

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

Yoichi Chikahara^{1,2}, Shinsaku Sakaue¹, Akinori Fujino¹, and Hisashi Kashima²

¹NTT Communication Science Laboratories

²Kyoto University

June 17, 2020

Abstract

Machine learning is increasingly used to make decisions for individuals in various fields, which require to achieve good prediction accuracy while ensuring fairness with respect to such sensitive features as race or gender. This problem, however, remains difficult in complex real-world scenarios. To effectively quantify unfairness in such scenarios, existing methods utilize *path-specific causal effects*. However, none of them can ensure fairness for each individual without making impractical assumptions. Specifically, these assumptions require us to formulate the true data-generating processes as the *causal model*, which requires an extremely deep understanding of data and is unrealistic in practice. In this paper, we propose a framework for learning an individually fair classifier without relying on the causal model. For this goal, we define the *probability of individual unfairness* (PIU) and solve an optimization problem that constrains its upper bound, which can be estimated from data without the causal model. We elucidate why this constraint can guarantee fairness for each individual. Experimental results demonstrate that our method learns an individually fair classifier at a slight cost of prediction accuracy.

1 Introduction

Machine learning is increasingly being used to make decisions for individuals in lending, hiring, and recidivism prediction. For these applications, it is crucial to make decisions that are not discriminatory with respect to such sensitive features as gender, race, or sexual orientation.

Although many researchers have studied how to make fair decisions while achieving high prediction accuracy [10, 14], it remains a challenge in complex real-world scenarios. For instance, let us consider the case of making hiring decisions for applicants for physically demanding jobs (e.g., firefighters). While it is discriminatory to reject applicants based on their gender, since the job requires physical strength, it is **not** discriminatory to reject them due to their physical strength. Since physical strength is affected by gender, rejecting applicants due to physical strength leads to different rejection rates between men and women. Although this difference is **not** unfair, traditional methods remove it to reject men and women at an equal

rate [10]. 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 significantly reduces the prediction accuracy.

To achieve high prediction accuracy, we need to remove only an unfair difference in decision outcomes. To do so, existing methods [7, 16, 19] measure unfairness for each individual by a *path-specific causal effect* [3], which we call an *unfair effect*. They utilize the two types of mean unfair effects over individuals. Some methods remove the *conditional mean unfair effect*, i.e., the mean unfair effect **only** over the individuals who have identical feature attributes [7, 16]. This ensures fairness for each individual; however, as pointed out by Wu et al. [30], this mean value can only be estimated under impractical assumptions. For instance, the method presented in [16] assumes that we can express the true data-generating processes as a *causal model*, which requires an extremely deep understanding of data and is unrealistic in practice. The *fair inference on outcome* method (FIO) [19] avoids such demanding assumptions by using the mean unfair effect over **all** individuals, which can be estimated from data. However, even when this mean is zero, depending on feature attributes, unfair effects might be largely positive for some individuals and largely negative for others, which is seriously discriminatory for these individuals.

The goal of this paper is to propose a learning framework that guarantees fairness for each individual without relying on a deep understanding of data or access to the causal model. For this goal, we define the *probability of individual unfairness* (PIU), which is the probability that an unfair effect is not zero for an individual, and learn the classifier so that PIU becomes small. However, when estimating PIU, we need a causal model, which is unavailable in practice. To overcome this problem, instead of PIU, we utilize its upper bound that can be consistently estimated from data and solve an optimization problem that constrains the upper bound to be close to zero. **Our contributions** are summarized as follows:

- We establish a framework for guaranteeing fairness for all individuals without requiring the causal model. For this goal, we derive the upper bound on PIU and constrain the upper bound value to be close to zero, which can be estimated from data.
- We elucidate why our constraint guarantees fairness for each individual in Sections 3.2.2 and 3.3. We also show how our framework can be extended to the cases where there are unobserved variables called *latent confounders* in Section 3.4.
- We experimentally show that the unfair effects of our method are much closer to zero than FIO [19] for all individuals, demonstrating that our framework achieves individual-level fairness.

2 Background

In this paper, we consider a binary classification problem for providing prediction $\hat{Y} \in \{0, 1\}$ on decision outcome $Y \in \{0, 1\}$ for each individual. Here \hat{Y} is provided by a classifier $h_\theta(\mathbf{X})$, where \mathbf{X} denotes a set of features of an individual, which contains sensitive feature $A \in \{0, 1\}$.

To measure the unfairness of \hat{Y} with respect to A , we utilize a path-specific causal effect [3], which is defined based on a *causal graph* and a *causal model*.

2.1 Causal graph and causal model

A causal graph is a directed acyclic graph (DAG) whose nodes and edges represent random variables and causal relationships, respectively. For instance, suppose that we predict hiring decisions \hat{Y} for physically demanding jobs. Then a causal graph is given as shown in Figure 1, where $A, Q, M \in \mathbf{X}$ represent gender (0 for female, 1 for male), qualification, and physical strength, respectively. By using this graph, we can express our prior knowledge on discrimination that decisions are unfair only if they are based on gender. To do so, we regard the direct pathway $A \rightarrow \hat{Y}$ as an unfair pathway π (i.e., $\pi = \{A \rightarrow \hat{Y}\}$).

Based on the causal graph, the causal model describes how each random variable is determined. For instance, when the causal graph in Figure 1 is given, each variable is expressed by the following functional relationships called *structural equations*:

$$A = f_A(\mathbf{U}_A), \quad Q = f_Q(\mathbf{U}_Q), \quad M = f_M(A, Q, \mathbf{U}_M), \quad \hat{Y} = h_\theta(A, Q, M). \quad (1)$$

Here, each variable $X \in \mathbf{X} = \{A, Q, M\}$ is determined by a deterministic function f_X whose input contains unobserved variables \mathbf{U}_X representing factors outside the causal graph, and \hat{Y} is determined by the classifier h_θ . For the formal definition of the causal model, see, Appendix A.1.

Formulating the causal model requires the knowledge of true data-generating processes, which is unavailable in practice. In contrast, a causal graph can be inferred from data in some cases [13]. For this reason, as with the existing methods [7, 19, 31], we assume that a causal graph is given.

2.2 Measuring unfairness by path-specific causal effects

To quantify the unfairness of prediction \hat{Y} for each individual, we use a path-specific causal effect along unfair pathways π (or a π -specific effect, in short).

For instance, to measure the unfairness of hiring decisions for physically demanding jobs, we use the π -specific effect with $\pi = A \rightarrow \hat{Y}$ (a.k.a., a *(natural) direct causal effect* [22]), which quantifies how greatly the prediction \hat{Y} is influenced only by changing gender A . As an example, we consider the case when changing gender from female to male; the opposite case can be considered in the same way. In this case, the π -specific effect is measured as the difference between the two predicted decision outcomes under *counterfactual situations*: when the gender is changed to female ($A = 0$), and when it is changed to male ($A = 1$) while physical strength M is set in the same way as when $A = 0$. These two predicted decision outcomes are represented by random variables called *potential outcomes* $\hat{Y}_{A \leftarrow 0}$ and $\hat{Y}_{A \leftarrow 1 \parallel \pi}$, respectively, which are defined by the following two causal models:

$$A = 0, \quad Q = f_Q(\mathbf{U}_Q), \quad M = f_M(0, Q, \mathbf{U}_M), \quad \hat{Y}_{A \leftarrow 0} = h_\theta(0, Q, M), \quad (2)$$

$$A = 1, \quad Q = f_Q(\mathbf{U}_Q), \quad M = f_M(0, Q, \mathbf{U}_M), \quad \hat{Y}_{A \leftarrow 1 \parallel \pi} = h_\theta(1, Q, M). \quad (3)$$

Compared with the causal model (1), several structural equations are replaced in (2) and (3), which are called *interventions*. In the causal model (2), the structural equation of A is replaced with $A = 0$, which is referred to the intervention $do(A = 0)$. Meanwhile, in (3), the structural equations of A and M are replaced with $A = 1$ and $M = f_M(0, Q, \mathbf{U}_M)$, respectively. Such a replacement is called the intervention $do(A = 1)$.

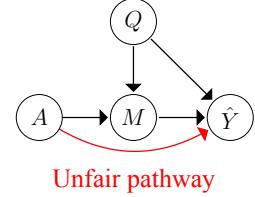


Figure 1: Causal graph that represents a scenario of making hiring decisions for physically demanding jobs.

along pathways π (see, [3] for details), and how to replace structural equations depends on π . Note that if we consider the effect of changing gender from male ($A = 1$) to female ($A = 0$), we can formulate the potential outcomes in the same way as $\hat{Y}_{A \leftarrow 1}$ and $\hat{Y}_{A \leftarrow 0 \parallel \pi}$.

As in this example, when unfair pathways are given as $\pi = A \rightarrow \hat{Y}$, we can eliminate the π -specific effect by making prediction \hat{Y} without A since it measures how greatly \hat{Y} depends on A values. However, when π contains multiple pathways from A to \hat{Y} (for example, see, Figure 3 (b)), we need to remove all features that are affected by A , which significantly reduces the prediction accuracy.

To address such general cases, we need to quantify the π -specific effect by $\hat{Y}_{A \leftarrow 1 \parallel \pi} - \hat{Y}_{A \leftarrow 0}$, which we call an unfair effect. Unfortunately, we cannot infer it for each individual because the joint distribution of $\hat{Y}_{A \leftarrow 0}$ and $\hat{Y}_{A \leftarrow 1 \parallel \pi}$ is defined by the causal model, which is unavailable in practice.

On the other hand, without using the causal model, we can estimate the mean unfair effect over all individuals [25], which is defined as the difference between marginal potential outcome probabilities:

$$\mathbb{E}_{\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1 \parallel \pi}} [\hat{Y}_{A \leftarrow 1 \parallel \pi} - \hat{Y}_{A \leftarrow 0}] = P(\hat{Y}_{A \leftarrow 1 \parallel \pi} = 1) - P(\hat{Y}_{A \leftarrow 0} = 1), \quad (4)$$

where the equality holds because of the linearity of the expectation and $\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1 \parallel \pi} \in \{0, 1\}$. To estimate the marginal probabilities $P(\hat{Y}_{A \leftarrow 0} = 1)$ and $P(\hat{Y}_{A \leftarrow 1 \parallel \pi} = 1)$ in (4), we can use the existing estimators (see, Appendix A.2 for an example).

