Learning Individually Fair Classifier with Path-Specific Causal-Effect Constraint

Yoichi Chikahara[1,3] **Shinsaku Sakaue**[2] **Akinori Fujino**[1] **Hisashi Kashima**[3]
[1]NTT [2]The University of Tokyo [3]Kyoto University

Abstract

Machine learning is used to make decisions for individuals in various fields, which require us to achieve good prediction accuracy while ensuring fairness with respect to sensitive features (e.g., race and gender). This problem, however, remains difficult in complex real-world scenarios. To quantify unfairness under such situations, existing methods utilize *path-specific causal effects*. However, none of them can ensure fairness for each individual without making impractical functional assumptions about the data. In this paper, we propose a far more practical framework for learning an individually fair classifier. To avoid restrictive functional assumptions, we define the *probability of individual unfairness* (PIU) and solve an optimization problem where PIU's upper bound, which can be estimated from data, is controlled to be close to zero. We elucidate why our method can guarantee fairness for each individual. Experimental results show that our method can learn an individually fair classifier at a slight cost of accuracy.

1 INTRODUCTION

Machine learning is increasingly being used to make critical decisions that severely affect people's lives (e.g., loan approvals (Khandani et al., 2010), hiring decisions (Houser, 2019), and recidivism predictions (Angwin et al., 2016)). The huge societal impact of such decisions on people's lives raises concerns about fairness because these decisions may be discriminatory with respect to *sensitive features*, including race, gender, religion, and sexual orientation.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Although many researchers have studied how to make fair decisions while achieving high prediction accuracy (Dwork et al., 2012; Feldman et al., 2015; Hardt et al., 2016), it remains a challenge in complex real-world scenarios. For instance, consider hiring decisions for physically demanding jobs. Although it is discriminatory to reject applicants based on gender, since the job requires physical strength, it is sometimes **not** discriminatory to reject the applicants due to physical strength. Since physical strength is often affected by gender, rejecting applicants due to physical strength leads to gender difference in the rejection rates. Although this difference due to physical strength is **not** always unfair, it is removed when using traditional methods (e.g., Feldman et al. (2015)). Consequently, even if there is a man who has a much more physical strength than a woman, these methods might reject him to accept her, which severely reduces the prediction accuracy.

To achieve high prediction accuracy, we need to remove only an unfair difference in decision outcomes. To measure this difference, existing methods utilize a path-specific causal effect (Avin et al., 2005), which we call an *unfair effect*. Using unfair effects, the *path-specific counterfactual fairness* (PSCF) method (Chiappa and Gillam, 2019) aims to guarantee fairness for each individual; however, achieving such an individual-level fairness is possible only when the data are generated by a restricted class of functions. By contrast, *fair inference on outcome* (FIO) (Nabi and Shpitser, 2018) does not require such demanding functional assumptions; however, it cannot ensure individual-level fairness.

The goal of this paper is to propose a learning framework that guarantees individual-level fairness without making impractical functional assumptions. For this goal, we train a classifier by forcing the *probability of individual unfairness* (PIU), defined as the probability that an unfair effect is non-zero, to be close to zero. This, however, is difficult to achieve because we cannot estimate PIU from data. To overcome this difficulty, we derive its upper bound that can be estimated from data and solve a penalized optimization problem where the upper-bound value is controlled to be close to zero.

Table 1: Comparison with existing methods

Method	Individually fair	Functional assumptions
Our method	Yes	Unnecessary
PSCF	Yes	Necessary
FIO	No	Unnecessary

Our contributions are summarized as follows:

- We establish a framework that guarantees fairness for each individual without restrictive functional assumptions on the data (Table 1). To achieve this, we make the PIU value close to zero by imposing a penalty that reduces its upper bound value, which can be estimated from data.
- We elucidate why imposing such a penalty guarantees individual-level fairness in Sections 4.3.3 and 4.4. We also show how our method can be extended to address cases where there are unobserved variables called *latent confounders* in Section 4.5.
- We experimentally show that our method makes much fairer predictions for each individual than the existing methods at a slight cost of prediction accuracy.

2 PRELIMINARIES

2.1 Problem Statement

In this paper, we consider a binary classification task. We train classifier h_θ with parameter θ to predict decision outcome $Y \in \{0, 1\}$ from the features of each individual $\boldsymbol{X}$, which contains sensitive feature $A \in \{0, 1\}$.

We seek classifier parameter θ that achieves a good balance between prediction accuracy and fairness with respect to sensitive feature A. Suppose that we have loss function L_θ and penalty function G_θ, which respectively measure prediction errors and unfairness based on θ. Formally, given n training instances $\{(\boldsymbol{x}_i, y_i)_{i=1}^n\}$, our learning problem is formulated as follows:

$$\min_\theta \quad \frac{1}{n}\sum_{i=1}^{n} L_\theta(\boldsymbol{x}_i, y_i) + \lambda G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n), \qquad (1)$$

where $\lambda \geq 0$ is a hyperparameter.

To achieve a high prediction accuracy, penalty function G_θ must be designed such that we can avoid imposing unnecessary penalizations. To do so, we utilize a *causal graph*, which is a directed acyclic graph (DAG) whose nodes and edges represent random variables and causal relationships, respectively (Pearl, 2009). We assume that a causal graph is provided by domain experts

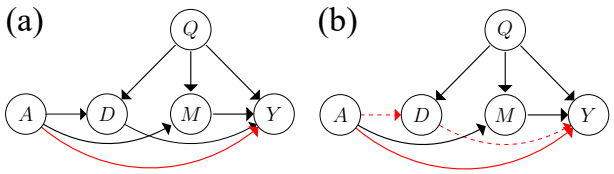

Figure 1: Causal graphs representing a scenario of hiring decisions for physically demanding jobs: Unfair pathways are (a): red solid edge $A \to Y$; (b): $A \to Y$ and red dashed pathway $A \to D \to Y$.

or can be inferred from data (Glymour et al., 2019); this assumption is common in many existing methods (Chiappa and Gillam, 2019; Kusner et al., 2017; Nabi and Shpitser, 2018; Zhang et al., 2017).

As an example of a causal graph, consider a scenario for hiring decisions for a physically demanding job. In this scenario, a causal graph might be given, as shown in Figure 1(a), where $A, Q, D, M \in \boldsymbol{X}$ represent gender, qualifications, the number of children, and physical strength, respectively. This graph expresses our knowledge that prediction Y is unfair only if it is based on gender A. To do so, we regard direct pathway $A \to Y$ as unfair pathway π (i.e., $\pi = \{A \to Y\}$).

In fact, in the above simple case, we can naively remove the unfairness by making a prediction without gender A; however, this is insufficient when we consider a complex scenario with multiple unfair pathways.

For instance, as shown in Figure 1(b), we may regard not only $A \to Y$ but also the pathway through the number of children D ($A \to D \to Y$) as unfair (because it is also discriminatory to reject women because of the possibility of bearing children). In this case, a naive approach to ensure fairness is to predict without A or D. This, however, might unnecessarily decrease the prediction accuracy. For instance, consider a case where number of children D is only slightly affected by gender A (e.g., the applicants had gender-equitable opportunities to take parental leave in the past) and largely influenced by other unobserved features that are important for prediction (e.g., communication skills). Then predicting without D will seriously decrease the accuracy while contributing almost nothing to fairness.

To address such cases, given unfair pathways π, we design penalty function G_θ by quantifying the unfairness based on data. To do so, we utilize path-specific causal effects, which are described in the next section.

2.2 Path-Specific Causal Effects

A path-specific causal effect measures how largely an observed variable influences another variable via path-

ways in a causal graph (Avin et al., 2005). Although prediction Y is not observed but is given by classifier h_θ, we can utilize this measure to quantify the influence of sensitive feature A on Y via unfair pathways π.

With a path-specific causal effect, this influence is measured by the difference of the two predictions, which is obtained by modifying input features $\boldsymbol{X}$. To illustrate these predictions, consider a case where sensitive feature A is gender. Then for each woman ($A = 0$), one prediction is made by directly taking her attributes as input, and another is made with *counterfactual* attributes, which would be observed if she were male ($A = 1$); for each man, these predictions are made using the counterfactual attributes that would be if he were female (see, Appendix A.1 for details). Although such counterfactual attributes are not observed, they can be computed by a *structural equation model* (SEM).

An SEM consists of *structural equations*, each of which expresses variable $V \in \{\boldsymbol{X}, Y\}$ by deterministic function f_V (Pearl, 2009). Each function f_V takes as input two types of variables. One is observed variables, which are the parents of V in a causal graph, and the other is unobserved noise U_V, which expresses random variable V using deterministic function f_V.

For instance, structural equations over $D, M \in \boldsymbol{X}$ in the causal graph in Figure 1(b) may be formulated as

$$\begin{aligned} D &= f_D(A, Q, U_D) = A + U_D Q, \\ M &= f_M(A, Q, U_M) = 3A + 0.5Q + U_M, \end{aligned} \tag{2}$$

where U_D is multiplicative noise and U_M is additive noise. By contrast, the structural equation over prediction Y is formulated using classifier h_θ. If h_θ is deterministic, it is expressed as $Y = h_\theta(A, Q, D, M)$; otherwise, $Y = h_\theta(A, Q, D, M, U_Y)$, where U_Y is a random variable used in the classifier. See Appendix A.2.1 for a formal definition of SEM in our setting.

Structural equations (2) can be used to compute the (counterfactual) attributes of D and M that are observed when $A = a$ ($a \in \{0, 1\}$) as

$$D(a) = a + U_D Q, \quad M(a) = 3a + 0.5Q + U_M. \tag{3}$$

If (3) is available, we can obtain attributes $D(0)$, $D(1)$, $M(0)$, and $M(1)$ for each individual.

Using these attributes, we can compute a path-specific causal effect for each individual, which we call an unfair effect. For instance, when measuring the influence via unfair pathways $\pi = \{A \to Y, A \to D \to Y\}$ in Figure 1(b), we define an unfair effect as the difference of two predictions $Y_{A \Leftarrow 1\|\pi} - Y_{A \Leftarrow 0}$, where $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1\|\pi}$ are called *potential outcomes* and given as

$$\begin{aligned} Y_{A \Leftarrow 0} &= h_\theta(0, Q, D(0), M(0)), \\ Y_{A \Leftarrow 1\|\pi} &= h_\theta(1, Q, D(1), M(0)). \end{aligned} \tag{4}$$

In (4), the inputs of $Y_{A \Leftarrow 0}$ are $A = 0$, $D(0)$, and $M(0)$, all of which are given using the same value, $a = 0$. By contrast, the inputs of $Y_{A \Leftarrow 1\|\pi}$ are formulated based on unfair pathways π; we use the value $a = 1$ **only** for A and D (i.e., $A = 1$ and $D(1)$), which correspond to the nodes on $\pi = \{A \to Y, A \to D \to Y\}$ (see Appendix A.3 for the formal definition).[1]

In practice, however, we cannot compute an unfair effect for each individual. This is because we cannot formulate an SEM since it requires a deep understanding of true data-generating processes; consequently, for instance, we can obtain $D(a)$ and $M(a)$ in (3) **only** for either $a = 0$ or $a = 1$ but **not both**. Due to this issue, existing methods use the (conditional) expected values of unfair effects, which can be estimated from data.

3 EXISTING METHODS AND THEIR WEAKNESSES

Using unfair effects, two types of existing methods have been proposed. Unfortunately, as presented in Table 1, each has a weakness. One requires restrictive functional assumptions, and the other cannot ensure individual-level fairness. Below we describe their details.

3.1 Methods for Ensuring Individual-Level Fairness

The PSCF method (Chiappa and Gillam, 2019) aims to satisfy the following individual-level fairness criterion:

Definition 1 (Wu et al. (2019b)) *Given unfair pathways π in a causal graph, classifier h_θ achieves a **(path-specific) individual-level fairness** if*

$$\mathbb{E}_{Y_{A \Leftarrow 0}, Y_{A \Leftarrow 1\|\pi}}[Y_{A \Leftarrow 1\|\pi} - Y_{A \Leftarrow 0} | \boldsymbol{X} = \boldsymbol{x}] = 0 \tag{5}$$

holds for any value of $\boldsymbol{x}$ of input features $\boldsymbol{X}$.

Condition (5) states that classifier h_θ is individually fair if the *conditional mean unfair effect* is zero, which is an average over individuals who have identical attributes for all features in $\boldsymbol{X}$. Since $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1\|\pi}$ are expressed using classifier parameter θ as shown in (4), we need to find appropriate θ values to satisfy (5).

Unfortunately, such θ values can be found only in restricted cases. As pointed out by Wu et al. (2019b), this is because we cannot always estimate the conditional mean unfair effect in (5). For instance, when potential

[1] We can also consider different potential outcomes $Y_{A \Leftarrow 1}$ and $Y_{A \Leftarrow 0\|\pi}$, where all inputs of $Y_{A \Leftarrow 1}$ are given using the value $a = 1$, and $Y_{A \Leftarrow 0\|\pi}$ is formulated using $a = 0$ only for the inputs that correspond to the nodes on pathways π.

outcomes are given as (4), since the conditional mean unfair effect is conditioned on A and D, estimating it requires the joint distribution of $D(0)$ and $D(1)$. This joint distribution, however, is unavailable because we cannot jointly obtain them as explained in Section 2.2.

