$\begin{aligned}\bar{b}' - \bar{b} &= \sum_j (w_j' - w_j)\cdot b_j\\ &= \sum_{j\neq i}(w_j' - w_j)\cdot b_j + \left((1-\sum_{j\neq i}w_j') - (1-\sum_{j\neq i}w_j)\right)b_i\\ &= \sum_{j\neq i}(w_j' - w_j)\cdot(b_j - b_i)\end{aligned}$$

Furthermore, let $\Delta = w_i - w_i'$, for $j\neq i$ we have:

$$w_j' - w_j = \frac{w_j}{1-\Delta} - w_j = w_j\cdot\frac{\Delta}{1-\Delta}$$

Therefore we have

$$\begin{aligned}\sum_{j\neq i}(w_j' - w_j)\cdot(b_j - b_i) &= \frac{\Delta}{1-\Delta}\cdot\sum_{j\neq i}w_j(b_j - b_i)\\ &= \frac{\Delta}{1-\Delta}\cdot\left((\bar{b} - w_i\cdot b_i) - (b_i - w_i\cdot b_i)\right)\\ &= \frac{\Delta}{1-\Delta}\cdot(\bar{b} - b_i) > 0\end{aligned}$$

To prove (2), we basically follow the same proof. The only difference is that now let $\Delta = w_i' - w_i$, then for $j \neq i$:

$$w_j' - w_j = \frac{w_j}{1+\Delta} - w_j = -w_j \cdot \frac{\Delta}{1+\Delta}$$

Then we have

$$\begin{aligned}
\bar{b}' - \bar{b} &= \sum_j (w_j' - w_j) \cdot b_j + w_i' \cdot (b_i' - b_i) \\
&= \sum_{j \neq i} (w_j' - w_j) \cdot (b_j - b_i) + w_i' \cdot (b_i' - b_i) \\
&= -\frac{\Delta}{1+\Delta} \cdot \sum_{j \neq i} w_j (b_j - b_i) + w_i' \cdot (b_i' - b_i) \\
&= -\frac{\Delta}{1+\Delta} \cdot \big((\bar{b} - w_i \cdot b_i) - (b_i - w_i \cdot b_i)\big) + w_i' \cdot (b_i' - b_i) \\
&= -\frac{\Delta}{1+\Delta} \cdot (\bar{b} - b_i) + w_i' \cdot (b_i' - b_i) < 0
\end{aligned}$$

□

D. Additional Experimental Results and Details

We include additional details of experiments in this section.

D.1. Dataset Details

The synthetic data is generated using a DAG with specified equations as follows:

$$\begin{aligned}
&X_1 \sim \text{Normal}(0, 1) \\
&A \sim \text{Bernoulli}(0.3) \\
&X_2 \sim \text{Normal}(A, 3) \\
&Z_1 \sim \text{Normal}(0, 1) \\
&X_3 \sim \text{Normal}(2 \cdot Z_1 - 1, 0.1) \\
&X_4 \sim \text{Bernoulli}(0.1) \\
&Y = sign(5 \cdot X_1 \cdot A + 0.2 \cdot X_2^3 + 0.5 \cdot A + 0.3 \cdot X_4 - X_3)
\end{aligned}$$

We use X_1, X_2, X_3, X_4, A as features, A as sensitive attributes, and Y as labels.

We split all tabular datasets randomly into 70% training, 15% validation, and 15% test set. We use the original data splitting in CelebA.

D.2. Experiment Details

We train the logistic regression on synthetic, Adult, and COMPAS using SGD with a learning rate 0.01. For CelebA, we train ResNet18 using Adam with a learning rate 0.001.

Generating Image Counterfactual Samples. When generating image counterfactual samples, we find directly using the generated images from W-GAN does not lead to a satisfactory mitigation performance because the distance between the counterfactual sample and the original sample is too small to impose a change that is large enough to improve fairness (tabular data has no such problem). Therefore we use a heuristic in CIF-based mitigation for image data. Using intervening X as the example, when we map a sample's feature from $X|C = c$ to $X|C = c'$, we get the counterfactual feature $\hat{x}_i = G_{c_i \to \hat{c}_i}(x_i)$. We then search from the real examples $X|C = c'$ to find the nearest neighbor (in the original model's feature space) of $\hat{x}_i$, *i.e.*

$$\hat{x}_i^{'} = \underset{x \sim X|C=c'}{\arg\min} \; ||g_{\hat{\theta}}(\hat{x}_i) - g_{\hat{\theta}}(x)||^2 \tag{23}$$