2.3 Overcoming weakness of existing methods

Since we can estimate the mean unfair effect in (4) from data, the existing FIO method aims to learn fair predictive models by constraining it to zero [19]. However, this cannot guarantee fairness for each individual. This is because even when the mean is zero, depending on the values of features \mathbf{X} , unfair effects might be largely positive for some individuals and largely negative for others, which are seriously discriminatory for them. Unfortunately, this issue cannot be resolved by naively constraining the mean of the absolute values of unfair effects. This is because estimating this quantity requires the joint distribution of potential outcomes, which cannot be inferred from data.

To guarantee individual-level fairness, i.e., to ensure that unfair effects are not affected by features \mathbf{X} , other methods such as the (path-dependent) counterfactual fairness [16, Section S4] remove the conditional mean unfair effect conditioned on \mathbf{X} , i.e., the average of unfair effects **only** over individuals who have identical feature attributes [7, 16]. As pointed out by Wu et al. [30], however, when the causal model is unknown, we cannot estimate this quantity from data without making additional restrictive assumptions on the underlying causal model (for detail, see, [26] and Appendix A.3). Without these assumptions, we need the data on each individual whose sensitive feature value is **both** $A = 0$ and $A = 1$ (e.g., both female and male), which are impossible to obtain in practice.

To resolve the shortcomings of the existing methods, we propose a framework for learning a classifier by constraining an unfair effect to be zero for **all** individuals; that is, by making potential outcomes take the same value (i.e., $\hat{Y}_{A \leftarrow 0} = \hat{Y}_{A \leftarrow 1 \parallel \pi}$) for **all** individuals. This condition is sufficient (but not necessary) for achieving individual-level fairness; that is, if potential outcomes take the same value for all individuals, unfair effect values are not affected by features \mathbf{X} , but the converse is not true. This means that we are imposing a stronger fairness constraint than the one required for achieving individual-level fairness. While this might

decrease the prediction accuracy, we experimentally confirmed that it resulted in only a slight decrease in test accuracy (see, Section 4). Compared with the existing methods for ensuring individual-level fairness [7, 16], our method has a clear advantage that it requires a much weaker assumption; we only need to estimate the marginal potential outcome probabilities in (4). Our method can also be extended to relax this assumption for addressing cases with unobserved variables called latent confounders (see, Section 3.4).

3 Proposed method

3.1 Eliminating individual unfairness based on PIU

To guarantee that prediction \hat{Y} is fair for each individual, we learn a classifier by constraining the potential outcomes (i.e., the predicted decision outcomes under counterfactual situations) to be the same for all individuals. To do so, we define the following quantity:

Definition 1 (Probability of Individual Unfairness (PIU)) *For unfair pathways π in a causal graph and potential outcomes $\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1 \parallel \pi} \in \{0, 1\}$, we define the probability of individual unfairness (PIU) by $P(\hat{Y}_{A \leftarrow 0} \neq \hat{Y}_{A \leftarrow 1 \parallel \pi})$.*

PIU is the probability that potential outcomes $\hat{Y}_{A \leftarrow 0}$ and $\hat{Y}_{A \leftarrow 1 \parallel \pi}$ become different. Unlike the conditional mean unfair effect described in Section 2.3, PIU is defined without conditioning on features \mathbf{X} . However, as with the conditional mean unfair effect, PIU can be used to guarantee individual-level fairness. Specifically, by constraining PIU to zero, we can guarantee that potential outcomes take the same value (i.e., $\hat{Y}_{A \leftarrow 0} = \hat{Y}_{A \leftarrow 1 \parallel \pi}$) with probability 1 regardless of the values of \mathbf{X} , which is sufficient to ensure individual-level fairness. Unfortunately, we cannot directly constrain PIU to zero. This is because to estimate PIU, we need the joint distribution of $\hat{Y}_{A \leftarrow 0}$ and $\hat{Y}_{A \leftarrow 1 \parallel \pi}$, which is unavailable since it is defined by the unknown causal model as described in Section 2.2.

To overcome this issue, instead of PIU, we utilize its upper bound, which can be estimated from data without using the causal model. Specifically, to make the PIU value close to zero, we impose a constraint on its upper bound, which is presented in the next section.

3.2 Constraining and penalizing upper bound on PIU

3.2.1 Upper bound formulation

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

Theorem 1 (Upper bound on PIU) *Suppose that potential outcomes $\hat{Y}_{A \leftarrow 0}$ and $\hat{Y}_{A \leftarrow 1 \parallel \pi}$ are binary. Then for any joint distribution $P(\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1 \parallel \pi})$, PIU is upper bounded as follows:*

$$P(\hat{Y}_{A \leftarrow 0} \neq \hat{Y}_{A \leftarrow 1 \parallel \pi}) \leq 2 P^I(\hat{Y}_{A \leftarrow 0} \neq \hat{Y}_{A \leftarrow 1 \parallel \pi}), \quad (5)$$

where P^I is the independent joint distribution, i.e., $P^I(\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1 \parallel \pi}) = P(\hat{Y}_{A \leftarrow 0}) P(\hat{Y}_{A \leftarrow 1 \parallel \pi})$.

The proof is detailed in Appendix B. Theorem 1 states that whatever joint distribution potential outcomes $\hat{Y}_{A \leftarrow 0}$ and $\hat{Y}_{A \leftarrow 1 \parallel \pi}$ follow, the resulting PIU value is at most twice the PIU value that is approximated with the independent joint distribution P^I .

This approximated PIU can be estimated from data without using the causal model. Since potential outcomes are binary (i.e., $\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1 \parallel \pi} \in \{0, 1\}$), it is equal to the probability that potential outcome values become $(\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1 \parallel \pi}) = (0, 1)$ or $(1, 0)$; in other words,

$$\begin{aligned} & P^I(\hat{Y}_{A \leftarrow 0} \neq \hat{Y}_{A \leftarrow 1 \parallel \pi}) \\ &= P(\hat{Y}_{A \leftarrow 1 \parallel \pi} = 1)(1 - P(\hat{Y}_{A \leftarrow 0} = 1)) + (1 - P(\hat{Y}_{A \leftarrow 1 \parallel \pi} = 1))P(\hat{Y}_{A \leftarrow 0} = 1). \end{aligned} \quad (6)$$

To estimate (6), we formulate the estimators of marginal probabilities $P(\hat{Y}_{A \leftarrow 0} = 1)$ and $P(\hat{Y}_{A \leftarrow 1 \parallel \pi} = 1)$ by using the conditional probability $c_\theta(\mathbf{X}) = P(\hat{Y} = 1 \mid \mathbf{X})$ that is provided by the classifier h_θ ; we let $c_\theta(\mathbf{X}) = h_\theta(\mathbf{X}) \in \{0, 1\}$ when classifier h_θ is deterministic. For instance, when the causal graph in Figure 1 is given, using the conditional probability c_θ and features of n individuals $\{\mathbf{x}_i\}_{i=1}^n = \{a_i, q_i, m_i\}_{i=1}^n$, we can express the weighted estimators in [15] as

$$\hat{p}_\theta^{A \leftarrow 0} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i c_\theta(0, q_i, m_i), \quad \hat{p}_\theta^{A \leftarrow 1 \parallel \pi} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i c_\theta(1, q_i, m_i) \text{ where } \hat{w}_i = \frac{\hat{P}(A = 0 | m_i, q_i)}{\hat{P}(A = 0 | q_i)},$$

where weight \hat{w}_i for $i \in \{1, \dots, n\}$ depends on the feature values q_i and m_i . To compute \hat{w}_i , we estimate conditional distributions $\hat{P}(A = 0 | m_i, q_i)$ and $\hat{P}(A = 0 | q_i)$ by fitting a parametric model (e.g., logistic regression or neural network) to training data beforehand.

3.2.2 Proposed objective function

To achieve a small PIU value, we learn the classifier h_θ by using the upper bound in (5). Since PIU is always smaller than its upper bound, by constraining the upper bound to be close to zero, we can find the classifier parameter θ that minimizes the prediction loss while achieving a small PIU value.

Let L be a loss function that measures the accuracy of classifier h_θ and let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be n training instances. Then we can formulate the learning problem by

$$\underset{\theta}{\text{minimize}} \frac{1}{n} \sum_{i=1}^n L(h_\theta(\mathbf{x}_i), y_i) \text{ subject to } 2(\hat{p}_\theta^{A \leftarrow 1 \parallel \pi}(1 - \hat{p}_\theta^{A \leftarrow 0}) + (1 - \hat{p}_\theta^{A \leftarrow 1 \parallel \pi})\hat{p}_\theta^{A \leftarrow 0}) \leq \delta, \quad (7)$$

where $\delta \in [0, 1]$ is a hyperparameter. The constraint requires the upper bound on PIU to be at most δ .

To simplify the optimization problem (7), we add a penalty term with hyperparameter $\lambda \geq 0$ to the loss function, which yields the following unconstrained optimization problem:

$$\underset{\theta}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n L(h_\theta(\mathbf{x}_i), y_i) + \lambda(\hat{p}_\theta^{A \leftarrow 1 \parallel \pi}(1 - \hat{p}_\theta^{A \leftarrow 0}) + (1 - \hat{p}_\theta^{A \leftarrow 1 \parallel \pi})\hat{p}_\theta^{A \leftarrow 0}). \quad (8)$$

To minimize (8), we used the stochastic gradient descent method [28] in our experiments. We discuss the computation time and convergence guarantees when minimizing (8) in Appendix C.

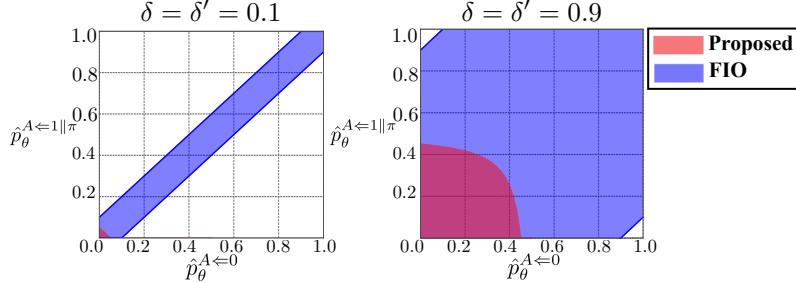


Figure 2: Feasible regions of our constraint (red) and that of FIO (blue) with $\delta = \delta' = 0.1, 0.9$.