Due to this issue, the PSCF method (and the one in (Kusner et al., 2017, Section S4)) can achieve individual-level fairness only when the data are generated from a restricted functional class of SEMs. Specifically, these existing methods assume that each variable V follows an additive noise model $V = f_V(\boldsymbol{pa}(V)) + U_V$, where $\boldsymbol{pa}(V)$ denotes the parents of V in the causal graph. Unfortunately, this model cannot express data-generating processes in many cases. For instance, it cannot express variable D in (2) due to multiplicative noise U_D. Traditionally, this assumption has been used to infer causal graphs (Hoyer et al., 2009; Shimizu et al., 2006). However, as mentioned in Glymour et al. (2019), since more recent causal graph discovery methods require much weaker assumptions (Zhang and Hyvärinen, 2009; Stegle et al., 2010), the presence of such an assumption severely restricts the scope of their applications.

3.2 Another Method for Removing Unfair Effects

To avoid the aforementioned restrictive functional assumption, the FIO method (Nabi and Shpitser, 2018) aims to remove the *mean unfair effect* over **all** individuals, which is expressed as

$$\begin{aligned}&\mathbb{E}_{Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}}[Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0}] \\ &= \mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1) - \mathrm{P}(Y_{A\Leftarrow 0} = 1).\end{aligned} \tag{6}$$

In (6), marginal probabilities $\mathrm{P}(Y_{A\Leftarrow 0} = 1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$ can be estimated under much weaker assumptions than the conditional mean unfair effect in (5). We detail these assumptions in Appendix B and how to derive the estimators in Appendix D.

However, removing this mean unfair effect does not imply individual-level fairness. This is because depending on input features $\boldsymbol{X}$, unfair effects might be largely positive for some individuals and largely negative for others, which is seriously discriminatory for these individuals. Note that we cannot resolve this issue simply using e.g., the mean of the absolute values of the unfair effects. This is because estimating such a quantity requires a joint distribution of $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$; however, this joint distribution is unavailable because we cannot obtain both $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ for each individual without an SEM.

4 PROPOSED METHOD

4.1 Overcoming Weaknesses of Existing Methods

To resolve the weaknesses of the existing methods, we propose a framework that guarantees individual-level fairness without restrictive functional assumptions.

For this goal, we aim to train a classifier by forcing an unfair effect to be zero for **all** individuals: i.e., making potential outcomes take the same value (i.e., $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 0$ or $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 1$) with probability 1 regardless of the values of input features $\boldsymbol{X}$. This is sufficient to satisfy the individual-level fairness condition (Definition 1) because it restricts the potential outcome values more severely than the latter condition, where potential outcomes $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi}$ can take 0 or 1 depending on $\boldsymbol{X}$'s values. Although such a fairness condition may be overly severe and might decrease the prediction accuracy, in Section 5 we experimentally show that our method can achieve comparable accuracy to the existing method for ensuring individual-level fairness (i.e., the PSCF method (Chiappa and Gillam, 2019)).

Compared with PSCF, our method has a clear advantage in that it requires much weaker assumptions. We only need to estimate the marginal potential outcome probabilities in (6), which only requires several conditional independence relations and the graphical condition on unfair pathways π (see Appendix B for our assumptions). Furthermore, we can relax these assumptions to address some cases where there are unobserved variables called latent confounders (Section 4.5).

4.2 Achieving Individual-Level Fairness with PIU

We aim to make potential outcomes take the same value for all individuals. To this end, we formulate penalty function G_θ based on the following quantity:

Definition 2 *For unfair pathways π in a causal graph and potential outcomes $Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi} \in \{0, 1\}$, we define the* ***probability of individual unfairness (PIU)*** *by* $\mathrm{P}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})$.

Intuitively, PIU is the probability that potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ take different values.

Unlike the conditional mean unfair effect in Definition 1, PIU is not conditioned on features $\boldsymbol{X}$ of each individual.

Nonetheless, PIU can be used to guarantee individual-level fairness. By constraining PIU to zero, we can

guarantee that potential outcomes take the same value (i.e., $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 0$ or $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 1$) with probability 1 regardless of the values of $\boldsymbol{X}$, which is sufficient to ensure individual-level fairness.

Unfortunately, we cannot directly impose constraints on PIU. This is because estimating the PIU value requires the joint distribution of $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$, which is unavailable as described in Section 3.2.

To overcome this issue, instead of PIU, we utilize its upper bound that can be estimated from data. Specifically, to make the PIU value close to zero, we formulate a penalty function that forces the upper bound on PIU to be nearly zero, which is described in the next section.

4.3 Penalty By Upper Bound on PIU

4.3.1 Upper Bound Formulation

To make the PIU value small, we utilize the following upper bound on PIU:

Theorem 1 (Upper bound on PIU) *Suppose that potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ are binary. Then for any joint distribution of potential outcomes $\mathrm{P}(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi})$, PIU is upper bounded as follows:*

$$\mathrm{P}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}) \leq 2\,\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}), \quad (7)$$

where P^I is an independent joint distribution, i.e., $\mathrm{P}^I(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}) = \mathrm{P}(Y_{A\Leftarrow 0})\,\mathrm{P}(Y_{A\Leftarrow 1\|\pi})$.

The proof is detailed in Appendix C. Theorem 1 states that whatever joint distribution potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ follow, the resulting PIU value is at most twice the PIU value that is approximated with independent joint distribution P^I.

Note that this upper bound can be larger than 1, and if so, the PIU value is not controlled because PIU is at most 1. However, since PIU is always smaller than its upper bound, by making the upper bound close to zero, we can guarantee that PIU is also close to zero.

4.3.2 Estimating Upper Bound

Using the observed data, we estimate the upper bound on PIU in (7), which is twice the value of the approximated PIU. Recall that this approximated PIU is the probability that potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ take different values when they are independent. Since potential outcomes are binary, it is expressed as the probability that potential outcome values are $(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}) = (0,1)$ or $(1,0)$; in other words,

$$\begin{aligned}&\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})\\ =&\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1)(1-\mathrm{P}(Y_{A\Leftarrow 0}=1))\\ &+(1-\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1))\,\mathrm{P}(Y_{A\Leftarrow 0}=1).\end{aligned} \quad (8)$$

Various estimators can be used to estimate marginal probabilities $\mathrm{P}(Y_{A\Leftarrow 0}=1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1)$ in (8). Among them, we utilize the computationally efficient estimator in Huber (2014), which can be computed in $O(n)$ time, where n is the number of training instances.

Let $c_\theta(\boldsymbol{X}) = \mathrm{P}(Y=1|\boldsymbol{X})$ denote the conditional distribution provided by classifier h_θ; we let $c_\theta(\boldsymbol{X}) = h_\theta(\boldsymbol{X}) \in \{0,1\}$ if h_θ is a deterministic classifier. For instance, suppose that the causal graph is given as shown in Figure 1(b) and that the features of n individuals are provided as $\{\boldsymbol{x}_i\}_{i=1}^n = \{a_i, q_i, d_i, m_i\}_{i=1}^n$. Then $\mathrm{P}(Y_{A\Leftarrow 0}=1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1)$ can be estimated as the following weighted averages:

$$\begin{aligned}\hat{p}_\theta^{A\Leftarrow 0} &= \frac{1}{n}\sum_{i=1}^n \mathbf{1}(a_i=0)\hat{w}_i c_\theta(a_i,q_i,d_i,m_i) \text{ and}\\ \hat{p}_\theta^{A\Leftarrow 1\|\pi} &= \frac{1}{n}\sum_{i=1}^n \mathbf{1}(a_i=1)\hat{w}'_i c_\theta(a_i,q_i,d_i,m_i),\end{aligned} \quad (9)$$

where $\mathbf{1}(\cdot)$ is an indicator function, and $\hat{w}_i$ and $\hat{w}'_i$ are the following weights for individual $i \in \{1,\dots,n\}$:

$$\hat{w}_i = \frac{1}{\hat{\mathrm{P}}(A=0|q_i)},$$

$$\hat{w}'_i = \frac{\hat{\mathrm{P}}(A=1|q_i,d_i)\hat{\mathrm{P}}(A=0|q_i,d_i,m_i)}{\hat{\mathrm{P}}(A=1|q_i)\hat{\mathrm{P}}(A=0|q_i,d_i)\hat{\mathrm{P}}(A=1|q_i,d_i,m_i)},$$

where $\hat{\mathrm{P}}$ is a conditional distribution, which we infer in the same way as Zhang and Bareinboim (2018a), i.e., by learning a statistical model (e.g., a neural network) from the training data beforehand.[2] We derive the estimators (9) in Appendix D.

In (9), marginal probabilities $\hat{p}_\theta^{A\Leftarrow 0}$ and $\hat{p}_\theta^{A\Leftarrow 1\|\pi}$ are estimated by taking a weighted average of conditional probability c_θ over individuals with $A=0$ and $A=1$, respectively, which is a widely used estimation technique called *inverse probability weighting* (IPW).

4.3.3 Formulating Penalty Function

To learn an individually fair classifier, we force the estimated value of the upper bound on PIU to be close

[2]Note that FIO infers conditional distributions not by learning statistical models beforehand but by simultaneously learning them with the predictive model of Y (Nabi and Shpitser, 2018). This is because unlike our method, it addresses not only training a classifier but also learning a generative model of joint distribution $\mathrm{P}(\mathbf{X}, Y)$.

to zero by minimizing the objective function (1) with the following penalty function G_θ:

$$G_\theta(\boldsymbol{x}_1, \dots \boldsymbol{x}_n) = \hat{p}_\theta^{A \Leftarrow 1\|\pi}(1 - \hat{p}_\theta^{A \Leftarrow 0}) + (1 - \hat{p}_\theta^{A \Leftarrow 1\|\pi})\hat{p}_\theta^{A \Leftarrow 0}. \quad (10)$$

For instance, if $\hat{p}_\theta^{A \Leftarrow 0}$ and $\hat{p}_\theta^{A \Leftarrow 1\|\pi}$ are given as the weighted estimators (9), to reduce the value of G_θ, our method imposes strong penalties on the predictions for individuals whose weights $\hat{w}_i$ and $\hat{w}'_i$ are large.

In our experiments, we minimize the objective function using the stochastic gradient descent method (Sutskever et al., 2013). We discuss the computation time and convergence guarantees in Appendix E.

From the penalty function (10), we can see why penalizing the upper bound on PIU guarantees individual-level fairness. As the penalty parameter value goes to infinity ($\lambda \to \infty$), the marginal probabilities $(\hat{p}_\theta^{A \Leftarrow 0}, \hat{p}_\theta^{A \Leftarrow 1\|\pi})$ approach $(0, 0)$ or $(1, 1)$. This guarantees that the potential outcomes take the same value with probability 1, which is sufficient to guarantee individual-level fairness.

4.4 Comparison with Existing Fairness Constraint

To show the effectiveness of penalty function (10), we compare it with the constraint of the FIO method.

Suppose that our penalty function forces the upper bound on PIU to satisfy the following condition:

$$\hat{p}_\theta^{A \Leftarrow 1\|\pi}(1 - \hat{p}_\theta^{A \Leftarrow 0}) + (1 - \hat{p}_\theta^{A \Leftarrow 1\|\pi})\hat{p}_\theta^{A \Leftarrow 0} \leq \delta. \quad (11)$$

Here we let constant δ be $\delta \in [0, 1]$ because otherwise we cannot force the PIU value to be less than 1.

Meanwhile, the FIO constraint limits the mean unfair effect (6) to lie in

$$-\delta' \leq \hat{p}_\theta^{A \Leftarrow 1\|\pi} - \hat{p}_\theta^{A \Leftarrow 0} \leq \delta', \quad (12)$$

where $\delta' \in [0, 1]$ is a hyperparameter. If $\delta' = 0$, it ensures $\hat{p}_\theta^{A \Leftarrow 0} = \hat{p}_\theta^{A \Leftarrow 1\|\pi}$.

Figure 2 shows the feasible region of our fairness condition (red) and the FIO constraint (blue), respectively, obtained by graphing the hyperbolic inequality in (11) and the linear inequality in (12). Here, to clarify their difference, we consider the case where $\delta = 2\delta'$.

When $\delta \approx 0$, our fairness condition only accepts region $(\hat{p}_\theta^{A \Leftarrow 0}, \hat{p}_\theta^{A \Leftarrow 1\|\pi}) \approx (0, 0)$ or $(1, 1)$, where the potential outcomes are likely to take the same value; hence, the unfair effect is likely to be zero. This demonstrates how effectively our condition removes an unfair effect for each individual. By contrast, with any δ' value, the

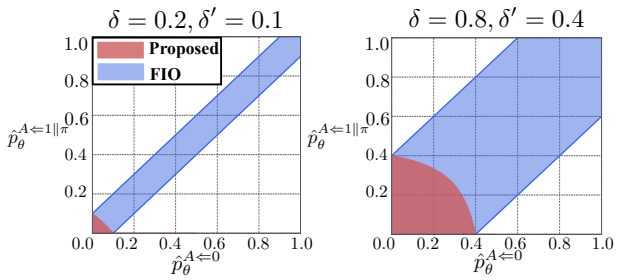

Figure 2: Feasible regions of our constraint (red) and FIO (blue) with $(\delta, \delta') = (0.2, 0.1), (0.8, 0.4)$

FIO constraint always accepts point $(\hat{p}_\theta^{A \Leftarrow 0}, \hat{p}_\theta^{A \Leftarrow 1\|\pi}) = (0.5, 0.5)$, where it is completely uncertain whether potential outcomes take the same value as detailed in Appendix F. This implies that FIO predictions might be unfair for some individuals, indicating that FIO cannot ensure individual-level fairness.

4.5 Extension for Dealing with Latent Confounders

So far, we have assumed that the marginal probabilities of potential outcomes can be estimated from data. This assumption, however, does not hold if there is a latent confounder (Pearl, 2009), i.e., an unobserved variable that is a parent of the observed variables in the causal graph. Although this is possible in practice, inferring marginal probabilities becomes much more challenging.