	A = 0	A = 1
Y = 0	0.45	0.35
Y = 1	0.15	0.55

Table 2. Group-dependent label noise rate added in the training samples in Adult data.

where $g_{\hat{\theta}}$ is the feature extractor of the original model. That is to say, we search from the pool of real samples belonging to the target group closest to the generated fake sample. Since now the counterfactual feature is another real sample in the training data, it is directly removing a sample and replace with another real sample, which induces a larger change than replacing with a fake sample that needs to be reasonably close to the original sample in the W-GAN's training constraint. The resulting counterfactual sample is $\hat{z}_i^{tr}(\hat{c}_i) = (\hat{x}_i^{'}, h_{\hat{\theta}}(\hat{x}_i^{'}), a_i, \hat{c}_i)$. In experiments, we cap the nearest neighbor search space to be 10% of the target group size to reduce the computational cost.

D.3. Additional Mitigation Results

Figure 10-12 show the model accuracy after applying CIF-based mitigation.

D.4. Generated Counterfactual Samples

Figure 13 and 14 show some random examples of generated images in CelebA when intervening A.

D.5. Distribution of Influence Values

Figure 15 shows the distribution of influence values computed on COMPAS corresponding to three fairness metrics.

D.6. Details of Experiments on Additional Applications

Fixing Mislabelling. The group-dependent label noise rate we add to the Adult training dataset is shown in Table 2. We follow a similar experimental setting in (Wang et al., 2021). After the label intervention, the bias increases significantly: DP increases from 16.1% to 49.6%, EOP increases from 31.4% to 77.3%, and EO increases from 19.1% to 63.3%.

We flag samples by choosing samples with top influence when Y is intervened and report the precision ($\frac{\#\text{flipped labels correctly detected}}{\#\text{flagged labels}}$) of our detection.

Defending against Poisoning Attacks. After the training samples are poisoned, the model unfairness increases as follows: DP increases from 16.1% to 61.4%, EOP increases from 31.4% to 69.4%, and EO increases from 19.1% to 65.2%.

Resampling Imbalanced Representations. After artificially unbalancing the training samples, the fairness gap increases as follows: DP increases from 16.1% to 42.6%, EOP increases from 31.4% to 63.7%, and EO increases from 19.1% to 47.2%.

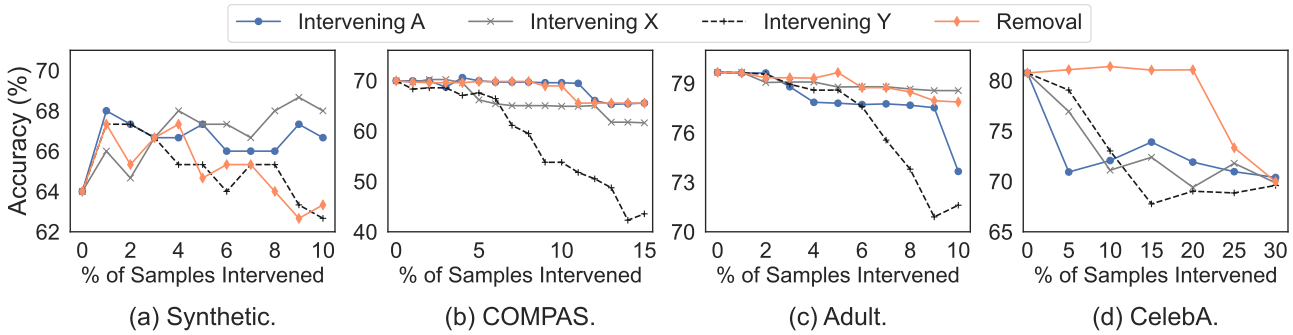


Figure 10. Model accuracy with CIF-based mitigation using fairness measure Demographic Parity (DP).