From the objective function (8), we can see the reason why penalizing the upper bound on PIU guarantees fairness for each individual. When setting $\lambda \rightarrow \infty$ in (8), the penalty term forces the marginal probabilities to be $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) = (0, 0)$ or $(1, 1)$. This guarantees that the potential outcomes take the same value with probability 1, which is sufficient to guarantee individual-level fairness. Roughly speaking, to ensure $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) \approx (0, 0)$ or $(1, 1)$, our method penalizes the prediction \hat{Y} more strongly for individuals who are more likely to receive unfair decision outcomes (see, Appendix D for details).

3.3 Comparison with existing fairness constraint

To show the effectiveness of our fairness constraint in (7), we compare it with the constraint of the FIO method [19]. Although our method uses not the constraint but the penalty term derived from it, which is shown in (8), the following comparison clarifies the difference from FIO.

Using a hyperparameter $\delta' \in [0, 1]$, the constraint of FIO limits the mean unfair effect (4) to lie in

$$-\delta' \leq \hat{p}_\theta^{A \leftarrow 1 \parallel \pi} - \hat{p}_\theta^{A \leftarrow 0} \leq \delta', \quad (9)$$

which ensures the marginal probabilities to be $\hat{p}_\theta^{A \leftarrow 0} = \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$ if we set $\delta' = 0$.

Figure 2 shows the feasible region of our constraint (red) and the constraint of FIO (blue), obtained by graphing the hyperbolic inequality in (7) and the linear inequality in (9), respectively. To clarify their difference, we consider the case where $\delta = \delta'$ by letting $\delta \in [0, 1]$.¹

Our constraint only accepts the region $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) \approx (0, 0)$ or $(1, 1)$, where the potential outcomes take the same value (i.e., $\hat{Y}_{A \leftarrow 0} = \hat{Y}_{A \leftarrow 1 \parallel \pi}$) with high probability. This demonstrates that our constraint with the hyperbolic inequality in (7) effectively makes potential outcomes take the same value for all individuals, which is sufficient to guarantee individual-level fairness. In contrast, the constraint of FIO always accepts the point $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) = (0.5, 0.5)$, where it is totally unknown whether potential outcomes take the same value (see, Appendix E for details); hence predictions might be unfair for some individuals. This implies that FIO cannot ensure individual-level fairness.

¹It may seem unreasonable to let $\delta \in [0, 1]$ because the left hand side of our constraint in (7) can become up to 2. Nevertheless, we do not consider $\delta \in (1, 2]$ because when using the constraint in (7) with such a value of δ , we impose no constraint on PIU since the PIU value, which is at most 1, is always less than δ if $\delta \in (1, 2]$.

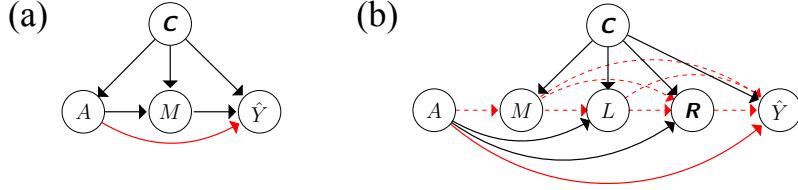


Figure 3: Causal graphs when making prediction \hat{Y} with (a) COMPAS dataset and (b) UCI Adult dataset. While direct pathways with red solid edges are unfair, the pathways through red dashed edges in (b) are unfair if they go from A to \hat{Y} via M (i.e., $A \rightarrow M \rightarrow \dots \rightarrow \hat{Y}$).

3.4 Extension for dealing with latent confounders

So far, to estimate the upper bound on PIU, we have assumed that marginal distributions of potential outcomes can be inferred from data. In fact, this assumption does not hold if observed features are affected by unobserved variables called *latent confounders* [21], which is possible in practice. Although inferring the marginal probabilities in such cases is extremely challenging, we can estimate the lower and upper bounds on them if a certain causal graph is given.

We describe how our method can be extended by utilizing such bounds in Appendix F and experimentally evaluate its performance in Appendix G.2.

4 Synthetic and real-world data experiments

We compared the performance of our method (**Proposed**) with **FIO** and with the baseline (**Unconstrained**), which only minimizes the training loss and does not use constraints or penalty terms. As classifiers, we used a feed-forward neural network that contains two linear layers with 100 and 50 hidden neurons. Due to space limitations, other settings are detailed in Appendix G.1.1.

Data and causal graphs. We prepared synthetic data that express a scenario of hiring decisions for physically demanding jobs. We randomly sampled the data from the causal model, whose formulation is described in Appendix G.1.2. Given such synthetic data, we used 5,000 samples to train the classifier and 1,000 samples to test the performance. To define the unfair effect, we used the causal graph in Figure 1 with unfair pathway $\pi = \{A \rightarrow \hat{Y}\}$. As regards real-world data, we used the COMPAS dataset [2] and the UCI Adult dataset [4], which were downloaded from [19]. The former contains the records of prisoners including race A and recidivism \hat{Y} , and the latter is US census data that contain the features of each individual, including gender A and income Y . To measure the fairness of prediction \hat{Y} on Y , following [19], we used the causal graphs shown in Figure 3, which are detailed in Appendix G.1.3. Unfair pathway π for COMPAS dataset is $A \rightarrow \hat{Y}$, and π for Adult dataset consists of $A \rightarrow \hat{Y}$ and all pathways from A to \hat{Y} via marital status M (i.e., $A \rightarrow M \rightarrow \dots \rightarrow \hat{Y}$).

Accuracy and fairness. We compared test accuracy and unfair effects of **Proposed** with **FIO** and with **Unconstrained**. To evaluate the unfair effects, we used the four statistics: (i) the mean unfair effect, (ii) the upper bound on PIU, (iii) the PIU, and (iv) the standard deviation in the conditional mean unfair effects. To compute (iii) and (iv), in synthetic data experiments, we used the causal model as described

Table 1: Accuracy and unfair effect on test data. The closer unfair effects are to zero, the fairer predicted decisions are. In (a), results are shown by (mean \pm standard deviation), computed based on 10 runs with randomly generated different datasets.

(a) Synthetic data					
Method	Test accuracy		Unfair effects		
	(%)	Mean	Upper bound on PIU	PIU	Std in conditional means
Proposed	89.2 \pm 1.0	$(1.63 \pm 0.52) \times 10^{-5}$	$(\mathbf{1.63 \pm 2.66}) \times 10^{-4}$	$(1.00 \pm 3.17) \times 10^{-4}$	$(2.26 \pm 7.16) \times 10^{-4}$
FIO	89.4 \pm 1.1	$(\mathbf{1.06 \pm 9.26}) \times 10^{-5}$	0.482 \pm 0.231	0.196 \pm 0.126	0.217 \pm 0.117
Unconstrained	90.3 \pm 1.24	$(6.02 \pm 2.40) \times 10^{-2}$	0.679 \pm 0.200	0.276 \pm 0.107	0.202 \pm 0.096

(b) Real-world data					
Data	Method	Test accuracy		Unfair effects	
		(%)	Mean	Upper bound on PIU	
COMPAS	Proposed	65.2	3.09×10^{-5}	1.29×10^{-3}	
	FIO	65.5	$-\mathbf{7.40 \times 10^{-6}}$	0.492	
	Unconstrained	66.3	3.02×10^{-2}	0.844	
Adult	Proposed	76.6	5.95×10^{-4}	1.11×10^{-4}	
	FIO	78.1	2.41×10^{-2}	0.111	
	Unconstrained	80.0	0.204	0.996	

in Appendix G.1.2. Note that we cannot evaluate them in real-world data experiments because the causal models are unknown for real-world data.

Tables 1a and 1b present the results on synthetic and real-world data, respectively. Here Table 1a shows the means and the standard deviations of test accuracy and unfair effects in 10 experiments with randomly generated training and test datasets.

With **Proposed**, all of the four statistics of unfair effects were much closer to zero than the other methods, indicating that **Proposed** made substantially fairer decisions. This is because by constraining the upper bound on PIU, **Proposed** forces unfair effects to be close to zero for all individuals, which guarantees the other statistics to be close to zero. Since **Proposed** imposes a stronger fairness constraint than that of **FIO**, its test accuracy became slightly lower than but almost comparable to **FIO**. These results demonstrate that at a slight cost of accuracy, our method can learn an individually fair classifier. To support this, we also present additional results in Appendix G.2.

Compared with **Proposed**, as shown in Table 1a, **FIO** provided much larger values for the three statistics: the upper bound on PIU, the PIU, and the standard deviation in the conditional mean unfair effects. This is because constraining the mean unfair effect does not imply reducing these three statistics. In particular, large values of the standard deviation in the conditional mean unfair effects indicate that unfair effects are largely affected by feature attributes of each individual, which indicates that significantly unfair predictions were made based on feature attributes. In Table 1b, although large upper bound values of PIU do not necessarily mean large values of the standard deviation in conditional mean unfair effects, unlike **Proposed**, we cannot guarantee that **FIO** made fair predictions for each individual, which is problematic in practice.

Penalty parameter effects. We further compared the performance of **Proposed** with **FIO** by using various penalty parameter values. Figure 4 presents the test accuracy and the upper bound on PIU for the

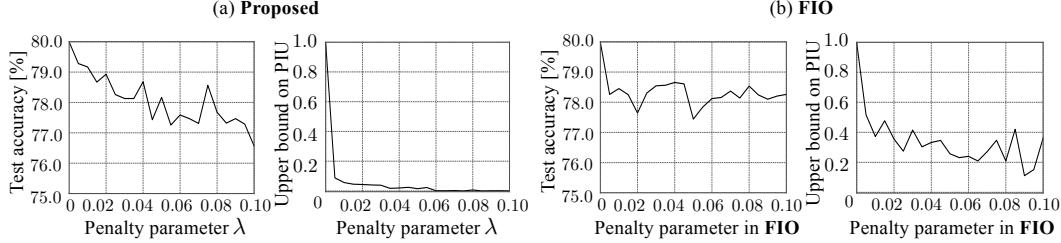


Figure 4: Test accuracy and upper bound on PIU of (a) **Proposed** and (b) **FIO** for UCI Adult dataset when setting penalty parameter of each method to 0, 0.005, 0.010, \dots , 0.100.