However, even in the presence of latent confounders, our method can ensure individual-level fairness if the lower and upper bounds on marginal probabilities are available. For instance, the bounds of Miles et al. (2017) can be used when there is only a single *mediator* (i.e., a feature affected by sensitive feature A), and it takes discrete values. In such cases, we can achieve individual-level fairness by reformulating the penalty function as follows.

Suppose that marginal probabilities $\mathrm{P}(Y_{A \Leftarrow 0} = 1)$ and $\mathrm{P}(Y_{A \Leftarrow 1\|\pi} = 1)$ are bounded by

$$\hat{l}_\theta^{A \Leftarrow 0} \leq \mathrm{P}(Y_{A \Leftarrow 0} = 1) \leq \hat{u}_\theta^{A \Leftarrow 0},$$
$$\hat{l}_\theta^{A \Leftarrow 1\|\pi} \leq \mathrm{P}(Y_{A \Leftarrow 1\|\pi} = 1) \leq \hat{u}_\theta^{A \Leftarrow 1\|\pi},$$

where $\hat{l}_\theta^{A \Leftarrow 0}$, $\hat{u}_\theta^{A \Leftarrow 0}$, $\hat{l}_\theta^{A \Leftarrow 1\|\pi}$, and $\hat{u}_\theta^{A \Leftarrow 1\|\pi}$ are the estimated lower and upper bounds, which we describe in Appendix G. Then for any marginal probability values, the upper bound on PIU in (7) is always smaller than twice the value of

$$G_\theta(\boldsymbol{x}_1, \dots \boldsymbol{x}_n) = \hat{u}_\theta^{A \Leftarrow 1\|\pi}(1 - \hat{l}_\theta^{A \Leftarrow 0}) + (1 - \hat{l}_\theta^{A \Leftarrow 1\|\pi})\hat{u}_\theta^{A \Leftarrow 0}. \quad (13)$$

Table 2: Test accuracy (%) on each dataset

Method	Synth	German	Adult
Proposed	80.0 ± 0.9	75.0	75.2
FIO	84.8 ± 0.6	78.0	81.2
PSCF	74.8 ± 1.6	76.0	73.4
Unconstrained	88.2 ± 0.9	81.0	83.2
Remove	76.9 ± 1.3	73.0	74.7

Therefore, by making this penalty function value nearly zero, we can achieve individual-level fairness. We experimentally confirmed that this framework makes fairer predictions than the original one in Appendix I.5.

5 EXPERIMENTS

We compared our method (**Proposed**) with the following four baselines: (1) **FIO** (Nabi and Shpitser, 2018), (2) **PSCF** (Chiappa and Gillam, 2019), which aims to achieve individual-level fairness by assuming that the data are generated by additive noise models, (3) **Unconstrained**, which does not use any constraints or penalty terms related to fairness, and (4) **Remove** (Kusner et al., 2017, Section S4), which removes unfair effects simply by making predictions without input features that correspond to the nodes on unfair pathways π. As classifiers of **Proposed**, **FIO**, **Unconstrained**, and **Remove**, we used a feed-forward neural network that contains two linear layers with 100 and 50 hidden neurons. Other settings are detailed in Appendix H.1.

Data and causal graphs: For a performance evaluation, we used a synthetic dataset and two real-world datasets: the German credit dataset and the Adult dataset (Bache and Lichman, 2013). We sampled the synthetic data from the SEM, whose formulation is described in Appendix H.2.1. To define the unfair effect, we used the causal graph in Figure 1(b). With the real-world datasets, we evaluated the performance as follows. With the German dataset, we predicted whether each loan applicant is risky (Y) from their features such as gender A and savings $\boldsymbol{S}$. With the Adult dataset, we predicted whether an annual income exceeds \$50,000 (Y) from features such as gender A and marital status M. To measure unfairness, following (Chiappa and Gillam, 2019), we used the causal graphs in Figure 3, which we detail in Appendix H.3.1.

Accuracy and fairness: We evaluated the test accuracy and four statistics of the unfair effects: (i) the mean unfair effect (6), (ii) the standard deviation in the conditional mean unfair effects in (5), (iii) the upper bound on PIU, and (iv) the PIU.

Table 2 and Figure 4 present the test accuracy and the four statistics of the unfair effects, respectively. In synthetic data experiments, we computed the means and the standard deviations based on 10 experiments with randomly generated training and test data. Figure 4 does not display two statistics (ii) and (iv) for the German and Adult datasets because computing them requires SEMs, which are unavailable for these real-world datasets. We do not show (i) or (iii) of **PSCF** because these are not well-defined for this method (see Appendix H.2.3 for details).

With **Proposed**, all the statistics of the unfair effects were sufficiently close to zero, demonstrating that it made fair predictions for all individuals. This is because by imposing a penalty on the upper bound on PIU, **Proposed** forced unfair effect values to be close to zero for all individuals, guaranteeing that the other statistics are close to zero.

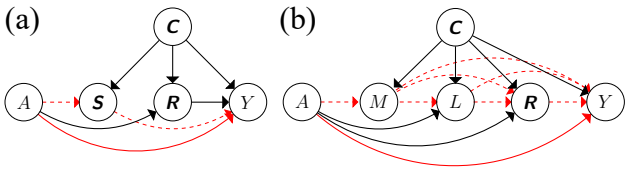

Figure 3: Causal graphs for (a) German credit dataset and (b) Adult dataset: Direct pathways with red solid edges are unfair. Red dashed pathway $A \to S \to Y$ is unfair in (a), and those that go through M (i.e., $A \to M \to \cdots \to Y$) are unfair in (b).

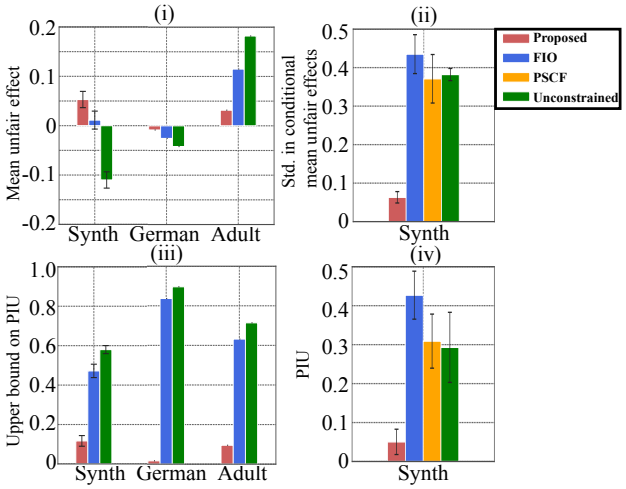

Figure 4: Four statistics of unfair effects on test data: The closer they are to zero, the fairer predictions are. Error bars of synthetic data (Synth) denote standard deviations in 10 runs with randomly generated data. With **Remove**, all statistics are zero (not shown). With **PSCF**, (i) and (iii) are not well-defined as described in Appendix H.2.3.

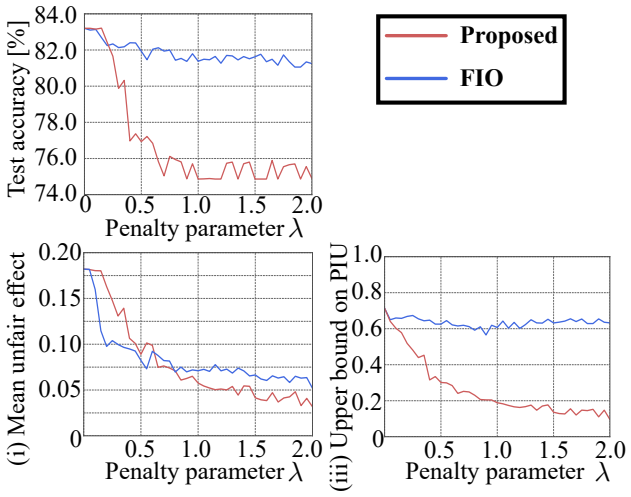

Figure 5: Test accuracy, mean unfair effect, and upper bound on PIU of (a) **Proposed** (red) and (b) **FIO** (blue) for Adult dataset when increasing penalty parameter value as $\lambda = 0, 0.05, 0.10, \cdots, 2.00$.

By contrast, regarding **FIO** and **PSCF**, the unfair effect values were much larger. With **FIO**, although the mean unfair effect (i.e., (i)) was close to zero, the other statistics deviated from zero, indicating that constraining the mean unfair effect did not ensure individual-level fairness. **PSCF** failed to reduce the value of the standard deviation in the conditional mean unfair effects (i.e., (ii)). This is because the data are not generated from additive noise models (see Appendix H.2.1 for the data), violating the functional assumption of **PSCF**. Since the large values of (ii) imply that unfair effects are greatly affected by the attributes of input features $\boldsymbol{X}$, these results indicate that **FIO** and **PSCF** made unfair predictions based on these attributes. With real-world datasets, **FIO** provided large values of the upper bound on PIU (i.e., (iii)). These upper bound values cast much doubt on whether the predictions of **FIO** are fair for each individual, which is problematic in practice.

The test accuracy of **Proposed** was lower than **FIO**, higher than **Remove**, and comparable to **PSCF**. Since **FIO** imposes a much weaker fairness constraint than **Proposed**, **Remove**, and **PSCF** (i.e., the methods designed for achieving individual-level fairness), it achieved higher accuracy. By contrast, since **Remove** eliminates all informative input features that are affected by the sensitive feature to guarantee individual-level fairness, it provided the lowest accuracy. The comparison of **Proposed** and **PSCF** indicates that although our method employs a more severe fairness condition than **PSCF**, it barely sacrifices prediction accuracy, demonstrating that it strikes a better balance between individual-level fairness and accuracy.

Penalty parameter effects: Using the Adult dataset, we further compared the performance of **Proposed** with **FIO** using various penalty parameter values. Regarding unfair effects, we evaluated two statistics (i) the mean unfair effect and (iii) the upper bound on PIU since the others are unavailable with this real-world dataset, as already described above.

Figure 5 presents the results. As the penalty parameter value increased, the test accuracy of **Proposed** decreased more sharply than **FIO** since **Proposed** imposed a stronger fairness condition. However, with **Proposed**, both (i) and (iii) dropped to nearly zero, while only (i) decreased with **FIO**. This implies that while it remains uncertain whether the predictions of **FIO** are individually fair, the predictions of **Proposed** are more reliable since it can successfully reduce the upper bound on PIU (i.e., (iii)) and guarantee individual-level fairness, which is helpful for practitioners.

Additional experimental results: To further demonstrate the effectiveness of our method, we present several additional experimental results in Appendix I. Through our experiments, we show that our method also worked well using other classifier than the neural network (Appendix I.1), present the statistical significance of the test accuracy (Appendix I.2), confirm that both our method and **PSCF** achieved individual-level fairness when the data satisfy the functional assumptions (Appendix I.3), demonstrate the tightness of our upper bound on PIU (Appendix I.4), and evaluate the performance of our extended framework for addressing latent confounders (Appendix I.5).

6 RELATED WORK

Causality-based fairness: Motivated by recent developments in inferring causal graphs (Chikahara and Fujino, 2018; Glymour et al., 2019), many causality-based approaches to fairness have been proposed (Chiappa and Gillam, 2019; Kilbertus et al., 2017; Kusner et al., 2017, 2019; Nabi and Shpitser, 2018; Nabi et al., 2019; Russell et al., 2017; Salimi et al., 2019; Wu et al., 2018, 2019a; Xu et al., 2019; Zhang et al., 2017; Zhang and Wu, 2017; Zhang et al., 2018; Zhang and Bareinboim, 2018a,b). However, few are designed for making individually fair predictions due to the difficulty of estimating the conditional mean unfair effect in (5). Although *path-specific counterfactual fairness* (PC-fairness) (Wu et al., 2019b) was proposed to estimate its lower and upper bounds, it only addresses measuring unfairness in data and not making fair predictions.

By contrast, we established a learning framework for making individually fair predictions under much weaker assumptions than the existing approaches.

Bounding PIU: To learn an individually fair classifier, our framework utilizes the upper bound on PIU in (7). Compared with the existing bounds described below, this upper bound has the following two advantages.

First, it can be used for binary potential outcomes. If we consider continuous potential outcomes, we can use several bounds on a functional of the joint distribution of potential outcomes (Fan et al., 2017; Firpo and Ridder, 2019) because PIU is also such a functional. However, these existing bounds cannot be used in our binary classification setting where PIU is formulated based on binary potential outcomes.

Second, it is much tighter than the result in Rubinstein and Singla (2017), which provides an upper bound on an expectation of random variables whose joint distribution is unavailable.[3] Since PIU can be written as an expectation (i.e., $\mathbb{E}_{Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}}[\mathbf{1}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})]$), we can apply this result to PIU; however, the bound becomes much looser than ours. Although the multiplicative constant in (7) is 2, this value becomes 200 with the bound of Rubinstein and Singla (2017). If we use such a loose upper bound, we need to impose an excessively severe penalty on it to ensure that PIU is close to zero. By contrast, using our upper bound, we can avoid imposing such a severe penalty, thus preventing an unnecessary decrease in prediction accuracy.

7 CONCLUSION

We proposed a learning framework for guaranteeing individual-level fairness without impractical functional assumptions. Based on the concept called path-specific causal effects, we defined PIU and derived its upper bound that can be estimated from data. By forcing this upper bound value to be nearly zero, our proposed method trains an individually fair classifier. We experimentally show that our method makes individually fairer predictions than the existing methods at a slight cost of accuracy, indicating that it strikes a better balance between fairness and accuracy.