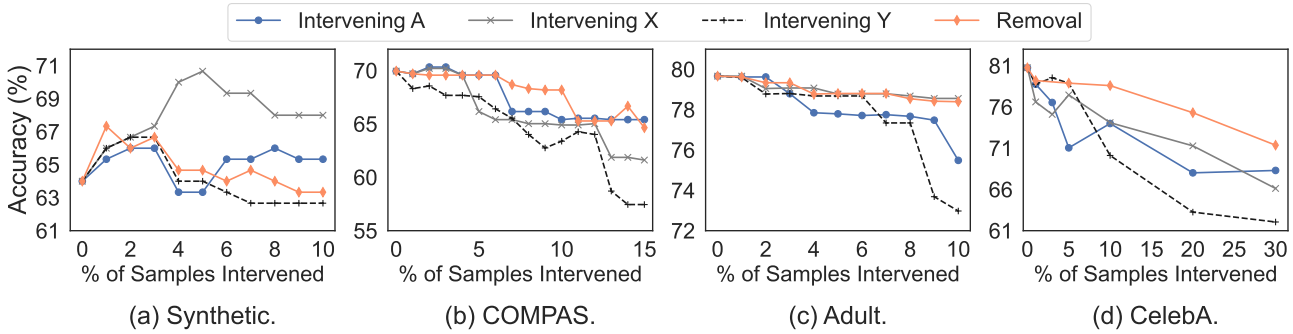


Figure 11. Model accuracy with CIF-based mitigation using fairness measure Equality of Opportunity (EOP).

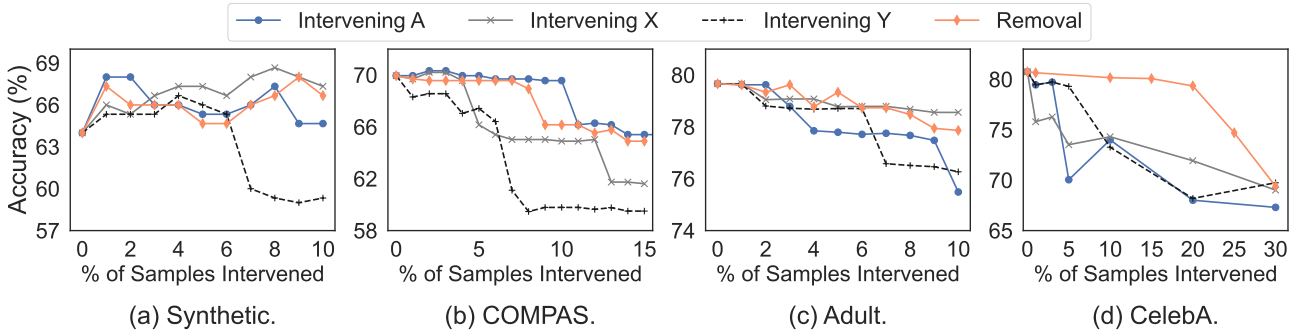


Figure 12. Model accuracy with CIF-based mitigation using fairness measure Equality of Odds (EO).

Figure 13. W-GAN generated images that map from male to female in CelebA.

Figure 14. W-GAN generated images that map from female to male in CelebA.

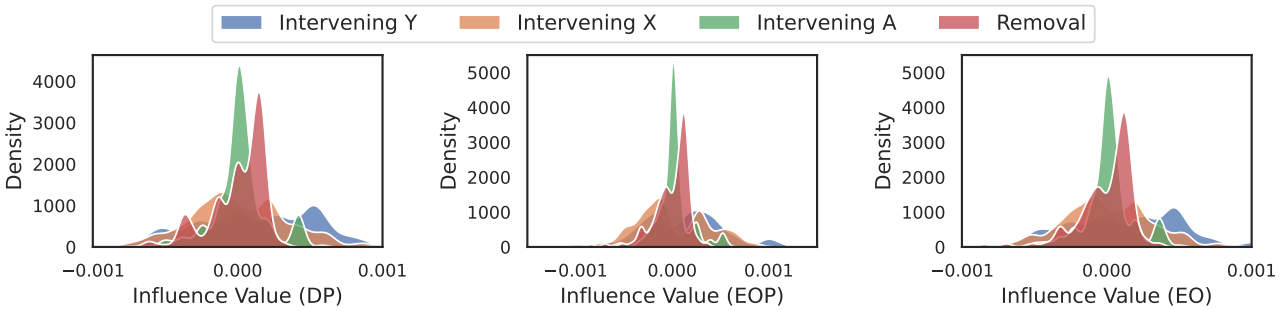


Figure 15. Distribution of influence values computed on COMPAS across three fairness metrics.