UCI Adult dataset. As we can see, as the penalty parameter value increased, the upper bound value of **Proposed** sharply dropped to (almost) zero, while that of **FIO** remained deviated from zero. As regards the test accuracy, the performance of **Proposed** slightly decreased; however, it remained comparable to that of **FIO**. These results demonstrate that our method effectively reduces upper bound on PIU, which enables us to achieve comparable prediction accuracy.

5 Related work

Causality-based fairness. Motivated by recent developments in inferring causal graphs [8, 17], a growing number of causality-based approaches to fairness have been proposed [7, 16, 19, 20, 24, 29, 31, 32, 33, 34]. However, none of them can ensure fairness for each individual without impractical assumptions. To relax the assumptions, *path-specific counterfactual fairness* (PC-fairness) [30] has been proposed, which estimates lower and upper bounds on the conditional mean unfair effect. However, it is designed to measure unfairness in data, not to learn a fair predictive model.

In contrast, we have established a framework for learning an individually fair classifier 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 (5). Compared with existing notions, this upper bound has the following two advantages.

First, it can be used for binary potential outcomes. PIU is a functional of the joint distribution of potential outcomes, and several bounds on such a functional have been derived [9, 12]. However, these bounds are designed for continuous potential outcomes and cannot be used in our setting.

Second, it is tighter than the existing upper bound [23]. The bound presented in [23] is designed for an expectation of random variables, whose upper bound has been studied in robust optimization [1]. Since PIU can be expressed as $\mathbb{E}_{\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1} \parallel \pi} [\mathbf{1}(\hat{Y}_{A \leftarrow 0} \neq \hat{Y}_{A \leftarrow 1} \parallel \pi)]$, where $\mathbf{1}$ is an indicator function, we can use this bound for PIU. However, when using it, the bound becomes much looser than ours. Specifically, while the multiplicative constant in (5) is 2, this value becomes 200 with the bound of [23]. If we use such a loose upper bound, we need to impose an excessively severe constraint on it to guarantee that the PIU value is close to zero. In contrast, by utilizing our upper bound, we can avoid imposing such a severe constraint, thus preventing an unnecessary decrease in the prediction accuracy.

6 Conclusion

We proposed a learning framework for guaranteeing fairness for each individual without requiring a deep understanding of data or access to the causal model. To develop such a framework, we derived the upper bound on PIU and solved an optimization problem for constraining its value to be close to zero. We elucidate why our constraint guarantees individual-level fairness from a perspective of feasible region of the constraint. Our experimental results suggest that while the existing FIO method makes decisions that are significantly unfair for some of the individuals, our method makes decisions that are fair for all individuals, demonstrating the effectiveness of our fairness constraint.

Broader Impact

Machine learning is increasingly used to make decisions that severely affect people’s life chances such as loan approvals, hiring decisions, and recidivism predictions [2, 5]. The huge societal impact of these decisions on people’s lives has raised growing concerns about fairness. This is because training data often exhibit a discriminatory bias if they include the records of past discriminatory decisions made by humans. This bias is difficult to quantify in complex real-world scenarios where many features are correlated with each other and are affected by the sensitive feature (e.g., the scenario of the UCI Adult dataset in Section 4); however, we can effectively measure it by utilizing the concept of causality called path-specific causal effects.

Based on this concept, we have proposed a framework for achieving fairness for each individual, which requires much weaker assumptions and is more widely applicable than the existing methods [7, 16]. Our framework enables us to guarantee that no individual suffers from discrimination because of their feature attributes. For instance, this will benefit banking, employers in companies, and criminal courts, who wish to improve the reliability of their decisions.

On the negative side, as with the existing methods [7, 19, 33], our framework does not fully address the cases where features of each individual are affected by unobserved variables called latent confounders. Although guaranteeing fairness in these cases is extremely challenging, they are possible in complex real-world scenarios. To tackle such scenarios, in Section 3.4, we presented an extension of our framework, which can be applied in some cases where a certain causal graph is given. Addressing more various cases with latent confounders constitutes our future work.

References

- [1] S. Agrawal, Y. Ding, A. Saberi, and Y. Ye. Correlation robust stochastic optimization. In *SODA*, pages 1087–1096, 2010.
- [2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias, 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [3] C. Avin, I. Shpitser, and J. Pearl. Identifiability of path-specific effects. In *IJCAI*, pages 357–363, 2005.
- [4] K. Bache and M. Lichman. UCI machine learning repository: Datasets. <http://archive.ics.uci.edu/ml/datasets>, 2013.

- [5] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. <http://www.fairmlbook.org>.
- [6] J. V. Burke, A. S. Lewis, and M. L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.
- [7] S. Chiappa and T. P. Gillam. Path-specific counterfactual fairness. In *AAAI*, pages 7801–7808, 2019.
- [8] Y. Chikahara and A. Fujino. Causal inference in time series via supervised learning. In *IJCAI*, pages 2042–2048, 2018.
- [9] Y. Fan, E. Guerre, and D. Zhu. Partial identification of functionals of the joint distribution of "potential outcomes". *Journal of Econometrics*, 197(1):42–59, 2017.
- [10] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268, 2015.
- [11] J. Ferrera. *An introduction to nonsmooth analysis*. Academic Press, 2013.
- [12] S. Firpo and G. Ridder. Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, 213(1):210–234, 2019.
- [13] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.
- [14] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *NeurIPS*, pages 3315–3323, 2016.
- [15] M. Huber. Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29(6):920–943, 2014.
- [16] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *NeurIPS*, pages 4066–4076, 2017.
- [17] H. Li, V. Cabeli, N. Sella, and H. Isambert. Constraint-based causal structure learning with consistent separating sets. In *NeurIPS*, pages 14257–14266, 2019.
- [18] C. H. Miles, P. Kanki, S. Meloni, and E. J. T. Tchetgen. On partial identification of the natural indirect effect. *Journal of Causal Inference*, 5(2), 2017.
- [19] R. Nabi and I. Shpitser. Fair inference on outcomes. In *AAAI*, pages 1931–1940, 2018. <https://github.com/raziehna/fair-inference-on-outcomes>.
- [20] R. Nabi, D. Malinsky, and I. Shpitser. Learning optimal fair policies. In *ICML*, pages 4674–4682, 2019.
- [21] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.
- [22] M. L. Petersen, S. E. Sinisi, and M. J. van der Laan. Estimation of direct causal effects. *Epidemiology*, pages 276–284, 2006.

- [23] A. Rubinstein and S. Singla. Combinatorial prophet inequalities. In *SODA*, pages 1671–1687, 2017.
- [24] C. Russell, M. J. Kusner, J. Loftus, and R. Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *NeurIPS*, pages 6414–6423, 2017.
- [25] I. Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, 37(6):1011–1035, 2013.
- [26] I. Shpitser and J. Pearl. Complete identification methods for the causal hierarchy. *JMLR*, 9(Sep):1941–1979, 2008.
- [27] I. Shpitser and J. Pearl. Effects of treatment on the treated: identification and generalization. In *UAI*, pages 514–521, 2009.
- [28] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
- [29] Y. Wu, L. Zhang, and X. Wu. On discrimination discovery and removal in ranked data using causal graph. In *KDD*, pages 2536–2544, 2018.
- [30] Y. Wu, L. Zhang, X. Wu, and H. Tong. PC-fairness: A unified framework for measuring causality-based fairness. In *NeurIPS*, pages 3399–3409, 2019.
- [31] J. Zhang and E. Bareinboim. Equality of opportunity in classification: A causal approach. In *NeurIPS*, 2018.
- [32] L. Zhang and X. Wu. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics*, 4(1):1–16, 2017.
- [33] L. Zhang, Y. Wu, and X. Wu. A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI*, 2017.
- [34] L. Zhang, Y. Wu, and X. Wu. Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

Appendix

A Background on causal inference

A.1 Definition of causal model

Given the causal graph, the causal model describes how each random variable is determined in two ways, depending on whether it is contained in *endogenous variables* \mathbf{V} (i.e., the variables expressed in the causal graph) or in *exogenous variables* \mathbf{U} (i.e., the others, which are often unobserved) [21]. Each variable $V \in \mathbf{V}$ is determined by the following structural equation:

$$V = f_V(\mathbf{Pa}_V, \mathbf{U}_V),$$

where $f_V \in \mathbf{F}$ is a deterministic function, where \mathbf{F} is the set of functions, $\mathbf{Pa}_V \subseteq \mathbf{V} \setminus V$ are the variables that are depicted as parents of V in the causal graph, and \mathbf{U}_V are the variables in \mathbf{U} . In particular, if we make prediction \hat{Y} by classifier $h_\theta(\mathbf{X})$ with the input features \mathbf{X} , the endogenous variables are $\mathbf{V} = \{\mathbf{X}, \hat{Y}\}$, and the structural equation of \hat{Y} is given by

$$\hat{Y} = f_{\hat{Y}}(\mathbf{X}, \mathbf{U}_{\hat{Y}}) = h_\theta(\mathbf{X}),$$

where $\mathbf{U}_{\hat{Y}}$ becomes an empty set ϕ (i.e., $\mathbf{U}_{\hat{Y}} = \phi$) if the classifier h_θ is deterministic. Meanwhile, each $U \in \mathbf{U}$ is determined by other factors that are outside the causal model. Given the three sets \mathbf{V} , \mathbf{U} , and \mathbf{F} , a causal model is formally defined by a triplet $\mathcal{M} = \{\mathbf{V}, \mathbf{U}, \mathbf{F}\}$.

A.2 Example of marginal potential outcome probability estimator

Although it is impossible to observe potential outcomes, their marginal probabilities can be represented as the distributions of the observed variables if several conditional independence relations hold [21].