References

Shipra Agrawal, Yichuan Ding, Amin Saberi, and Yinyu Ye. Correlation robust stochastic optimization. In *SODA*, pages 1087–1096, 2010.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3]Such an upper bound is known as *correlation gap* in the field of robust optimization. (Agrawal et al., 2010).

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *IJCAI*, pages 357–363, 2005.

K. Bache and M. Lichman. UCI machine learning repository: Datasets. http://archive.ics.uci.edu/ml/datasets, 2013.

James V Burke, Adrian S Lewis, and Michael L Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

Silvia Chiappa and Thomas PS Gillam. Path-specific counterfactual fairness. In *AAAI*, pages 7801–7808, 2019.

Yoichi Chikahara and Akinori Fujino. Causal inference in time series via supervised learning. In *IJCAI*, pages 2042–2048, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS*, pages 214–226, 2012.

Yanqin Fan, Emmanuel Guerre, and Dongming Zhu. Partial identification of functionals of the joint distribution of "potential outcomes". *Journal of Econometrics*, 197(1):42–59, 2017.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268, 2015.

Juan Ferrera. *An introduction to nonsmooth analysis*. Academic Press, 2013.

Sergio Firpo and Geert Ridder. Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, 213(1):210–234, 2019.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *NeurIPS*, pages 3315–3323, 2016.

Kimberly A Houser. Can AI solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making. *Stan. Tech. L. Rev.*, 22:290, 2019.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *NeurIPS*, pages 689–696, 2009.

Martin Huber. Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29(6):920–943, 2014.

Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning

algorithms. *Journal of Banking and Finance*, 34(11): 2767–2787, 2010.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *NeurIPS*, pages 656–666, 2017.

Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. Making decisions that reduce discriminatory impacts. In *ICML*, pages 3591–3600, 2019.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NeurIPS*, pages 4066–4076, 2017.

John Langford and Robert Schapire. Tutorial on practical prediction theory for classification. *JMLR*, 6 (10):273–306, 2005.

Caleb H Miles, Phyllis Kanki, Seema Meloni, and Eric J. Tchetgen Tchetgen. On partial identification of the natural indirect effect. *Journal of Causal Inference*, 5(2), 2017.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *AAAI*, pages 1931–1940, 2018. https://github.com/raziehna/fair-inference-on-outcomes.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *ICML*, pages 4674–4682, 2019.

Judea Pearl. Direct and indirect effects. In *UAI*, pages 411–420, 2001.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Aviad Rubinstein and Sahil Singla. Combinatorial prophet inequalities. In *SODA*, pages 1671–1687, 2017.

Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *NeurIPS*, pages 6414–6423, 2017.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *SIGMOD*, pages 793–810, 2019.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7(Oct):2003–2030, 2006.

Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *NeurIPS*, pages 1687–1695, 2010.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pages 1139–1147, 2013.

Yongkai Wu, Lu Zhang, and Xintao Wu. On discrimination discovery and removal in ranked data using causal graph. In *KDD*, pages 2536–2544, 2018.

Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *IJCAI*, 2019a.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. PC-fairness: A unified framework for measuring causality-based fairness. In *NeurIPS*, pages 3399–3409, 2019b.

Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *IJCAI*, pages 1452–1458, 2019.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *NeurIPS*, 2018a.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making–the causal explanation formula. In *AAAI*, 2018b.

K Zhang and A Hyvärinen. On the identifiability of the post-nonlinear causal model. In *UAI*, pages 647–655, 2009.

Lu Zhang and Xintao Wu. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics*, 4(1):1–16, 2017.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI*, 2017.

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.

Appendix

A PATH-SPECIFIC CAUSAL EFFECTS

In this section, we provide a formal definition of path-specific causal effects (Avin et al., 2005). We begin by reviewing a formulation of path-specific causal effects in Appendix A.1 and then formally define SEM and the concept of *intervention* (Pearl, 2009) in Appendix A.2, both of which are needed to define path-specific causal effects. Finally, we define path-specific causal effects in Appendix A.3.

A.1 Revisiting Example

To take an example of path-specific causal effects, we revisit a scenario of hiring decisions for physically demanding jobs with the causal graph in Figure 1(b).

In this scenario, as described in Section 2.2, the path-specific causal effect on prediction Y is formulated as the difference $Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0}$, where $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ are the following potential outcomes:

$$Y_{A\Leftarrow 0} = h_\theta(0, Q, D(0), M(0)), \quad Y_{A\Leftarrow 1\|\pi} = h_\theta(1, Q, D(1), M(0)). \tag{4}$$

Here $D(a)$ and $M(a)$ ($a \in \{0,1\}$) represent the (counterfactual) attributes of number of children D and physical strength M. Using structural equations (2), these attributes are formulated as

$$D(a) = a + U_D Q, \quad M(a) = 3a + 0.5Q + U_M. \tag{3}$$

To illustrate these attributes, suppose that a woman $i \in \{1, \dots, n\}$ has the attributes $\boldsymbol{x}_i = \{A = 0, Q = q_i, D = d_i, M = m_i\}$. For her, attributes $D(0)$ and $M(0)$ are observed as d_i and m_i, respectively; that is,

$$D(0) = d_i = u_{D,i} q_i, \quad M(0) = m_i = 0.5 q_i + u_{M,i}$$

where $u_{D,i}$ and $u_{M,i}$ are the values of unobserved noises U_D and U_M for woman i, respectively. Meanwhile, $D(1)$ and $M(1)$ express the counterfactual attributes of D and M that would be if she were male, respectively, which are formulated as

$$D(1) = 1 + u_{D,i} q_i, \quad M(1) = 3 + 0.5 q_i + u_{M,i}.$$

Therefore, for her, potential outcome $Y_{A\Leftarrow 0}$ in (4) is provided by directly using her attributes $A = 0$, $Q = q_i$, $D(0) = d_i$, and $M(0) = m_i$. Meanwhile, $Y_{A\Leftarrow 1\|\pi}$ is obtained using her attributes $Q = q_i$ and $M(0) = m_i$ and counterfactual attributes $A = 1$ and $D(1) = 1 + u_{D,i} q_i$.

To take another example, suppose that a man $j \in \{1, \dots, n\}$ ($j \neq i$) has the attributes $\boldsymbol{x}_j = \{A = 1, Q = q_j, D = d_j, M = m_j\}$. For him, attributes $D(1)$ and $M(1)$ are observed as d_j and m_j, respectively; in other words,

$$D(1) = d_j = 1 + u_{D,j} q_j, \quad M(1) = m_j = 3 + 0.5 q_j + u_{M,j}.$$

In contrast, $D(0)$ and $M(0)$ represent the counterfactual attributes of D and M that would be if he were female, respectively, which are expressed as

$$D(0) = u_{D,j} q_j. \quad M(0) = m_j = 0.5 q_j + u_{M,j}.$$

Consequently, for this man, potential outcome $Y_{A\Leftarrow 0}$ in (4) is provided by using his attribute $Q = q_j$ and counterfactual attributes $A = 0$, $D(0) = u_{D,j} q_j$, and $M(0) = 0.5 q_j + u_{M,j}$. By contrast, $Y_{A\Leftarrow 1\|\pi}$ is obtained by using his attributes $A = 1$, $Q = q_j$, $D(1) = d_j$ and counterfactual attribute $M(0) = 0.5 q_j + u_{M,j}$.

Without an SEM, since unobserved noise values (i.e., $u_{D,i}$, $u_{D,j}$, $u_{M,i}$ and $u_{M,j}$) are unavailable, we cannot obtain counterfactual attributes $D(1)$ and $M(1)$ for woman i and $D(0)$ and $M(0)$ for man j. However, if the SEM is available, we can compute them by sampling these noises from their distribution $\mathrm{P}(\boldsymbol{U})$ and computing the counterfactual attributes based on (3). In the next section, we formally define an SEM and formulate these counterfactual attributes.

A.2 Concepts of Causal Inference

A.2.1 SEM

SEM $\mathcal{M}$ is formally defined as quadruplet $\mathcal{M} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, \mathrm{P}(\boldsymbol{U}))$, where $\boldsymbol{U}$ is a set of unobserved noise variables called *exogenous variables*, $\boldsymbol{V}$ is a set of observed variables called *endogenous variables*, which are represented in a causal graph, $\boldsymbol{F}$ is a set of deterministic functions, and $\mathrm{P}(\boldsymbol{U})$ is the joint distribution over $\boldsymbol{U}$ (Pearl, 2009).

For each variable $V \in \boldsymbol{V}$, the SEM describes how its variable value is determined, formulated as a structural equation:

$$V = f_V(\boldsymbol{pa}(V), \boldsymbol{U}_V), \tag{A1}$$

where $f_V \in \boldsymbol{F}$ is a deterministic function, $\boldsymbol{pa}(V) \subseteq \boldsymbol{V} \backslash V$ are the variables that are the parents of V in the causal graph, and $\boldsymbol{U}_V \subseteq \boldsymbol{U}$.

In our setting, we consider SEM $\mathcal{M}^p$, whose endogenous variables $\boldsymbol{V} = \{\boldsymbol{X}, Y\}$ contain prediction Y. In this SEM, each input feature variable $V \in \boldsymbol{X}$ is expressed by structural equation (A1). By contrast, as described in Section 2.2, the structural equation over prediction Y is expressed by classifier h_θ. If it is deterministic, the structural equation is given by

$$Y = h_\theta(\boldsymbol{X}),$$

and if h_θ is a probabilistic classifier, it is expressed by

$$Y = h_\theta(\boldsymbol{X}, \boldsymbol{U}_Y),$$

where $\boldsymbol{U}_Y \subseteq \boldsymbol{U}$ denotes the unobserved random noises used in the classifier. These formulations of the structural equation over prediction Y can be regarded as a special case of (A1), where deterministic function $f_Y \in \boldsymbol{F}$ is replaced with classifier h_θ.

Note that this SEM differs from an SEM in the standard setting of causal inference, which expresses a true generating process of the observed data. While the former SEM includes prediction Y, the latter contains observed decision outcome Y, whose observations are included in the data and represented by a structural equation (A1).

A.2.2 Interventions

To define potential outcomes $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$, we need to formulate the counterfactual attributes of each individual that would be if sensitive feature attribute were changed (e.g., $D(a)$ and $M(a)$ in (3)), which requires an operation for SEM $\mathcal{M}^p$, called an intervention (Pearl, 2009).

To express the counterfactual situations where sensitive feature attribute is changed, we consider intervention $do(A = a)$ on sensitive feature A, which forces A to take a certain value, $a \in \{0, 1\}$. This intervention is formally defined as a replacement of the structural equation over A with $A = a$.

Such a replacement modifies the structural equations over the descendant variables of A, and the counterfactual attributes can be formulated by this modified SEM. For instance, when the causal graph in Figure 1(b) is given, intervention $do(A = a)$ modifies the structural equations over D and M as (3), which provides counterfactual attributes $D(a)$ and $M(a)$.

In general, an SEM modified by an intervention is called an *interventional SEM*. In what follows, let $\mathcal{M}^p_{A=a}$ denote an interventional SEM that is obtained by performing intervention $do(A = a)$.

A.3 Defining Path-Specific Causal Effects

Using interventional SEM $\mathcal{M}^p_{A=a}$ ($a \in \{0, 1\}$) defined in Appendix A.2.2, we define path-specific causal effects (see the original paper (Avin et al., 2005) for details). As already described in Section 2.2, a path-specific causal effect is formulated by the difference between two potential outcomes, $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$, as $Y_{A \Leftarrow 1 \| \pi} - Y_{A \Leftarrow 0}$. In what follows, we formally define these potential outcomes.

Definition: To define potential outcome $Y_{A\Leftarrow 0}$, we consider interventional SEM $\mathcal{M}^p_{A=0}$, which is obtained by simply performing intervention $do(A=0)$. Suppose that this SEM expresses each variable $V \in \{\boldsymbol{X}, Y\}$ by the following structural equation:

$$V = f_V(\boldsymbol{pa}(V)_{A=0}, \boldsymbol{U}_V), \tag{A2}$$

where $\boldsymbol{pa}(V)_{A=0}$ denotes variables $\boldsymbol{pa}(V)$ (i.e., parents of variable V), whose values are determined by interventional SEM $\mathcal{M}^p_{A=0}$. Then potential outcome $Y_{A\Leftarrow 0}$ is defined as prediction Y, whose structural equation is expressed by (A2) where function f_Y is given by classifier h_θ.

By contrast, to define potential outcome $Y_{A\Leftarrow 1\|\pi}$, we need an SEM that is modified using interventional SEMs $\mathcal{M}^p_{A=0}$ and $\mathcal{M}^p_{A=1}$. To formulate this modified SEM, for each variable $V \in \{\boldsymbol{X}, Y\}$, we partition its parents $\boldsymbol{pa}(V)$ into two subsets, $\boldsymbol{pa}(V) = \{\boldsymbol{pa}(V)^\pi, \boldsymbol{pa}(V)^{\overline{\pi}}\}$, where $\boldsymbol{pa}(V)^\pi$ is the members of $\boldsymbol{pa}(V)$ connected with V on unfair pathways π, and $\boldsymbol{pa}(V)^{\overline{\pi}}$ is a complementary set (i.e., $\boldsymbol{pa}(V)^{\overline{\pi}} = \boldsymbol{pa}(V) \backslash \boldsymbol{pa}(V)^\pi$). Based on these two subsets, we consider the following structural equation over $V \in \{\boldsymbol{X}, Y\}$:

$$V = f_V(\boldsymbol{pa}(V)^\pi_{A=1}, \boldsymbol{pa}(V)^{\overline{\pi}}_{A=0}, \boldsymbol{U}_V), \tag{A3}$$

where $\boldsymbol{pa}(V)^\pi_{A=1}$ is a set of the variables in $\boldsymbol{pa}(V)^\pi$ whose values are determined by interventional model $\mathcal{M}^p_{A=1}$, and $\boldsymbol{pa}(V)^{\overline{\pi}}_{A=0}$ is a set of the variables in $\boldsymbol{pa}(V)^{\overline{\pi}}$ whose values are provided by $\mathcal{M}^p_{A=0}$. Then potential outcome $Y_{A\Leftarrow 1\|\pi}$ is defined as prediction Y, whose structural equation is represented by (A3).