For instance, given the causal graph in Figure 1, we can use the marginal probability estimator presented in [15]. To do so, by using the classifier $h_\theta(A, Q, M)$, we compute the conditional probability $c_\theta(A, Q, M) = P(\hat{Y} = 1 | A, Q, M)$; we let $c_\theta(A, Q, M) = h_\theta(A, Q, M) \in \{0, 1\}$ if h_θ is a deterministic classifier. Using c_θ , this estimator is formulated as

$$P(\hat{Y}_{A=0} = 1) = \mathbb{E}_{Q,M}[w c_\theta(0, Q, M)], \quad P(\hat{Y}_{A=1} = 1) = \mathbb{E}_{Q,M}[w c_\theta(1, Q, M)]. \quad (\text{A1})$$

Here

$$w = \frac{P(A = 0 | M, Q)}{P(A = 0 | Q)}$$

denotes a weight depending on conditional distributions $P(A | M, Q)$ and $P(A | Q)$, which can be estimated by fitting a parametric model (e.g., logistic regression) to training data beforehand.

Based on the estimator (A1), using the features of n individuals $\{\mathbf{x}_i\}_{i=1}^n = \{a_i, q_i, m_i\}$, we can estimate the marginal probabilities as described in Section 3.2.1.

A.3 Difficulty in estimating conditional mean unfair effect

To measure unfairness for each individual, as described in Section 2.3, several existing methods use the conditional mean unfair effect [7, 16, 30]. Letting \mathbf{X} be features of each individual, this quantity is defined as the difference between the following two conditional probabilities of potential outcomes:

$$\mathbb{E}_{\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1} \mid \pi} [\hat{Y}_{A \leftarrow 1 \mid \pi} - \hat{Y}_{A \leftarrow 0} \mid \mathbf{X} = \mathbf{x}] = P(\hat{Y}_{A \leftarrow 1 \mid \pi} = 1 \mid \mathbf{X} = \mathbf{x}) - P(\hat{Y}_{A \leftarrow 0} = 1 \mid \mathbf{X} = \mathbf{x}). \quad (\text{A2})$$

Unfortunately, these conditional probabilities conditioned on the features \mathbf{X} are difficult to estimate because the sensitive feature A is included in \mathbf{X} (i.e., $A \in \mathbf{X}$). For instance, letting A be gender, if we consider the mean unfair effect only over the men with identical attributes $\mathbf{X} = \mathbf{x}$, we need the probability that these men would receive favorable outcomes if they were female ($A = 0$), i.e., $P(\hat{Y}_{A \leftarrow 0} = 1 \mid \mathbf{X} = \mathbf{x})$, where $\mathbf{X} = \mathbf{x}$ contains $A = 1$ (male). However, since no individual can be both male ($A = 1$) and female ($A = 0$), estimating this probability requires additional assumptions.

According to the previous results in [27], if we consider the conditional probability that is conditioned only on A , (e.g., $P(\hat{Y}_{A \leftarrow 0} = 1 \mid A = 1)$), which is known as the expression of *effect of treatment on the treated* (ETT), we can infer it from data under reasonable assumptions. However, if we consider more general cases such as (A2), as shown by [26], we cannot consistently estimate the conditional probability without making a restrictive assumption on the causal model. Specifically, we need to assume that exogenous variables \mathbf{U} are not shared with the causal model that are provided by interventions. This assumption does not hold for all the causal model with $\mathbf{U} \setminus \mathbf{U}_A \neq \emptyset$, where \emptyset is an empty set (e.g., the causal model in (1)). See, [26] for further details of this assumption.

B Proof of Theorem 1

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

Then we have

$$\begin{aligned} p_{10} + p_{11} &= \alpha, \\ p_{00} + p_{01} &= 1 - \alpha, \\ p_{01} + p_{11} &= \beta, \quad \text{and} \\ p_{10} + p_{00} &= 1 - \beta. \end{aligned} \quad (\text{A3})$$

As described in Section 3.2.1, the right-hand side in (5) in Theorem 1 can be represented by marginal probabilities. With the above notations, this 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 (A3) are non-negative, p_{01} and p_{10} become at most $\min\{\beta, 1 - \alpha\}$ and $\min\{\alpha, 1 - \beta\}$, respectively, yielding

$$p_{01} + p_{10} \leq \min\{\beta, 1 - \alpha\} + \min\{\alpha, 1 - \beta\}. \quad (\text{A4})$$

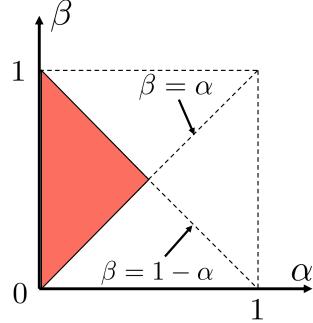


Figure A1: Red area represents region where marginal probability values α and β satisfy $\alpha \leq \beta \leq 1 - \alpha$.

Hence, it suffices to prove

$$\min\{\beta, 1 - \alpha\} + \min\{\alpha, 1 - \beta\} \leq 2\beta(1 - \alpha) + 2\alpha(1 - \beta). \quad (\text{A5})$$

Since both sides in (A5) 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 A1 as the red region.

In this case, since $\min\{\beta, 1 - \alpha\} = \beta$ and $\min\{\alpha, 1 - \beta\} = \alpha$, (A5) is reduced to

$$\begin{aligned} \beta + \alpha &\leq 2\beta(1 - \alpha) + 2\alpha(1 - \beta) \\ \iff \alpha + \beta - 4\alpha\beta &\geq 0. \end{aligned} \quad (\text{A6})$$

Since $\alpha + \beta \leq 1$ holds in this case, we have the inequality $\alpha + \beta - (\alpha + \beta)^2 \geq 0$. Using this inequality, the inequality (A6) can be proven as follows:

$$\alpha + \beta - 4\alpha\beta = \alpha + \beta - (\alpha + \beta)^2 + (\alpha - \beta)^2 \geq 0. \quad (\text{A7})$$

Thus, we prove Theorem 1.

C Computation time and convergence guarantees for minimizing (8)

To minimize the objective function (8), we use the stochastic gradient descent method [28].

When using it, we computed the penalty term and its gradient over samples in each mini-batch. The computation time that is required to evaluate the objective function (8) and to compute its gradient is as much as the one to evaluate the training loss in (8) and to compute its gradient, respectively.

Whether we can guarantee that the gradient descent method converges depends on the choice of classifier h_θ . For instance, if it is a neural network classifier, we cannot guarantee that the stochastic gradient descent method [28] converges because the objective function (8) 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 by using e.g., the gradient sampling method [6] because in this case, the gradient of the objective function becomes *locally Lipschitz continuous* [11, Chapter 2].

D Influence of proposed fairness constraint on predictions

This section illustrates how our fairness constraint affects the predicted decision outcome \hat{Y} for each individual.

As discussed in Section 3.2.2, when setting the hyperparameter $\lambda \rightarrow \infty$ in the objective function (8), our fairness constraint forces marginal potential outcome probabilities to be $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) = (0, 0), (1, 1)$. In practice, since we set λ at a finite value (i.e., $\lambda < \infty$), our constraint makes the marginal probabilities to be close to those values, i.e., $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) \approx (0, 0), (1, 1)$.

For instance, when estimating $\hat{p}_\theta^{A \leftarrow 0}$ and $\hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$ based on the existing estimators presented in (A1) in Appendix A.2, our method imposes the following constraint:

$$\hat{p}_\theta^{A \leftarrow 0} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i c_\theta(0, q_i, m_i) \approx 0, \quad \hat{p}_\theta^{A \leftarrow 1 \parallel \pi} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i c_\theta(1, q_i, m_i) \approx 0, \quad \text{or} \quad (\text{A8})$$

$$\hat{p}_\theta^{A \leftarrow 0} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i c_\theta(0, q_i, m_i) \approx 1, \quad \hat{p}_\theta^{A \leftarrow 1 \parallel \pi} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i c_\theta(1, q_i, m_i) \approx 1, \quad (\text{A9})$$

where $c_\theta(A, Q, M) = \text{P}(\hat{Y} \mid A, Q, M)$ is the conditional probability of \hat{Y} , which is provided by classifier h_θ , and $w_i > 0$ ($i \in \{1, \dots, n\}$) is a weight, which is formulated as follows:

$$\hat{w}_i = \frac{\hat{\text{P}}(A = 0 \mid m_i, q_i)}{\hat{\text{P}}(A = 0 \mid q_i)}. \quad (\text{A10})$$

From the constraint (A8), we can see that for each $i \in \{1, \dots, n\}$, if the weight w_i is large, we need to impose a strong constraint to ensure $c_\theta(0, q_i, m_i) \approx 0$ and $c_\theta(1, q_i, m_i) \approx 0$. In contrast, from (A9), we can find that if w_i is large, we need to impose a severe constraint to guarantee $c_\theta(0, q_i, m_i) \approx 1$ and $c_\theta(1, q_i, m_i) \approx 1$.

In the above example, the weight value w_i in (A10) depends on the values of features q_i and m_i . To illustrate when w_i becomes large, let us consider the case of making hiring decisions of physically demanding jobs. In this case, the weight w_i becomes large for an individual i whose physical strength M is likely to be the one of a woman ($A = 0$) and small for those whose physical strength M is unlikely to be the one of a woman. Therefore, our method strongly constrains the prediction \hat{Y} for individuals whose physical strengths seem to be those of females to prevent them from receiving an unfair prediction.

With the weighted estimators of marginal potential outcome probabilities, our method effectively imposes a fairness constraint for each individual based on the estimated weight value.

E Further comparison of fairness constraints

As described in Section 3.3, our fairness constraint effectively makes potential outcomes take the same value (i.e., $\hat{Y}_{A \leftarrow 0} = \hat{Y}_{A \leftarrow 1 \parallel \pi}$) while the constraint of FIO does not. To illustrate this, we compare the possible values of PIU under each fairness constraint.

E.1 Comparison of possible PIU values

To discuss the true value of PIU, instead of the estimated marginal probabilities $\hat{p}_\theta^{A \leftarrow 0}$ and $\hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$ used in Section 3.3, we consider the true ones $\alpha = P(\hat{Y}_{A \leftarrow 0} = 1)$ and $\beta = P(\hat{Y}_{A \leftarrow 1 \parallel \pi} = 1)$. In addition, we use the notations of the joint probabilities in Appendix B. Specifically, we let the true joint probabilities of $(\hat{Y}_{A \leftarrow 0}, \hat{Y}_{A \leftarrow 1 \parallel \pi}) = (0, 0), (0, 1), (1, 0)$, 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 by using the marginal probabilities α and β as follows:

$$|\alpha - \beta| \leq p_{01} + p_{10} \leq \min\{\beta, 1 - \alpha\} + \min\{\alpha, 1 - \beta\}, \quad (\text{A11})$$

which we prove in Appendix E.2.

When using our constraint, as presented in Figure 2, the marginal probabilities are constrained as $(\alpha, \beta) \approx (0, 0)$ or $(1, 1)$. If α and β satisfy this, since both lower and upper bounds in (A11) become close to zero, we can see that PIU is constrained to almost zero (i.e., $p_{01} + p_{10} \approx 0$).

On the other hand, as described in Section 3.3, the constraint of FIO always accepts the marginal probabilities $(\alpha, \beta) = (0.5, 0.5)$. At this point, the lower and upper bounds in (A11) become 0 and 1, respectively; that is, $0 \leq p_{01} + p_{10} \leq 1$. This implies that it is totally unknown whether the PIU value is high since the joint probabilities are unknown in practice. Therefore, FIO cannot ensure that potential outcomes take the same value for all individuals, which is insufficient to guarantee individual-level fairness.

E.2 Proof of (A11)

We prove the lower bound in (A11). Since α and β are marginal probabilities, we have

$$\begin{aligned} p_{10} + p_{11} &= \alpha, \\ p_{01} + p_{11} &= \beta, \end{aligned}$$

which are equivalent to

$$\begin{aligned} p_{10} &= \alpha - p_{11}, \\ p_{01} &= \beta - p_{11}, \end{aligned}$$

respectively. By summing up both, we have

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

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

$$p_{01} + p_{10} \geq \alpha + \beta - 2 \min\{\alpha, \beta\} = |\alpha - \beta|. \quad (\text{A12})$$

Combined with the upper bound on $p_{01} + p_{10}$ in (A4), we prove (A11).

F Upper bound on PIU in the presence of latent confounders

In this section, we show how to formulate the upper bound on PIU when there are unobserved variables called latent confounders in a given causal graph, which is possible in real-world scenarios. Formally, a latent confounder is defined as an unobserved exogenous variable that affects two endogenous variables.

In the presence of latent confounders, it is extremely challenging to infer the marginal potential outcome probabilities $P(\hat{Y}_{A \leftarrow 0} = 1)$ and $P(\hat{Y}_{A \leftarrow 1 \parallel \pi} = 1)$. However, existing studies have shown that we can infer lower and upper bounds on marginal probabilities in some cases.

Let the lower and upper bounds on marginal probabilities $P(\hat{Y}_{A \leftarrow 0} = 1)$ and $P(\hat{Y}_{A \leftarrow 1 \parallel \pi} = 1)$ be

$$l_\theta^{A \leftarrow 0} \leq P(\hat{Y}_{A \leftarrow 0} = 1) \leq u_\theta^{A \leftarrow 0}, \quad (\text{A13})$$

$$l_\theta^{A \leftarrow 1 \parallel \pi} \leq P(\hat{Y}_{A \leftarrow 1 \parallel \pi} = 1) \leq u_\theta^{A \leftarrow 1 \parallel \pi}, \quad (\text{A14})$$

respectively. Then, to guarantee individual-level fairness in the presence of latent confounders, we can reformulate our upper bound constraint in (7) as

$$2(\hat{u}_\theta^{A \leftarrow 1 \parallel \pi}(1 - \hat{l}_\theta^{A \leftarrow 0}) + (1 - \hat{l}_\theta^{A \leftarrow 1 \parallel \pi})\hat{u}_\theta^{A \leftarrow 0}) \leq \delta. \quad (\text{A15})$$

By using a penalty term derived from (A15), as with (8), we can formulate an unconstrained optimization problem.

In what follows, as an example of the estimators of the lower and upper bounds $\hat{l}_\theta^{A \leftarrow 0}$, $\hat{u}_\theta^{A \leftarrow 0}$, $\hat{u}_\theta^{A \leftarrow 0}$, and $\hat{u}_\theta^{A \leftarrow 1 \parallel \pi}$, we present the existing estimators in [18].

F.1 Example of lower and upper bounds on marginal potential outcome probabilities

According to the previous result in [18], the lower and upper bounds on marginal potential outcome probabilities can be obtained for certain causal graphs.

Figure A2 presents an example of a causal graph. Here $A \in \{0, 1\}$ and $\hat{Y} \in \{0, 1\}$ are binary random variables, M is a discrete one, R can be a continuous or discrete one, and H is a latent confounder that affects the two variables R and Y .² Note that although H is an exogenous variable, which is often not shown in a causal graph, we express it as a gray node in Figure A2.

²As regards A , although the authors of [18] dealt with a general case, where A can be continuous or discrete, for simplicity, we consider a binary case, i.e., $A \in \{0, 1\}$

Regarding the direct effect along $A \rightarrow \hat{Y}$ as π , the previous work [18] showed that lower and upper bounds on $P(\hat{Y}_{A \leftarrow 1} \mid \pi = 1)$ are expressed as

$$\begin{aligned}\hat{l}^{A \leftarrow 1 \parallel \pi} &= \sum_m \max\{0, P(M = m \mid A = 0) - 1 + \sum_r P(\hat{Y} = 1 \mid A = 1, m, r) P(R = r \mid A = 1)\}, \\ \hat{u}^{A \leftarrow 1 \parallel \pi} &= \sum_m \min\{P(M = m \mid A = 0), \sum_r P(\hat{Y} = 1 \mid A = 1, m, r) P(R = r \mid A = 1)\},\end{aligned}$$

respectively. By using the conditional distribution $c_\theta(1, M, R) = P(\hat{Y} = 1 \mid A = 1, M, R)$ provided by the classifier output $h_\theta(1, M, R)$, these bounds can be estimated as the following functions of classifier parameter θ :

$$\hat{l}_\theta^{A \leftarrow 1 \parallel \pi} = \sum_m \max\{0, \hat{P}(M = m \mid A = 0) - 1 + \sum_r c_\theta(1, m, r) \hat{P}(R = r \mid A = 1)\} \quad (\text{A16})$$

$$\hat{u}_\theta^{A \leftarrow 1 \parallel \pi} = \sum_m \min\{\hat{P}(M = m \mid A = 0), \sum_r c_\theta(1, m, r) \hat{P}(R = r \mid A = 1)\}, \quad (\text{A17})$$

respectively. Here the conditional distributions $\hat{P}(M = m \mid A = 0)$ and $\hat{P}(R = r \mid A = 1)$ can be estimated by fitting a regression model (e.g., logistic regression or neural networks) to training data beforehand.

As with (A16) and (A17), we can formulate the estimated lower and upper bounds on marginal probability $P(\hat{Y}_{A \leftarrow 0} = 1)$ as follows:

$$\hat{l}_\theta^{A \leftarrow 0} = \sum_m \max\{0, \hat{P}(M = m \mid A = 0) - 1 + \sum_r c_\theta(0, m, r) \hat{P}(R = r \mid A = 0)\} \quad (\text{A18})$$

$$\hat{u}_\theta^{A \leftarrow 0} = \sum_m \min\{\hat{P}(M = m \mid A = 0), \sum_r c_\theta(0, m, r) \hat{P}(R = r \mid A = 0)\}, \quad (\text{A19})$$

respectively.

Note that if we use the above lower and upper bounds, it is not easy to solve the optimization problem with convergence guarantees because these lower and upper bounds in the penalty term are not differentiable. Leveraging other bounds on marginal probabilities to formulate an objective function that is differentiable and easy to optimize constitutes our future work.

G Experimental settings and additional experiments

G.1 Experimental settings

G.1.1 Settings of our method (Proposed) and existing method (FIO)

To evaluate the performance of **Proposed** and **FIO**, as their classifier h_θ , we used a two-layered feed-forward neural network that contains linear layers with 100 and 50 hidden neurons. As their activation functions, we used a sigmoid function.

To train classifier h_θ , we employed cross-entropy loss as loss function L , used 1,000 training samples as a minibatch, and stopped the training after 1,000 epochs.

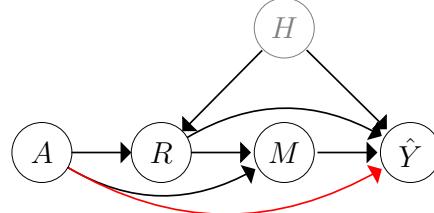


Figure A2: Example of causal graph containing latent confounder H (gray node), which affects both R and Y . Red pathway represents unfair pathway π .

To compare the best performance of each method, we selected the hyperparameter values of **Proposed** and **FIO**. For each of the two methods, we used a grid search with 0.1 grid size to select the value of penalty parameter from $[0.0, 2.0]$.

G.1.2 Settings in synthetic data experiments

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

$$\begin{aligned}
 A &= U_A, \quad U_A \sim \text{Bernoulli}(0.6), \\
 Q &= U_Q, \quad U_Q \sim \mathcal{N}(5, 2.5^2), \\
 M &= 3A + \lfloor 0.4Q \rfloor + \lfloor U_M \rfloor, \quad U_M \sim \mathcal{N}(3, 1.5^2), \\
 Y &= h(A, Q, M).
 \end{aligned} \tag{A20}$$

Here Bernoulli and \mathcal{N} represent the Bernoulli and Gaussian distributions, respectively, $\lfloor \cdot \rfloor$ is a floor function that returns an integer by removing the decimal places, and the function h expresses a logistic regression model that provides the conditional distribution

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

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

To compute the mean unfair effect and the upper bound on PIU, we estimated marginal potential outcome probabilities $\hat{p}_\theta^{A \leftarrow 0}$ and $\hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$ by using the estimators described in Section 3.2.1.