In this way, $Y_{A\Leftarrow 1\|\pi}$ is formulated by performing intervention $do(A=1)$ only on the variables that involve unfair pathways π (i.e., $\boldsymbol{pa}(V)^\pi$), and by taking the difference between $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$, we can measure the influence via the pathways π.

Example: We provide a simple formulation example of potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ based on the causal graph in Figure 1 (a).

Since the parents of prediction Y in this causal graph are $\boldsymbol{pa}(Y) = \boldsymbol{X} = \{A, Q, D, M\}$, by simply performing intervention $do(A=0)$ on these variables, potential outcome $Y_{A\Leftarrow 0}$ is formulated as

$$Y_{A\Leftarrow 0} = h_\theta(0, Q, D(0), M(0), \boldsymbol{U}_Y), \tag{A4}$$

where classifier inputs $\{0, Q, D(0), M(0)\}$ correspond to $\boldsymbol{pa}(Y)_{A=0}$, whose values are given by interventional SEM $\mathcal{M}^p_{A=0}$.

Potential outcome $Y_{A\Leftarrow 1\|\pi}$ is defined based on unfair pathway $\pi = \{A \to Y\}$. Since prediction Y is not connected with Q, D or M on unfair pathway π but connected with A, the parents of prediction Y (i.e., $\boldsymbol{pa}(Y) = \boldsymbol{X}$) is partitioned into two subsets, $\boldsymbol{pa}(Y)^{\overline{\pi}} = \{Q, D, M\}$ and $\boldsymbol{pa}(Y)^\pi = \{A\}$. Letting the values of these variable subsets be determined by interventional models $\mathcal{M}^p_{A=0}$ and $\mathcal{M}^p_{A=1}$, respectively, potential outcome $Y_{A\Leftarrow 1\|\pi}$ is expressed as

$$Y_{A\Leftarrow 1\|\pi} = h_\theta(1, Q, D(0), M(0), \boldsymbol{U}_Y), \tag{A5}$$

where classifier inputs $\{Q, D(0), M(0)\}$ and $\{1\}$ correspond to $\boldsymbol{pa}(Y)^{\overline{\pi}}_{A=0}$ and $\boldsymbol{pa}(Y)^\pi_{A=1}$, respectively.

Given two potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$, by taking their difference (i.e., $Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0}$), we can measure the influence via direct pathway $\pi = \{A \to Y\}$, which is called *natural direct causal effects* (Pearl, 2001).

Note that we can completely remove this direct causal effect by making a prediction without sensitive feature A (Kusner et al., 2017, Section S4). However, if we consider multiple unfair pathways, we need to remove all the features that are descendants of A in the causal graph, which may seriously decrease the prediction accuracy, as described in Section 2.1.

B ASSUMPTIONS

To estimate the marginal probabilities of potential outcomes $\mathrm{P}(Y_{A\Leftarrow 0} = 1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$, our method uses two standard assumptions, both of which are widely used in the existing methods (Chiappa and Gillam, 2019; Nabi and Shpitser, 2018; Zhang and Wu, 2017; Zhang et al., 2017).

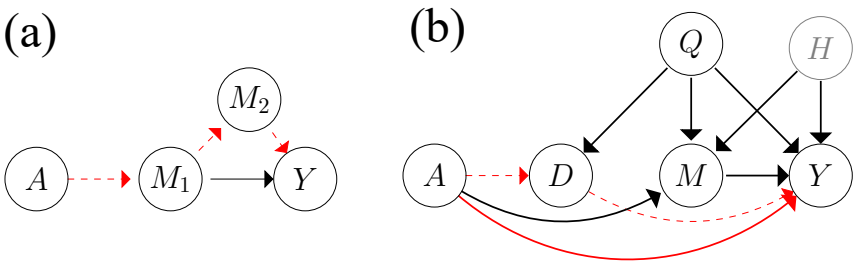

Figure A1: Two causal graphs that violate Assumptions 1 and 2, respectively. Unfair pathways are (a): $\pi = \{A \to M_1 \to M_2 \to Y\}$ and (b): $\pi = \{A \to Y, A \to D \to Y\}$.

One is an assumption on unfair pathways π, which is expressed using the following graphical condition called the *recanting witness criterion*:

Definition 3 (Recanting witness criterion (Avin et al., 2005)) *Given pathways π, let Z be a node in the causal graph that satisfies the following:*

1. *There is a pathway from A to Z ($A \to \cdots \to Z$) in π.*
2. *There is a pathway from Z to Y ($Z \to \cdots \to Y$) in π.*
3. *There is another pathway from Z to Y ($Z \to \cdots \to Y$) that is in the causal graph but not in π.*

Then pathways π satisfy the recanting witness criterion with node Z, which is called a witness.

For example, consider the causal graph in Figure A1(a), where the unfair pathway is $\pi = \{A \to M_1 \to M_2 \to Y\}$. Clearly, pathway π satisfies the recanting witness criterion with witness M_1.

According to Avin et al. (2005), to estimate marginal probabilities $\mathrm{P}(Y_{A \Leftarrow 0} = 1)$ and $\mathrm{P}(Y_{A \Leftarrow 1 \| \pi} = 1)$, we need to assume that pathways π do **not** satisfy the recanting witness criterion in Definition 3; that is,

Assumption 1 *Pathways π do* **not** *satisfy the recanting witness condition.*

The other is a common assumption in causal inference called *conditional ignorability* (Rosenbaum and Rubin, 1983), which requires conditional independence relations between variables.

We formulate this assumption based on the estimators in a previous work (Huber, 2014), which we use in our method as presented in (9). As an example, we show a formulation based on the causal graph in Figure 1(b).

In what follows, we use the same notations as those in the original paper (Huber, 2014). Let potential outcomes $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$ denote

$$Y_{A \Leftarrow 0} = Y(0, D(0), M(0)), \quad Y_{A \Leftarrow 1 \| \pi} = Y(1, D(1), M(0)), \tag{A6}$$

respectively, where $D(0)$, $D(1)$, and $M(0)$ express counterfactual attributes formulated by (3). Then the conditional ignorability is expressed as follows:

Assumption 2 (Conditional ignorability (Huber, 2014)) *For all $a, a', a'' \in \{0, 1\}$ and d, m, q in the supports of D, M, and Q, the following four relations hold:*

$$\{Y(a, d, m), D(a'), M(a'')\} \perp\!\!\!\perp A | Q = q \tag{A7}$$
$$Y(a', d, m) \perp\!\!\!\perp D | A = a, Q = q \tag{A8}$$
$$Y(a', d, m) \perp\!\!\!\perp M | A = a, Q = q \tag{A9}$$
$$\mathrm{P}(A = a | Q = q, D = d, M = m) > 0. \tag{A10}$$

In Assumption 2, three relations, (A7), (A8), and (A9), are needed to express the potential outcome using the observed data distribution. As mentioned by Huber (2014), these relations are not satisfied if there is an

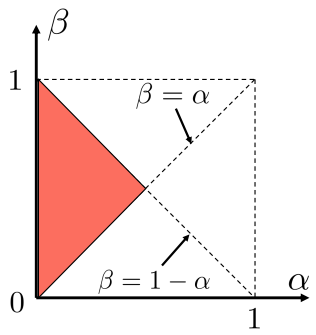

Figure A2: Red area represents region where marginal probability values α and β satisfy $\alpha \leq \beta \leq 1 - \alpha$

unobserved variable called a latent confounder, which is an unobserved parent of the observed variables. For instance, when the causal graph in Figure A1(b) is given, the aforementioned relations do not hold due to a latent confounder, H (gray node). However, even in the presence of latent confounders, in some cases, our method can achieve individual-level fairness using an extended penalty function, as described in Section 4.5.

The last relation (A10) is used to avoid a division by zero when computing the estimators (9).

C PROOF OF THEOREM 1

We first introduce several notations. Let the marginal potential outcome probabilities that satisfy $Y_{A \Leftarrow 0} = 1$ and $Y_{A \Leftarrow 1 \| \pi} = 1$ be α and β, and their joint probabilities, $(Y_{A \Leftarrow 0}, Y_{A \Leftarrow 1 \| \pi}) = (0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$, be p_{00}, p_{01}, p_{10}, and p_{11}.

Then we have

$$p_{10} + p_{11} = \alpha, \quad p_{00} + p_{01} = 1 - \alpha, \quad p_{01} + p_{11} = \beta, \quad \text{and} \quad p_{10} + p_{00} = 1 - \beta. \tag{A11}$$

As described in Section 4.3.1, the right-hand side in (7) in Theorem 1 can be represented by marginal probabilities. With the above notations, it can be written as $2(\beta(1 - \alpha) + \alpha(1 - \beta))$.

Therefore, our goal is to prove

$$p_{01} + p_{10} \leq 2(\beta(1 - \alpha) + \alpha(1 - \beta)).$$

Since all the joint probabilities in (A11) are non-negative, p_{01} and p_{10} become at most $\min\{\beta, 1 - \alpha\}$ and $\min\{\alpha, 1 - \beta\}$, respectively, yielding:

$$p_{01} + p_{10} \leq \min\{\beta, 1 - \alpha\} + \min\{\alpha, 1 - \beta\}. \tag{A12}$$

Hence, it suffices to prove

$$\min\{\beta, 1 - \alpha\} + \min\{\alpha, 1 - \beta\} \leq 2\beta(1 - \alpha) + 2\alpha(1 - \beta). \tag{A13}$$

Since both sides in (A13) are symmetrical with respect to $\beta = \alpha$ and $\beta = 1 - \alpha$, it is sufficient to consider the case when $\alpha \leq \beta \leq 1 - \alpha$, which is illustrated in Figure A2 as the red region.

In this case, since $\min\{\beta, 1 - \alpha\} = \beta$ and $\min\{\alpha, 1 - \beta\} = \alpha$, (A13) is reduced to

$$\beta + \alpha \leq 2\beta(1 - \alpha) + 2\alpha(1 - \beta) \quad \alpha + \beta - 4\alpha\beta \geq 0. \tag{A14}$$

Since $\alpha + \beta \leq 1$ holds in this case, we have inequality $\alpha + \beta - (\alpha + \beta)^2 \geq 0$. Using this inequality, inequality (A14) can be proven as follows:

$$\alpha + \beta - 4\alpha\beta = \alpha + \beta - (\alpha + \beta)^2 + (\alpha - \beta)^2 \geq 0. \tag{A15}$$

Thus, we prove Theorem 1.

D DERIVATION OF (9)

Following the original paper (Huber, 2014), we derived the following formulation of the existing estimators of marginal potential outcome probabilities:

$$\hat{p}_\theta^{A\Leftarrow 0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(a_i = 0)\hat{w}_i c_\theta(a_i, q_i, d_i, m_i), \quad \hat{p}_\theta^{A\Leftarrow 1\|\pi} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(a_i = 1)\hat{w}'_i c_\theta(a_i, q_i, d_i, m_i), \tag{9}$$

where $c_\theta(\boldsymbol{X}) = \mathrm{P}(Y = 1|\boldsymbol{X})$ is the conditional distribution given by classifier h_θ, $\mathbf{1}(\cdot)$ is an indicator function, and $\hat{w}_i$ and $\hat{w}'_i$ are the following weights:

$$\hat{w}_i = \frac{1}{\hat{\mathrm{P}}(A = 0|q_i)}, \quad \hat{w}'_i = \frac{\hat{\mathrm{P}}(A = 1|q_i, d_i)\hat{\mathrm{P}}(A = 0|q_i, d_i, m_i)}{\hat{\mathrm{P}}(A = 1|q_i)\hat{\mathrm{P}}(A = 0|q_i, d_i)\hat{\mathrm{P}}(A = 1|q_i, d_i, m_i)},$$

where $\hat{\mathrm{P}}$ is the conditional distribution that is estimated by learning the statistical models (e.g., neural networks) to the training data beforehand.

Following the notations in the original paper (Huber, 2014), let potential outcomes denote $Y_{A\Leftarrow 0} = Y(0, D(0), M(0))$ and $Y_{A\Leftarrow 1\|\pi} = Y(1, D(1), M(0))$, respectively.