To evaluate the values of PIU and the standard deviation in conditional mean unfair effects, we sampled potential outcomes $\hat{Y}_{A \leftarrow 0}$ and $\hat{Y}_{A \leftarrow 1 \parallel \pi}$ by using the causal model (A20). We sampled $\hat{Y}_{A \leftarrow 0}$ from the following causal model:

$$\begin{aligned}
 A &= 0, \\
 Q &= U_Q, \quad U_Q \sim \mathcal{N}(5, 2.5^2), \\
 M &= 3A + \lfloor 0.4Q \rfloor + \lfloor U_M \rfloor, \quad U_M \sim \mathcal{N}(3, 1.5^2), \\
 \hat{Y}_{A \leftarrow 0} &= h_\theta(A, Q, M),
 \end{aligned}$$

where h_θ is the classifier. As regards $\hat{Y}_{A \leftarrow 1 \parallel \pi}$, we sampled it from

$$\begin{aligned} A &= 1, \\ Q &= U_Q, \quad U_Q \sim \mathcal{N}(5, 2.5^2), \\ M &= \lfloor 0.4Q \rfloor + \lfloor U_M \rfloor, \quad U_M \sim \mathcal{N}(3, 1.5^2), \\ \hat{Y}_{A \leftarrow 1 \parallel \pi} &= h_\theta(A, Q, M). \end{aligned}$$

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

$$\hat{P}(\hat{Y}_{A \leftarrow 0} \neq \hat{Y}_{A \leftarrow 1 \parallel \pi}) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(\hat{y}_{A \leftarrow 0, i} \neq \hat{y}_{A \leftarrow 1 \parallel \pi, i}),$$

where $\mathbf{1}$ is an indicator function that takes 1 if $\hat{y}_{A \leftarrow 0, i} \neq \hat{y}_{A \leftarrow 1 \parallel \pi, i}$ ($i \in \{1, \dots, n\}$) and 0 otherwise.

Meanwhile, we computed the standard deviation in conditional mean unfair effects $\hat{\sigma}$ by separating n individuals into K subgroups of individuals who have the same values of features \mathbf{X} , taking the mean unfair effects over individuals in each subgroup, and computing their standard deviation. Let the individuals in the k -th subgroup ($k = 1, \dots, K$) have the identical feature attributes $\mathbf{X} = \mathbf{x}^k$, where the superscript k represents the k -th subgroup. Then, using $\{(\hat{y}_{A \leftarrow 0, i}, \hat{y}_{A \leftarrow 1 \parallel \pi, i})\}_{i=1}^n$, we estimated the standard deviation of conditional mean unfair effects over K subgroups as follows:

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

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

$$\hat{\mu}^k = \frac{1}{n^k} \sum_{i \in \{1, \dots, n\} | \mathbf{x}_i = \mathbf{x}^k} \mathbf{1}(\hat{y}_{A \leftarrow 0, i}, \hat{y}_{A \leftarrow 1 \parallel \pi, i}),$$

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.,

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

G.1.3 Settings in real-world data experiments

In real-world data experiments, we used two datasets, i.e., the COMPAS dataset [2] and the UCI Adult dataset [4], whose causal graphs are displayed in Figure 3.

COMPAS is a decision support system for predicting recidivism (i.e., whether a prisoner repeats his or her crime), which is being used across the US to determine whether to release prisoners. The records of prisoners whose recidivism was predicted by COMPAS are publicly available, and they are known as the COMPAS dataset [2].

Using this dataset, we evaluated recidivism prediction performance. We used 4,751 samples to train our classifier and 529 samples to test the performance. To evaluate the fairness, following [19], we used the causal graph in Fig. 3 (a), where A denotes race (African-American or Caucasian), \hat{Y} expresses the predicted recidivism, M represents past criminal records, and \mathbf{C} corresponds to demographic information (age and gender). We regarded race A as a sensitive feature and the direct causal effect along $A \rightarrow \hat{Y}$ as an unfair effect.

To estimate potential outcome probabilities $\hat{p}_\theta^{A \leftarrow 0}$ and $\hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$ in the COMPAS dataset, we used the existing estimators [15]. Given the features of n individuals $\{a_i, m_i, \mathbf{c}_i\}_{i=1}^n$, they are formulated by

$$\hat{p}_\theta^{A \leftarrow 0} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i h_\theta(0, m_i, \mathbf{c}_i), \quad \hat{p}_\theta^{A \leftarrow 1 \parallel \pi} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i h_\theta(1, m_i, \mathbf{c}_i) \text{ where } \hat{w}_i = \frac{\hat{P}(A = 0 | m_i, \mathbf{c}_i)}{\hat{P}(A = 0 | \mathbf{c}_i)},$$

where h_θ represents the classifier. We estimated conditional distributions $\hat{P}(A = 0 | m_i, \mathbf{c}_i)$ and $\hat{P}(A = 0 | \mathbf{c}_i)$ by fitting the logistic regression model to the data beforehand.

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

With this dataset, we evaluated the performance when predicting whether or not 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 [19], we used the causal graph in Fig. 3 (b). We regarded gender A as a sensitive feature. Other features are marital status M , level of education L , information about occupation \mathbf{R} (e.g., working hours per week), age and nationality \mathbf{C} , and predicted income \hat{Y} . Unfair pathways π are the direct pathway $A \rightarrow \hat{Y}$ and the pathways from A to \hat{Y} via M (i.e., $A \rightarrow M \rightarrow \dots \rightarrow \hat{Y}$).

To compute $\hat{p}_\theta^{A \leftarrow 0}$ and $\hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$ with this dataset, we used the estimators that are derived in the same way as those in the previous work [15]. Using the features of n individuals $\{a_i, m_i, l_i, \mathbf{r}_i, \mathbf{c}_i\}_{i=1}^n$, they are computed by

$$\hat{p}_\theta^{A \leftarrow 0} = \frac{1}{n} \sum_{i=1}^n \hat{w}_{A \leftarrow 0, i} h_\theta(0, m_i, l_i, \mathbf{r}_i, \mathbf{c}_i), \quad \hat{p}_\theta^{A \leftarrow 1 \parallel \pi} = \frac{1}{n} \sum_{i=1}^n \hat{w}_{A \leftarrow 1 \parallel \pi, i} h_\theta(1, m_i, l_i, \mathbf{r}_i, \mathbf{c}_i), \quad (\text{A21})$$

where each of the weights $\hat{w}_{A \leftarrow 0, i}$ and $\hat{w}_{A \leftarrow 1 \parallel \pi, i}$ is given by

$$\hat{w}_{A \leftarrow 0, i} = \frac{\hat{P}(A = 0 | m_i, l_i, \mathbf{r}_i, \mathbf{c}_i)}{\hat{P}(A = 0 | \mathbf{c}_i)}, \quad \hat{w}_{A \leftarrow 1 \parallel \pi, i} = \frac{\hat{P}(A = 1 | m_i, \mathbf{c}_i)}{P(A = 1 | \mathbf{c}_i)} \cdot \frac{\hat{P}(A = 0 | m_i, l_i, \mathbf{r}_i, \mathbf{c}_i)}{\hat{P}(A = 0 | m_i, \mathbf{c}_i)}.$$

We estimated the three conditional distributions $\hat{P}(A | m, l, \mathbf{r}, \mathbf{c})$, $\hat{P}(A = 0 | \mathbf{c})$, and $\hat{P}(A = 1 | m, \mathbf{c})$ by fitting the logistic regression model to data beforehand.

G.1.4 Computing infrastructure

We used PyTorch 1.1.0 as an implementation of the optimization algorithm [28].

To perform synthetic data experiments and real-world data experiments with the COMPAS dataset, we used a 64-bit macOS machine with 2.7GHz Intel Core i7 CPUs and 16GB RAM. Since the UCI Adult dataset is large, to experiment with it, we used a 64-bit CentOS machine with 2.6GHz Xeon E5-2697A-v4 (x2) CPUs and 512GB RAM.

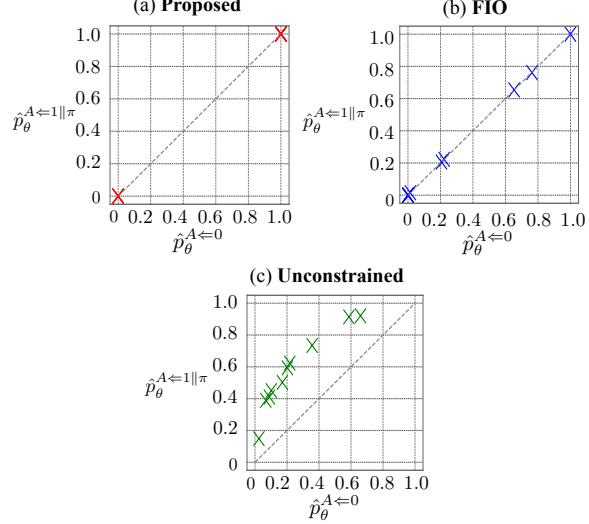


Figure A3: Estimated marginal potential outcome probabilities: (a) **Proposed** (red), (b) **FIO** (blue), and (c) **Unconstrained** (green). Each point (denoted by x) represents a pair of values of estimated marginal probabilities in each of 10 experiments with randomly generated synthetic training datasets. Gray dashed diagonal line represents line $\hat{p}_\theta^{A \leftarrow 0} = \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$.

G.2 Additional experiments

In this section, we describe additional experimental results to show the effectiveness of our method.

G.2.1 Estimated marginal potential outcome probabilities

To achieve individual-level fairness, our method guarantees that potential outcomes take the same value for all individuals. To do so, as described in Section 3.2.2, our method learns a classifier by using a penalty term that forces the marginal potential outcome probabilities to be $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) \approx (0, 0)$ or $(1, 1)$.

To experimentally verify this, we show the values of the estimated marginal probabilities $\hat{p}_\theta^{A \leftarrow 0}$ and $\hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$ for **Proposed**, **FIO**, and **Unconstrained** by using the same synthetic data as the one described in synthetic data experiments in Section 4. Specifically, for each of the 10 randomly generated training datasets that are used in those experiments, we computed the marginal probabilities by using the learned classifier.

Figure A3 presents the results. As expected, the estimated marginal probabilities of **Proposed** were $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) \approx (0, 0)$ or $(1, 1)$, which can be seen from Figure A4. In contrast, since **FIO** imposes the constraint $\hat{p}_\theta^{A \leftarrow 0} \approx \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$, the estimated marginal probabilities were almost on the line $\hat{p}_\theta^{A \leftarrow 0} = \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}$ (i.e., the gray dashed diagonal line in Figure A3). With these results, we can confirm that the constraint of **Proposed** only accepts the region that satisfies $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) \approx (0, 0)$ or $(1, 1)$. As described in Appendix E.1, this is why our method can achieve small PIU values.

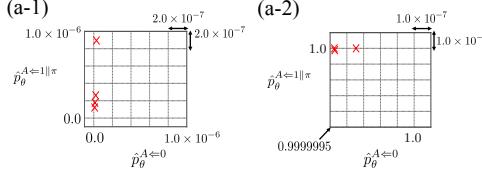


Figure A4: Estimated marginal potential outcome probabilities of **Proposed** in Figure A3 (a) are shown around $(\hat{p}_\theta^{A \leftarrow 0}, \hat{p}_\theta^{A \leftarrow 1 \parallel \pi}) = (0, 0)$ in (a-1) and $(1, 1)$ in (a-2).

Table 2: Accuracy and unfair effect on test data when using logistic regression model. The closer unfair effects are to zero, the fairer predicted decisions are. In (a), results are shown by (mean \pm standard deviation), computed based on 10 runs with randomly generated different datasets.

(a) Synthetic data

Method	Test accuracy (%)	Unfair effects			
		Mean	Upper bound on PIU	PIU	Std in conditional mean
Proposed	71.1 ± 6.9	$(1.25 \pm 1.91) \times 10^{-4}$	$(7.18 \pm 4.75) \times 10^{-4}$	$(3.64 \pm 5.05) \times 10^{-4}$	$(3.04 \pm 4.59) \times 10^{-3}$
FIO	72.8 ± 4.3	$(-1.77 \pm 1.04) \times 10^{-3}$	0.761 ± 0.183	$(8.65 \pm 1.78) \times 10^{-2}$	0.111 ± 0.056
Unconstrained	73.0 ± 3.4	$(-9.40 \pm 3.06) \times 10^{-2}$	0.857 ± 0.102	0.135 ± 0.024	0.173 ± 0.042

(b) Real-world data

Data	Method	Test accuracy (%)	Unfair effects	
			Mean	Upper bound on PIU
COMPAS	Proposed	66.3	6.41×10^{-5}	1.21×10^{-11}
	FIO	66.4	1.39×10^{-5}	0.969
	Unconstrained	67.7	-6.22×10^{-2}	0.994
Adult	Proposed	70.2	-2.64×10^{-11}	6.00×10^{-9}
	FIO	72.2	-3.52×10^{-2}	0.501
	Unconstrained	82.4	0.197	0.894

G.2.2 Experiments with logistic regression model

In experiments in Section 4, as classifier h_θ of each method, we used the two-layered feed-forward neural network.

In this section, we present the performance of each method when using the logistic regression model as h_θ . For **FIO**, we trained the logistic regression model using the implementation of the authors of [19] and set the hyperparameter for the fairness constraint by following this implementation. Other settings are provided in the same way as the experiments in Section 4.

Table 2 presents the results. Again, with **Proposed**, all of the four statistics of unfair effects were much closer to zero than with **FIO** while the test accuracy was comparable to that of **FIO**. Combined with the results in Table 1, we can conclude that for both of the neural network classifier and the logistic regression model, our method could guarantee fairness for each individual at a slight sacrifice in prediction accuracy.

Table 3: Comparison of test accuracy with naive baseline on synthetic data. Results are shown by (mean \pm standard deviation), computed based on 10 runs with randomly generated different datasets. Although mean unfair effect, PIU, and standard deviation in conditional mean unfair effects of **Naive Baseline** are not shown, they are guaranteed to be zero as described in Appendix G.2.3.

Method	Classifier	Test accuracy		Unfair effects		
		(%)	Mean	PIU	Std in conditional means	
Proposed Naive baseline	Logistic	71.1 \pm 6.9 63.1 \pm 1.1	$(1.25 \pm 1.91) \times 10^{-4}$ -	$(3.64 \pm 5.05) \times 10^{-4}$ -	$(3.04 \pm 4.59) \times 10^{-3}$ -	
Proposed Naive baseline	NN	89.2 \pm 1.0 88.7 \pm 1.0	$(1.63 \pm 0.52) \times 10^{-5}$ -	$(1.00 \pm 3.17) \times 10^{-4}$ -	$(2.26 \pm 7.16) \times 10^{-4}$ -	

G.2.3 Additional baseline comparison

We compared our method with a naive baseline that learns the classifier without gender A .

As described in Section 2.2, when unfair pathway π is only direct pathway $A \rightarrow \hat{Y}$, we can completely remove the unfair effect by making prediction \hat{Y} without using sensitive feature A .

Compared with this naive baseline, our method has the following difference. While it can control the trade-off between fairness and accuracy by appropriately choosing the value of penalty parameter λ , this naive baseline cannot. Although making decisions without A makes the unfair effects be exactly zero for all individuals, since this corresponds to setting $\lambda \rightarrow \infty$ in our method, the prediction accuracy becomes lower than our method.

To verify this, we performed additional synthetic data experiments. Using the same synthetic data as the ones described in Section 4, we trained the two-layered neural network classifier and compared the mean and the standard deviation of test accuracies in 10 experiments with randomly generated training and test datasets.

Table 3 presents the results, where the results of **Proposed** are the same as those in Table 1a and Table 2a. We do not show the mean unfair effect, PIU, and the standard deviation in conditional mean unfair effects of **Naive baseline** because they are guaranteed to be zero as already described.

As expected, **Proposed** achieved higher test accuracies than **Naive baseline** while keeping the three statistics of unfair effects to be almost zero. These results demonstrate that our method can strike a better balance between fairness and prediction accuracy than the naive strategy that learns a classifier without sensitive feature A .

G.2.4 Performance on synthetic data with latent confounder

In Appendix F, we show how our method can be extended to address the case with latent confounders. In this section, we present the performance of our proposed extension (**Proposed**_{ex}), which uses a penalty term that is derived from the constraint (A15) instead of the one in (8).

To evaluate the performance, we used the causal graph in Figure A2 and prepared synthetic data that

are sampled from the following causal model:

$$\begin{aligned}
A &= U_A, \quad U_A \sim \text{Bernoulli}(0.6), \\
R &= 3A + \lfloor 10H \rfloor + \lfloor U_R \rfloor, \quad H \sim \mathcal{N}(1, 0.5^2), \quad U_R \sim \mathcal{N}(1, 0.5^2), \\
M &= A + R + \lfloor U_M \rfloor, \quad U_M \sim \mathcal{N}(1, 105^2), \\
Y &= h(A, R, M, H),
\end{aligned} \tag{A22}$$

where function h expresses a logistic regression model that provides the following conditional distribution:

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

Given such a synthetic dataset, we trained the same two-layered neural network classifier as the one described in Appendix G.1.1 by using 500 samples with the minibatch size 50. To evaluate the performance, we used 10,000 samples.

Table 4 presents the results, where the mean and the standard deviation of test accuracy and unfair effects are shown based on 10 experiments with randomly generated training and test datasets. Note that we do not show the mean unfair effect and the upper bound on PIU in Table 4 since we cannot fairly compare their values for each method. This is because they are computed based on the estimators of marginal potential outcome probabilities that are different between **Proposed**_{ex} and the two methods **Proposed** and **FIO**. Specifically, we used the lower and upper bounds on marginal probabilities (A18) and (A19) for **Proposed**_{ex} while we used the estimator presented in (A1) for **Proposed** and **FIO**.

Since **Proposed**_{ex} imposes a stronger fairness constraint than **Proposed**, its test accuracy was slightly lower than the other three methods. However, since it uses more reliable estimators of marginal potential outcome probabilities than **Proposed** and **FIO**, its PIU values and standard deviation in the conditional mean unfair effects were closer to zero than these methods. These results show the effectiveness of our proposed extension for dealing with a case with latent confounders.

As described in Appendix F, estimating marginal potential outcome probabilities in such a case remains an open problem. However, if their reliable estimators are given, our proposed extension will enable us to utilize them to achieve fairness for each individual in such complex cases.

Table 4: Accuracy and unfair effect on synthetic data with latent confounder. The closer unfair effects are to zero, the fairer predicted decisions are. Results are shown by (mean \pm standard deviation), computed based on 10 runs with randomly generated different datasets.

Method	Classifier	Test accuracy (%)	Unfair effects	
			PIU	Std in conditional means
Proposed_{ex}		90.9 \pm 2.7	$(1.66 \pm 1.40) \times 10^{-3}$	$(4.52 \pm 2.99) \times 10^{-2}$
Proposed	Logistic	91.3 \pm 2.5	$(1.68 \pm 1.30) \times 10^{-3}$	$(6.75 \pm 4.21) \times 10^{-2}$
FIO		93.0 \pm 0.4	$(8.06 \pm 2.67) \times 10^{-3}$	$(8.76 \pm 3.11) \times 10^{-2}$
Unconstrained		93.9\pm 0.5	$(1.03 \pm 0.33) \times 10^{-2}$	0.104 \pm 0.023
Proposed_{ex}		92.1 \pm 0.5	$(1.25 \pm 1.79) \times 10^{-9}$	$(2.13 \pm 2.12) \times 10^{-9}$
Proposed	NN	92.3 \pm 1.0	$(0.69 \pm 2.58) \times 10^{-5}$	$(0.66 \pm 3.49) \times 10^{-4}$
FIO		92.5 \pm 0.6	$(0.43 \pm 1.94) \times 10^{-3}$	$(2.59 \pm 4.16) \times 10^{-2}$
Unconstrained		94.4\pm 1.0	$(4.77 \pm 1.67) \times 10^{-2}$	0.103 \pm 0.050