Then given the causal graph in Figure 1(b), marginal probability $\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$ can be written as

$$\begin{aligned}
&\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1) \\
&= \mathrm{P}(Y(1, D(1), M(0)) = 1) \\
&= \mathbb{E}_Q[\mathbb{E}_{D(1)|Q}[\mathbb{E}_{M(0)|Q,D(1)}[\mathrm{P}(Y(1, d, m) = 1|A = 1, Q = q, D(1) = d, M(0) = m)]]].
\end{aligned}$$

Using Assumption 2 in Appendix B, this can be rewritten as

$$\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1) = \mathbb{E}_Q[\mathbb{E}_{D|A=1,Q}[\mathbb{E}_{M|A=0,Q,D}[\mathrm{P}(Y(1, d, m) = 1|A = 1, Q = q, D = d, M = m)]]],$$

With Bayes' theorem, this can be expressed as

$$\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1) = \mathbb{E}_Q[\mathbb{E}_{D|Q}[\mathbb{E}_{M|Q,D}[\omega' \mathrm{P}(Y = 1|dA = 1, Q = q, D = d, M = m)]]],$$

where ω' is expressed as follows:

$$\omega' = \frac{\mathrm{P}(A = 1|Q = q, D = d)\,\mathrm{P}(A = 0|Q = q, D = d, M = m)}{\mathrm{P}(A = 1|Q = q)\,\mathrm{P}(A = 0|Q = q, D = d)}.$$

With indicator function $\mathbf{1}(\cdot)$, this can be formulated as

$$\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1) = \mathbb{E}[\mathbf{1}(A = 1)w' \mathrm{P}(Y = 1|A = 1, q, d, m)], \tag{A16}$$

where weight w' is expressed as

$$w' = \frac{1}{\mathrm{P}(A = 1|Q = q, D = d, M = m)}\omega'.$$

In a similar manner, marginal probability $\mathrm{P}(Y_{A\Leftarrow 0} = 1)$ can be represented as

$$\mathrm{P}(Y_{A\Leftarrow 0} = 1) = \mathbb{E}[\mathbf{1}(A = 0)w\,\mathrm{P}(Y = 1|A = 0, q, d, m)], \tag{A17}$$

where weight w is formulated as

$$w = \frac{1}{\mathrm{P}(A = 0|Q = q)}.$$

Given empirical distribution, by plugging conditional distribution c_θ into $\mathrm{P}(Y = 1|A = 1, Q = q, D = d, M = m)$, we can estimate (A17) and (A16) as (9) and derive the estimators (9).

E COMPUTATION TIME AND CONVERGENCE GUARANTEE

To minimize the objective function (1), we use the stochastic gradient descent method (Sutskever et al., 2013).

With this method, we computed the penalty term and its gradient over the samples in each mini-batch. The computation time, which is required to evaluate the objective function (1) and its gradient is as much as the time needed to evaluate the training loss in (1) and its gradient, respectively.

Whether we can guarantee that the gradient descent method converges depends on the choice of classifier h_θ. For instance, if we choose a neural network classifier, we cannot guarantee that the stochastic gradient descent method (Sutskever et al., 2013) converges because the objective function (1) becomes nonconvex, and its gradient does not become *Lipschitz continuous*; that is, the maximum rate of change in the gradient is not bounded. However, if the neural network only contains activation functions whose gradients are Lipschitz continuous (e.g., the sigmoid function), we can optimize the objective function with convergence guarantees using e.g., the gradient sampling method (Burke et al., 2005), because in this case, the gradient of the objective function becomes *locally Lipschitz continuous* (Ferrera, 2013, Chapter 2).

F PIU VALUES THAT SATISFY FAIRNESS CONDITIONS

As described in Section 4.4, our method effectively makes potential outcomes take the same value (i.e., $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi}$) while FIO does not. To illustrate this, in this section, we compare the possible PIU values that satisfy the fairness condition of each method.

F.1 Comparing Possible PIU Values

We first introduce several notations. Let the (true) marginal probabilities of potential outcomes be $\alpha = \mathrm{P}(Y_{A\Leftarrow 0} = 1)$ and $\beta = \mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$, and the (true) joint probabilities of $(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}) = (0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ be p_{00}, p_{01}, p_{10}, and p_{11}.

With these notations, PIU can be formulated as $p_{01} + p_{10}$, and its lower and upper bound can be expressed using marginal probabilities α and β as

$$|\alpha - \beta| \leq p_{01} + p_{10} \leq \min\{\beta, 1-\alpha\} + \min\{\alpha, 1-\beta\}, \tag{A18}$$

which we prove in Appendix F.2.

With our method, as presented in Figure 2, the marginal probabilities are forced to be $(\alpha, \beta) \approx (0,0)$ or $(1,1)$. If α and β satisfy this condition, since both lower and upper bounds in (A18) become close to zero, PIU is constrained to almost zero (i.e., $p_{01} + p_{10} \approx 0$).

By contrast, as described in Section 4.4, FIO always accepts the marginal probabilities $(\alpha, \beta) = (0.5, 0.5)$. At this point, the lower and upper bounds in (A18) become 0 and 1: $0 \leq p_{01} + p_{10} \leq 1$. This implies that it is completely unknown whether the PIU value is high since the joint probabilities are unknown in practice. Therefore, FIO cannot ensure that the potential outcomes take the same value for all individuals, which is insufficient to guarantee individual-level fairness.

To support the above discussion on possible PIU values, in what follows, we prove the lower and upper bound on PIU (i.e., (A18)).

F.2 Proof of (A19)

Since we already proved the upper bound in (A12), below we derive the lower bound in (A18). Since α and β are marginal probabilities, we have

$$p_{10} + p_{11} = \alpha, \quad p_{01} + p_{11} = \beta,$$

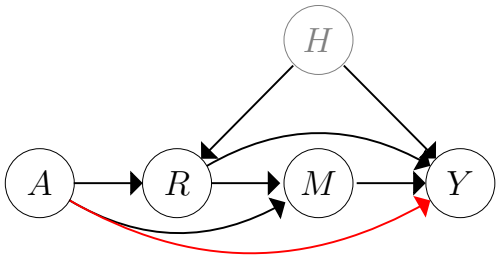

Figure A3: Example of causal graph containing latent confounder H (gray node), which affects both R and Y. Red pathway represents unfair pathway π.

which are equivalent to

$$p_{10} = \alpha - p_{11}, \quad p_{01} = \beta - p_{11},$$

respectively. By summing up both, we have

$$p_{01} + p_{10} = \alpha + \beta - 2p_{11}.$$

Since joint probability p_{11} is less than marginal probabilities α and β, we have $p_{11} \leq \min\{\alpha, \beta\}$. Therefore,

$$p_{01} + p_{10} \geq \alpha + \beta - 2\min\{\alpha, \beta\} = |\alpha - \beta|. \tag{A19}$$

Combined with the upper bound on $p_{01} + p_{10}$ in (A12), we prove (A18).

G ADDRESSING LATENT CONFOUNDERS

This section details how our method can be extended to ensure individual-level fairness when there are unobserved variables called latent confounders.

As described in Section 4.5, although it is extremely challenging to estimate the marginal potential outcome probabilities in this case, we can sometimes guarantee individual-level fairness using lower and upper bounds on them. If these lower and upper bounds are given, we can achieve this by reformulating the penalty function as follows:

$$G_\theta(\boldsymbol{x}_1, \dots \boldsymbol{x}_n) = \hat{u}_\theta^{A \Leftarrow 1 \| \pi}(1 - \hat{l}_\theta^{A \Leftarrow 0}) + (1 - \hat{l}_\theta^{A \Leftarrow 1 \| \pi})\hat{u}_\theta^{A \Leftarrow 0}, \tag{13}$$

where $\hat{l}_\theta^{A \Leftarrow 0}$ and $\hat{u}_\theta^{A \Leftarrow 0}$ are the estimated lower and upper bounds on marginal probability $\mathrm{P}(Y_{A \Leftarrow 0} = 1)$, respectively, and $\hat{l}_\theta^{A \Leftarrow 1 \| \pi}$ and $\hat{u}_\theta^{A \Leftarrow 1 \| \pi}$ are the estimated lower and upper bounds on marginal probability $\mathrm{P}(Y_{A \Leftarrow 1 \| \pi} = 1)$, respectively.

In general, we cannot estimate lower and upper bounds $\hat{l}_\theta^{A \Leftarrow 0}$, $\hat{u}_\theta^{A \Leftarrow 0}$, $\hat{u}_\theta^{A \Leftarrow 0}$, and $\hat{u}_\theta^{A \Leftarrow 1 \| \pi}$. However, when a certain causal graph is given, we can estimate them from data. In what follows, as an example, we detail the existing estimators from a previous work (Miles et al., 2017).

G.1 Example of Existing Estimators

According to a previous result (Miles et al., 2017), we can estimate the lower and upper bounds on marginal potential outcome probabilities when the causal graph in Figure A3 is given. Here $A \in \{0, 1\}$ and $Y \in \{0, 1\}$ are binary random variables, M is a discrete one, R can be either a continuous or a discrete one, and H is a latent confounder that affects two variables R and Y.[4] Here unfair pathway π is given as direct pathway $A \to Y$ (i.e., $\pi = \{A \to Y\}$).

[4]Regarding A, although Miles et al. (2017) dealt with a general case, where A can be continuous or discrete, we consider a binary case, i.e., $A \in \{0, 1\}$

With this causal graph, a previous work (Miles et al., 2017) expressed lower and upper bounds on $\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1)$ as

$$\hat{l}^{A\Leftarrow 1\|\pi} = \sum_m \max\{0,\ \mathrm{P}(M=m|A=0) - 1 + \sum_r \mathrm{P}(Y=1|A=1,m,r)\,\mathrm{P}(R=r|A=1)\},$$
$$\hat{u}^{A\Leftarrow 1\|\pi} = \sum_m \min\{\mathrm{P}(M=m|A=0),\ \sum_r \mathrm{P}(Y=1|A=1,m,r)\,\mathrm{P}(R=r|A=1)\}.$$

With conditional distribution $c_\theta(1,M,R)=\mathrm{P}(Y=1|A=1,M,R)$ provided by classifier h_θ, these bounds can be estimated as the following functions of classifier parameter θ:

$$\hat{l}_\theta^{A\Leftarrow 1\|\pi} = \sum_m \max\{0,\ \hat{\mathrm{P}}(M=m|A=0) - 1 + \sum_r c_\theta(1,m,r)\hat{\mathrm{P}}(R=r|A=1)\} \tag{A20}$$
$$\hat{u}_\theta^{A\Leftarrow 1\|\pi} = \sum_m \min\{\hat{\mathrm{P}}(M=m|A=0),\ \sum_r c_\theta(1,m,r)\hat{\mathrm{P}}(R=r|A=1)\}, \tag{A21}$$

respectively. Here conditional distributions $\hat{\mathrm{P}}(M=m|A=0)$ and $\hat{\mathrm{P}}(R=r|A=1)$ can be estimated by learning statistical models (e.g., logistic regression or neural networks) from the training data beforehand.

As with (A20) and (A21), we can formulate the estimated lower and upper bounds on marginal probability $\mathrm{P}(Y_{A\Leftarrow 0}=1)$ as

$$\hat{l}_\theta^{A\Leftarrow 0} = \sum_m \max\{0,\ \hat{\mathrm{P}}(M=m|A=0) - 1 + \sum_r c_\theta(0,m,r)\hat{\mathrm{P}}(R=r|A=0)\} \tag{A22}$$
$$\hat{u}_\theta^{A\Leftarrow 0} = \sum_m \min\{\hat{\mathrm{P}}(M=m|A=0),\ \sum_r c_\theta(0,m,r)\hat{\mathrm{P}}(R=r|A=0)\}, \tag{A23}$$

respectively.

Note that if we use the above lower and upper bounds, solving the optimization problem with convergence guarantees becomes complicated because these lower and upper bounds in the penalty term are not differentiable. Our future work will leverage other bounds on marginal probabilities to formulate an objective function that is differentiable and easy to optimize.

H EXPERIMENTAL SETTINGS

This section details the experimental settings presented in Section 5.

H.1 Settings of Each Method

For classifier h_θ for **Proposed**, **FIO**, **Unconstrained**, and **Remove**, we used a feed-forward neural network that consists of two linear layers with 100 and 50 hidden neurons, respectively. We used sigmoid activation functions and formulated the output layer by a log softmax function. For the loss function, we used cross-entropy loss. To train this classifier, we set the minibatch size to $1,000$ for the synthetic data and the Adult dataset. With the German credit dataset, since the number of training samples is less than $1,000$, we set it to 100. We stopped the training after $1,000$ epochs.

Unlike these methods, **PSCF** use multiple neural networks to learn a predictive model for each variable in $\{\boldsymbol{X}, Y\}$ (see, Appendix H.2.3 for the details of these predictive models). We use the same network architecture as that of the original paper (Chiappa and Gillam, 2019).

To compare the best performances, we selected the hyperparameter values of **Proposed**, **FIO**, and **PSCF**. For each method, we used a grid search with 0.05 grid size to select the value of the penalty parameter from $[0.0, 2.0]$.

H.2 Settings in Synthetic Data Experiments

H.2.1 Data

We prepared synthetic data that represent a scenario of hiring decisions for physically demanding jobs, whose causal graph is shown in Figure 1(b). We sampled gender $A \in \{0, 1\}$, qualification Q, number of children D, physical strength M, and hiring decision outcome $Y \in \{0, 1\}$ from the following SEM:

$$\begin{aligned} A &= U_A, \quad U_A \sim \mathrm{Bernoulli}(0.6), \\ Q &= \lfloor U_Q \rfloor, \quad U_Q \sim \mathcal{N}(2, 5^2), \\ D &= A + \lfloor 0.5 Q U_D \rfloor, \quad U_D \sim \mathrm{Tr}\mathcal{N}(2, 1^2, 0.1, 3.0), \\ M &= 3A + 0.4 Q U_M, \quad U_M \sim \mathrm{Tr}\mathcal{N}(3, 2^2, 0.1, 3.0), \\ Y &= h(A, Q, D, M), \end{aligned} \tag{A24}$$

where Bernoulli, $\mathcal{N}$, and $\mathrm{Tr}\mathcal{N}$ represent the Bernoulli, Gaussian, and truncated Gaussian distributions, respectively, and $\lfloor \cdot \rfloor$ is a floor function that returns an integer by removing the decimal places. To output hiring decision outcome Y, we used function h, which is a logistic regression model that provides the following conditional distribution:

$$\mathrm{P}(Y = 1 | A, Q, M) = \mathrm{Bernoulli}(\varsigma(-10 + 5A + Q + D + M)),$$

where $\varsigma(x) = 1/(1 + \exp(-x))$ is a standard sigmoid function.

Note that in (A24), the structural equations over D and M are not expressed by additive noise models (Hoyer et al., 2009) because they contain multiplicative noises U_D and U_M.

In experiments, we used $5,000$ samples to train the classifier and $1,000$ samples to test the performance.

H.2.2 Computing Unfair Effects

With such synthetic data, we computed the four statistics of unfair effects for **Proposed**, **FIO**, and **Unconstrained** as follows.

To compute (i) the mean unfair effect and (iii) the upper bound on PIU, we estimated marginal potential outcome probabilities $\hat{p}_\theta^{A \Leftarrow 0}$ and $\hat{p}_\theta^{A \Leftarrow 1 \| \pi}$ with estimators (9).

To obtain (ii) the standard deviation in the conditional mean unfair effects and (iv) the PIU value, we sampled potential outcomes $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$ based on the SEM (A24). Specifically, we sampled $Y_{A \Leftarrow 0}$ from the following (interventional) SEM:

$$\begin{aligned} A &= 0, \\ Q &= \lfloor U_Q \rfloor, \quad U_Q \sim \mathcal{N}(2, 5^2), \\ D(0) &= \lfloor 0.5 Q U_D \rfloor, \quad U_D \sim \mathrm{Tr}\mathcal{N}(2, 1^2, 0.1, 3.0), \\ M(0) &= 0.4 Q U_M, \quad U_M \sim \mathrm{Tr}\mathcal{N}(3, 2^2, 0.1, 3.0), \\ Y_{A \Leftarrow 0} &= h_\theta(0, Q, D(0), M(0)), \end{aligned}$$

where h_θ is the classifier. We sampled $Y_{A \Leftarrow 1 \| \pi}$ from

$$\begin{aligned} A &= 1, \\ Q &= \lfloor U_Q \rfloor, \quad U_Q \sim \mathcal{N}(2, 5^2), \\ D(1) &= 1 + \lfloor 0.5 Q U_D \rfloor, \quad U_D \sim \mathrm{Tr}\mathcal{N}(2, 1^2, 0.1, 3.0), \\ M(0) &= 0.4 Q U_M, \quad U_M \sim \mathrm{Tr}\mathcal{N}(3, 2^2, 0.1, 3.0), \\ Y_{A \Leftarrow 1 \| \pi} &= h_\theta(1, Q, D, M). \end{aligned}$$

Then using n pairs of these samples $\{(y_{A\Leftarrow 0,i}, y_{A\Leftarrow 1\|\pi,i})\}_{i=1}^n$, we evaluated the PIU value by

$$\hat{\mathrm{P}}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_{A\Leftarrow 0,i} \neq y_{A\Leftarrow 1\|\pi,i}),$$

where $\mathbf{1}(\cdot)$ is an indicator function that takes 1 if $y_{A\Leftarrow 0,i} \neq y_{A\Leftarrow 1\|\pi,i}$ ($i \in \{1,\dots,n\}$) and 0 otherwise.

We computed the standard deviation in conditional mean unfair effects as follows. We separated n individuals into K subgroups of individuals who have the same values of features $\boldsymbol{X}$, took the mean unfair effects over individuals in each subgroup, and computed their standard deviation. Let the individuals in the k-th subgroup ($k = 1,\dots,K$) have identical feature attributes $\boldsymbol{X} = \boldsymbol{x}^k$, where superscript k represents the k-th subgroup. Then using $\{(y_{A\Leftarrow 0,i}, y_{A\Leftarrow 1\|\pi,i})\}_{i=1}^n$, we estimated the standard deviation of the conditional mean unfair effects over the K subgroups as

$$\hat{\sigma} = \sqrt{\frac{\sum_{k=1}^{K} \hat{\mu}^k - \hat{\mu}}{K}}. \tag{A25}$$

Here $\hat{\mu}^k$ is the estimated conditional mean unfair effect in the k-th subgroup of individuals with identical attributes $\boldsymbol{X} = \boldsymbol{x}^k$, i.e.,

$$\hat{\mu}^k = \frac{1}{n^k}\sum_{i\in\{1,\dots,n\}|\boldsymbol{x}_i=\boldsymbol{x}^k} \mathbf{1}(y_{A\Leftarrow 0,i} \neq y_{A\Leftarrow 1\|\pi,i}),$$

where n^k is the number of individuals in the k-th subgroup and $\hat{\mu}$ is the mean of $\hat{\mu}^k$ over $k = 1,\dots,K$, i.e.,

$$\hat{\mu} = \frac{1}{K}\sum_{k}^{K} \hat{\mu}^k.$$

H.2.3 Unfair Effects of PSCF

With **PSCF**, we did not evaluate the two statistics of unfair effects, (i) the mean unfair effect and (iii) the upper bound on PIU, since they are not well-defined for this method.

This is because these statistics measure the unfairness of the learned predictive model of Y (i.e., classifier h_θ in our method); however, **PSCF** aims to ensure fairness using *unfair* predictive models.

To do so, **PSCF** approximates the SEM by learning predictive models of each variable in $\mathrm{P}(\boldsymbol{X}, Y)$, which is unfair due to the discriminatory bias in the observed data, and removes the unfairness by sampling the *fair* feature values based on the approximated SEM.

For instance, in the case of synthetic data experiments, **PSCF** approximates the SEM (A24) as follows. Using latent variable H_D, **PSCF** learns the predictive models of A, Q, D, M, and Y and the distribution of H_D, which are expressed as follows:

$$\begin{aligned}
&A = \mathrm{P}_\theta(A),\\
&Q = \mathrm{P}_\theta(Q),\\
&D = \mathrm{P}_\theta(D|A,Q,H_D), \quad H_D = \mathrm{P}_\theta(H_D),\\
&M = \mathrm{P}_\theta(M|A,Q),\\
&Y = \mathrm{P}_\theta(Y|A,Q,D,M),
\end{aligned} \tag{A26}$$

where P_θ denotes a (conditional) distribution, which is parameterized as a neural network. Here latent variable H_D approximates the additive noise in the structural equation over D.

By using the approximated SEM (A26), **PSCF** aims to make fair predictions as follows. For each individual $i \in \{1,\dots,n\}$ with attributes $\{a_i, q_i, d_i, m_i\}$, **PSCF** samples their *fair* attribute of D by

$$\hat{d}(0)_i \sim \mathrm{P}_\theta(D|A=0,q_i,h_{D,i}), \quad h_{D,i} \sim \mathrm{P}_\theta(H_D). \tag{A27}$$

Using this attribute, **PSCF** makes a prediction using the following Monte Carlo estimate:

$$\hat{y}_i^{PSCF} = \frac{1}{J}\sum_{j=1}^{J} \hat{y}_{i,j}^{PSCF} \quad \text{where} \quad \hat{y}_{i,j}^{PSCF} \sim \mathrm{P}_\theta(Y|A=0, q_i, d(0)_i, m_i). \tag{A28}$$

Here J is the number of Monte Carlo samples, which is set to $J = 5$ in our experiments.

According to Chiappa and Gillam (2019), if there is slight mismatch between the approximated SEM (A26) and the true SEM (A24), **PSCF** can eliminate the conditional mean unfair effect and achieve individual-level fairness. Intuitively, this is because in (A28) and (A27), A's values are fixed in the approximated structural equations over Y and D, which involve unfair pathways $\pi = \{A \to Y, A \to D \to Y\}$.

In this way, **PSCF** aims to achieve fairness by sampling the fair feature values using unfair predictive models. Therefore, we cannot measure the unfairness of this method using the two statistics (i) and (iii), which measure the unfairness of the predictive model.

In contrast, it is appropriate to measure the unfairness based on two other statistics, i.e., (ii) the standard deviation in the conditional mean unfair effects and (iv) the PIU value. Since both are formulated using the true SEM, they can be used to quantify the unfairness due to SEM's approximation error.

To compute the unfairness of **PSCF** using these two statistics, we made a prediction in the same way as (A28), except that we used the true SEM (A24). Let such predicted values be $\{y_i^{PSCF}\}_{i=1}^n$. Then using n pairs of predicted values $\{(y_i^{PSCF}, \hat{y}_i^{PSCF})\}_{i=1}^n$, we estimated (iv) the PIU value as

$$\hat{\mathrm{P}}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_i^{PSCF} \neq \hat{y}_i^{PSCF}),$$

and (ii) the standard deviation of conditional mean unfair effects in the same way as (A25) except that $\hat{\mu}^k$ is estimated as

$$\hat{\mu}^k = \frac{1}{n^k}\sum_{i\in\{1,\ldots,n\}|\boldsymbol{x}_i=\boldsymbol{x}^k} \mathbf{1}(y_i^{PSCF} \neq \hat{y}_i^{PSCF}).$$

H.3 Settings in Real-World Data Experiments

H.3.1 Data and Causal Graphs

In real-world data experiments, we used two datasets, the German credit and Adult datasets (Bache and Lichman, 2013).

The German credit dataset consists of the records of 1,000 loan applicants and includes the attributes of each individual, such as gender, amount of savings, and age.

Using this dataset, we predicted whether each loan applicant is a good or bad credit risk. We used 900 samples for training and 100 samples to test the performance.

To evaluate the unfairness of the predictions, following Chiappa and Gillam (2019), we used the causal graph in Figure 3(a). Here, A denotes gender, Y expresses credit risk, $\boldsymbol{S}$ represents financial information (i.e., amount, checking account balance, and house ownership), $\boldsymbol{R}$ stands for credit amount and repayment duration, and $\boldsymbol{C}$ corresponds to such attributes of each individual as age and loan purpose. We regarded gender A as a sensitive feature and pathways $A \to Y$ and $A \to \boldsymbol{S} \to Y$ as unfair.

On the other hand, the Adult dataset is comprised of US census data that contain the features of individuals including gender, occupation, and income.

With this dataset, we predicted whether each individual has an annual income exceeding 50,000 US dollars. We employed 34,001 samples to train our classifier and 10,870 samples to test the performance.

Following Chiappa and Gillam (2019), we used the causal graph in Fig. 3(b). We regarded gender A as a sensitive feature. Other features are marital status M, education L, occupation information $\mathbf{R}$ (e.g., weekly working hours),

age and nationality $\mathbf{C}$, and predicted income Y. Unfair pathways π are direct pathway $A \to Y$ and pathways from A to Y through M (i.e., $A \to M \to \cdots \to Y$).

H.3.2 Computing Unfair Effects

To evaluate the two statistics of unfair effects (i.e., (i) and (iii)), we computed the marginal probabilities of potential outcomes $\hat{p}_\theta^{A \Leftarrow 0}$ and $\hat{p}_\theta^{A \Leftarrow 1 \| \pi}$ based on the existing estimators (Huber, 2014).

With the German credit dataset, using the attributes of n individuals $\{a_i, \boldsymbol{c}_i, \boldsymbol{s}_i, \boldsymbol{r}_i\}_{i=1}^n$, we estimated the marginal probabilities as

$$\hat{p}_\theta^{A \Leftarrow 0} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(a_i = 0)\hat{w}_i c_\theta(a_i, \boldsymbol{c}_i, \boldsymbol{s}_i, \boldsymbol{r}_i), \quad \hat{p}_\theta^{A \Leftarrow 1 \| \pi} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(a_i = 1)\hat{w}'_i c_\theta(a_i, \boldsymbol{c}_i, \boldsymbol{s}_i, \boldsymbol{r}_i), \tag{A29}$$

respectively, where weights $\hat{w}_i$ and $\hat{w}'_i$ are expressed as

$$\hat{w}_i = \frac{1}{\hat{\mathrm{P}}(A = 0|\boldsymbol{c}_i)}, \quad \hat{w}'_i = \frac{\hat{\mathrm{P}}(A = 1|\boldsymbol{c}_i, \boldsymbol{s}_i)\hat{\mathrm{P}}(A = 0|\boldsymbol{c}_i, \boldsymbol{s}_i, \boldsymbol{c}_i)}{\hat{\mathrm{P}}(A = 1|\boldsymbol{c}_i)\hat{\mathrm{P}}(A = 0|\boldsymbol{c}_i, \boldsymbol{s}_i)\hat{\mathrm{P}}(A = 1|\boldsymbol{c}_i, \boldsymbol{s}_i, \boldsymbol{c}_i)},$$

To compute these weights, we inferred the conditional probabilities of A by fitting the logistic regression model to the training data beforehand.

For the Adult dataset, given the attributes of n individuals $\{a_i, m_i, l_i, \mathbf{r}_i, \boldsymbol{c}_i\}_{i=1}^n$, we estimated the marginal probabilities as

$$\hat{p}_\theta^{A \Leftarrow 0} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(a_i = 0)\hat{w}_i c_\theta(0, m_i, l_i, \boldsymbol{r}_i, \boldsymbol{c}_i), \quad \hat{p}_\theta^{A \Leftarrow 1 \| \pi} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(a_i = 1)\hat{w}'_i c_\theta(1, m_i, l_i, \boldsymbol{r}_i, \boldsymbol{c}_i),$$

where weights $\hat{w}_i$ and $\hat{w}'_i$ are provided by

$$\hat{w}_i = \frac{1}{\hat{\mathrm{P}}(A = 0|\boldsymbol{c}_i)}, \quad \hat{w}'_i = \frac{\hat{\mathrm{P}}(A = 1|m_i, \boldsymbol{c}_i)\hat{\mathrm{P}}(A = 0|m_i, l_i, \boldsymbol{r}_i, \boldsymbol{c}_i)}{\hat{\mathrm{P}}(A = 1|\boldsymbol{c}_i)\hat{\mathrm{P}}(A = 0|m_i, \boldsymbol{c}_i)\hat{\mathrm{P}}(A = 1|m_i, l_i, \boldsymbol{r}_i, \boldsymbol{c}_i)}.$$

To obtain these weight values, we estimated each conditional probability of A by fitting the logistic regression model to the data beforehand.

H.4 Computing Infrastructure

In our experiments, we used PyTorch 1.6.0 as an implementation of the optimization algorithm (Sutskever et al., 2013) and a 64-bit CentOS machine with 2.6GHz Xeon E5-2697A-v4 (x2) CPUs and 512-GB RAM.

I ADDITIONAL EXPERIMENTAL RESULTS

To further demonstrate the effectiveness of our proposed method, in this section, we provide several additional experimental results.

This section is organized as follows. In Appendix I.1, we show that our method works well even with a simpler classifier than the neural network used in Section 5. In Appendix I.2, we investigate the statistical significance of the test accuracy using the test set bound (Langford and Schapire, 2005). In Appendix I.3, we confirm that when the data satisfy the functional assumptions of the PSCF method (Chiappa and Gillam, 2019), both our method and PSCF achieve individual-level fairness. In Appendix I.4, we show the tightness of the upper bound on PIU by confirming that the performance does not differ so much even when having an oracle access to the true PIU value. Finally, in Appendix I.5, we evaluate the performance of our extended framework described in Appendix G.

Table 3: Test accuracy (%) on each dataset when using logistic regression model

Method	Synth	German	Adult
Proposed	78.2 ± 1.5	76.0	75.2
FIO	83.4 ± 1.2	77.5	79.0
Unconstrained	87.8 ± 0.8	78.8	79.7
Remove	76.1 ± 0.9	73.8	74.4

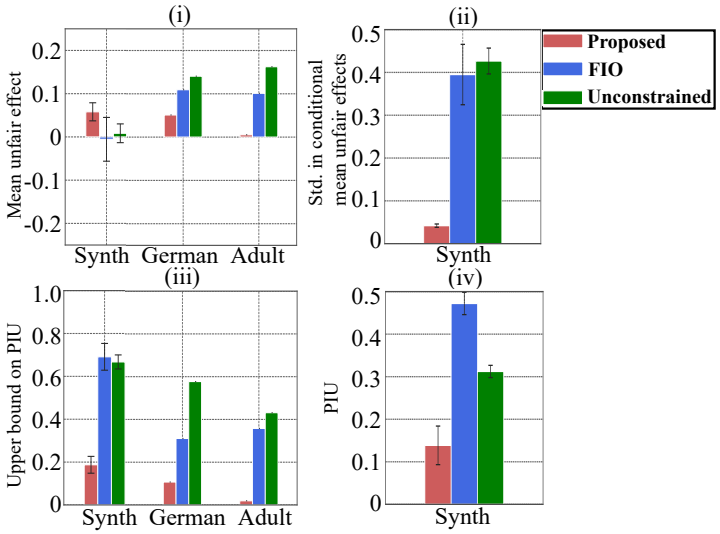

Figure A4: Four statistics of unfair effects on test data when using logistic regression model. The closer they are to zero, the fairer predictions are. Error bars express standard deviations in 10 runs with different datasets. With **Remove**, all statistics are zero (not shown).

I.1 Results with Logistic Regression Model

In Section 5, we evaluated the performance of our method with a two-layered feed-forward neural network as a classifier. To confirm that our method also works well with a simpler classifier, we show its experimental results with logistic regression model.

We compared the performance of our method (**Proposed**) with three baselines: **FIO** (Nabi and Shpitser, 2018), **Unconstrained**, and **Remove** (Kusner et al., 2017). Here, we did not use **PSCF** (Chiappa and Gillam, 2019) as a baseline because it is a neural network-based model.

Table 3 and Figure A4 present the test accuracy and the four statistics of unfair effects. As with the results with the neural network shown in Section 5, the unfair effects of **Proposed** were much closer to zero than **FIO** and **Unconstrained**, and its test accuracy exceeded **Remove**. These experimental results indicate that our method can achieve a good performance without using a complex classifier such as a neural network.

I.2 Confidence Intervals on Error Rates on Real-World Data

In Tables 2 and 3, we present the test accuracy of each method. These results contain the standard deviations on synthetic datasets, which are computed using randomly generated data; however, they do not include those on real-world datasets. For this reason, in this section, we evaluated the statistical significance of the test accuracy on real-world datasets.

We computed the confidence interval of the test error using the test set bound (Langford and Schapire, 2005) (with error rate $\delta = 0.05$), which is a widely-used error measure for binary classification. Here, as with Appendix I.1, we used logistic regression model as a classifier.

Table 4: Test error and test set bound (%) on each real-world dataset when using logistic regression model.

Method	German		Adult	
	Test error	Test set bound	Test error	Test set bound
Proposed	24.0	[18.0, 32.1]	24.8	[24.1, 25.5]
FIO	22.5	[17.1, 31.0]	21.0	[20.4, 21.7]
Unconstrained	21.2	[15.4, 28.8]	20.3	[19.7, 20.9]
Remove	26.2	[19.8, 34.2]	26.6	[25.9, 27.3]

Table 5: Test accuracy (%) on synthetic data that satisfy functional assumptions of PSCF method

Method	Test accuracy (%)
Proposed	72.5 ± 1.1
PSCF	72.5 ± 0.4
Unconstrained	80.0 ± 1.3

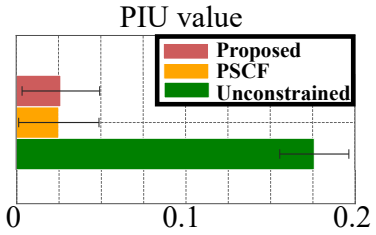

Figure A5: PIU values on test data. The closer they are to zero, the fairer predictions are. Error bars express standard deviations in 10 runs with randomly generated datasets.

Table 4 shows the results. With the German dataset, the test errors did not significantly differ due to the small sample size. However, with the Adult dataset, **Proposed** achieved a significantly lower test error than **Remove** and kept the unfair effect values close to zero (Figure A4). Thus, we confirmed the statistical significance of the test accuracy on the Adult dataset.

I.3 Performance on Data That Satisfy Functional Assumptions

In Section 5, using synthetic data that do not satisfy the functional assumptions of the PSCF method, we show that when those assumptions are violated, PSCF cannot achieve individual-level fairness. By contrast, in this section, we provide experimental results using synthetic data that satisfy the functional assumptions and confirm that with such data, both our method and PSCF can learn an individually fair classifier.

I.3.1 Data

Since the PSCF method assumes that the data are generated by additive noise models (Hoyer et al., 2009), following this assumption, we prepared the synthetic data. We prepared $5,000$ training samples and $1,000$ test samples using the following additive noise model:

$$\begin{aligned} A &= U_A, \quad U_A \sim \text{Bernoulli}(0.6), \\ Q &= \lfloor U_Q \rfloor, \quad U_Q \sim \mathcal{N}(2.5, 5^2), \\ D &= A + \lfloor 0.1Q \rfloor + \lfloor U_D \rfloor, \quad U_D \sim \mathcal{N}(1, 0.5^2), \\ M &= 3A + \lfloor 0.4Q \rfloor + \lfloor U_M \rfloor, \quad U_M \sim \mathcal{N}(1, 0.5^2), \\ Y &= h(A, Q, D, M). \end{aligned} \tag{A30}$$

Table 6: Test accuracy and PIU value on synthetic data

Method	Test accuracy (%)	PIU ($\times 10^{-2}$)
Proposed	80.0 ± 0.9	5.04 ± 3.25
Oracle	78.7 ± 0.9	2.63 ± 0.90

I.3.2 Results

With the above data, we compared the performance of **Proposed** with **PSCF** and **Unconstrained**.

Table 5 and Figure A5 show the test accuracies and the PIU values. The test accuracies of **Proposed** and **PSCF** were almost the same, which were lower than **Unconstrained**. Their PIU values were much close to zero than **Unconstrained**. These results demonstrate that if the data satisfy the functional assumptions, **Proposed** and **PSCF** can achieve almost the same performance and make individually fair predictions.

I.4 Effectiveness of Proposed Upper Bound on PIU

To demonstrate the tightness of our upper bound on PIU, we compared our **Proposed** with **Oracle**, which uses true PIU values as penalties during the training phase with the same penalty parameter value. As with the experiments presented in Section 5, we used a two-layered neural network as a classifier.

Table 6 presents the test accuracy and the PIU value on the synthetic dataset. Both the test accuracy and the PIU value did not differ much, even if we have an oracle access to the true PIU values. These results show the effectiveness of our upper bound on PIU.

I.5 Evaluating Extended Framework

In this section, we show the empirical performance of our extended framework (**Proposed**$_{ex}$), which addresses cases with latent confounders (see Appendix G for the details of our extended framework).

I.5.1 Experimental Settings

We evaluated the performance with synthetic data that contain a latent confounder. Following the causal graph in Figure A3, we prepared such data by sampling from the following SEM:

$$\begin{aligned} &A = U_A, \quad U_A \sim \text{Bernoulli}(0.6), \\ &R = 3A + \lfloor 10H \rfloor + \lfloor U_R \rfloor, \quad U_R \sim \mathcal{N}(1, 0.5^2), \\ &M = A + R + \lfloor U_M \rfloor, \quad U_M \sim \mathcal{N}(1, 0.5^2), \\ &Y = h(A, R, M, H), \end{aligned} \tag{A31}$$

where H denotes a latent confounder, which is sampled by $H \sim \mathcal{N}(1, 0.5^2)$, and function h expresses a logistic regression model that provides the following conditional distribution:

$$\mathrm{P}(Y = 1 | A, R, M, H) = \text{Bernoulli}(\varsigma(-10 + 5A + R + M + 10H)).$$

Given such a synthetic dataset, we used $5,000$ samples for training and $1,000$ samples to test the performance. Other settings are given in the same way as Appendix H.1.

To evaluate the unfair effects, we used the causal graph in Figure A3 with unfair pathway $\pi = \{A \to Y\}$. With SEM (A31), we computed (ii) the standard deviation in the conditional mean unfair effects and (iv) the PIU in a similar manner as in the synthetic data experiments in Section 5. For a fair comparison, we did not evaluate the other two statistics (i.e., (i) and (iii)) because they depend on marginal potential outcome probabilities, whose estimators are formulated in different ways between **Proposed**$_{ex}$ and other methods.

Table 7: Test accuracy on synthetic data with latent confounder: Results are shown by (mean ± standard deviation), computed based on 10 runs with randomly generated different datasets.

Method	Test accuracy (%)
Proposed$_{ex}$	95.7 ± 0.5
Proposed	96.1 ± 0.6
FIO	96.3 ± 0.6
PSCF	93.8 ± 0.9
Unconstrained	97.2 ± 0.6
Remove	94.0 ± 0.6

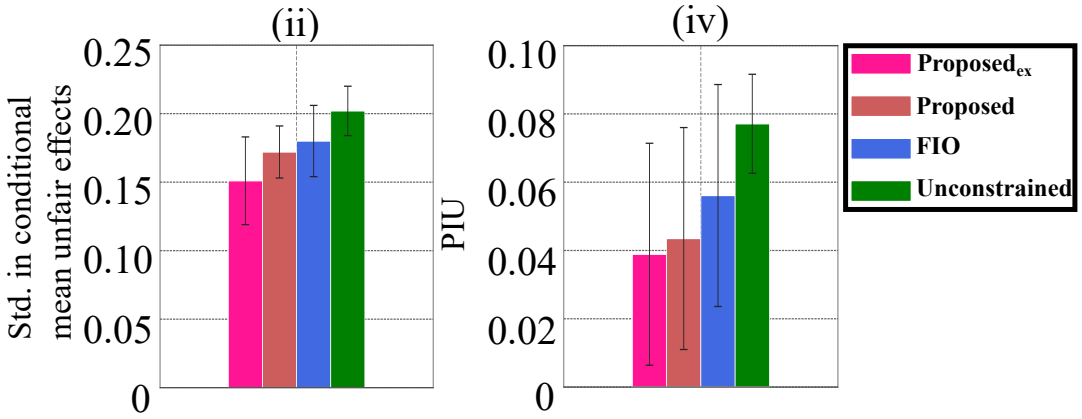

Figure A6: Two statistics of unfair effects on test data. The closer they are to zero, the fairer predictions are. Error bars express standard deviations in 10 runs with randomly generated datasets. As described in Appendix I.5.1, with **PSCF** and **Remove**, both statistics are zero (not shown).

Note that in this experiment, as with **Remove**, the unfair effects of **PSCF** become exactly zero. This is because in this case, **PSCF** makes prediction Y by fixing A's value to zero for all individuals, which completely removes the unfair effect along $\pi = \{A \to Y\}$.

I.5.2 Results

We present the test accuracy in Table 7 and the unfair effects in Figure A6. The results are shown as the mean and standard deviation in 10 experiments with randomly generated data.

With **Proposed**$_{ex}$, both the statistics of the unfair effects were closer to zero than **Proposed** and **FIO** because it uses more reliable estimators of marginal potential outcome probabilities. These results demonstrate that our proposed extension makes fairer predictions than these methods.

The test accuracy of **Proposed**$_{ex}$ exceeded **PSCF** and **Remove**, both of which completely eliminates unfair effects as described in Appendix I.5.1, indicating that our **Proposed**$_{ex}$ can strike a better balance between prediction accuracy and fairness than these two methods.

These results imply that with our proposed extension, we can effectively address cases with latent confounders. Although achieving a good balance between accuracy and fairness in such cases remains an open problem, if there are reliable estimators of lower and upper bounds on marginal potential outcome probabilities, our proposed extension will enable us to strike a good balance between individual-level fairness and prediction accuracy.