Learning Individually Fair Classifier with Causal-Effect Constraint

Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino
NTT Communication Science Laboratories

February 18, 2020

Abstract

Machine learning is increasingly being used in various applications that make decisions for individuals. For such applications, we need to strike a balance between achieving good prediction accuracy and making fair decisions with respect to a sensitive feature (e.g., race or gender), which is difficult in complex real-world scenarios. Existing methods measure the unfairness in such scenarios as *unfair causal effects* and constrain its mean to zero. Unfortunately, with these methods, the decisions are not necessarily fair for all individuals because even when the mean unfair effect is zero, unfair effects might be positive for some individuals and negative for others, which is discriminatory for them. To learn a classifier that is fair for all individuals, we define unfairness as the *probability of individual unfairness* (PIU) and propose to solve an optimization problem that constrains an upper bound on PIU. We theoretically illustrate why our method achieves individual fairness. 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. Such applications require us to make decisions that are not discriminatory with respect to a sensitive feature, e.g., gender, race, or sexual orientation. This requirement is indispensable because data often exhibit a discriminatory bias including a correlation between the decision outcome and a sensitive feature if they include the records of past discriminatory decisions made by humans.

Although many researchers have studied how to make fair decisions while achieving high prediction accuracy Feldman et al. [2015], Hardt et al. [2016], 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., construction). While it is discriminatory to reject applicants because of their gender, since the job requires physical strength, it is sometimes **not** discriminatory to reject them because of their physical strength. Since physical strength is affected by gender, rejecting applicants because of physical strength leads to a difference in the rejection rates for men and women. Although such a difference is **not** unfair, it is eliminated when using traditional methods that are designed to reject male and female applicants at an equal rate (e.g., Feldman et al. [2015]). This indicates that these methods might provide the same decision outcome for a man and a woman with substantially different degrees of physical strength, which may significantly reduce the prediction accuracy.

The goal of this paper is to achieve high prediction accuracy while removing only an unfair difference in decision outcomes. To quantify such an unfair difference in complex real-world scenarios, we utilize the notion of *unfair causal effects*, which measures the unfairness for each individual. In fact, several methods have already been proposed that constrain the mean unfair effect across multiple individuals to be zero Nabi and Shpitser [2018], Chiappa and Gillam [2019]. However, imposing such a constraint does not ensure that the decisions are fair for every individual. This is because even when the mean unfair effect value is zero, unfair effect values might be largely positive for some individuals and largely negative for others, which indicates that the decisions are seriously discriminatory for these individuals.

Our contribution is to build an approach that guarantees the classifier to be fair for all individuals in complex real-world scenarios. To achieve this, we define the *probability of individual unfairness* (PIU), i.e., the probability that an unfair effect is not zero for an individual, and consider an optimization problem that constrains PIU to zero. Unfortunately, it is impossible to directly solve such an optimization problem because PIU cannot be consistently estimated from data. To overcome this difficulty, we upper bound the PIU value by a quantity that can be consistently estimated from data. Then we solve an alternative optimization problem that constrains the upper bound to be close to zero. We theoretically discuss why our constraint can guarantee fairness for all individuals. The experimental results show that at a slight cost of prediction accuracy, the unfair effects of our method become much closer to zero for all individuals than those of the existing method, thus demonstrating the effectiveness of our fairness constraint.

2 Background

To judge whether decisions are fair, we utilize the following concept of causal effects.

2.1 Unfair Causal Effects

To define causal effects, we use a directed acyclic graph (DAG) called a *causal graph*, whose nodes and edges represent random variables and causal relationships, respectively. Although a causal graph may not be provided in advance, we can infer it from data by using existing causal discovery methods (e.g., Spirtes et al. [2000]).

For instance, when making hiring decisions for physically demanding jobs, a causal graph can be depicted as shown in Figure 1 where A, M, Q, and Y denote random variables that respectively represent gender (0 for female and 1 for male), physical strength, qualification information, and decision outcome (0 for rejection and 1 for acceptance).

Taking A as a sensitive feature, we assume that it is fair to reject applicants based on physical strength M, but unfair to reject them based on gender A. The latter leads to a difference in decision outcome Y between men and women, which we consider as unfair causal effects (or unfair effects for short).

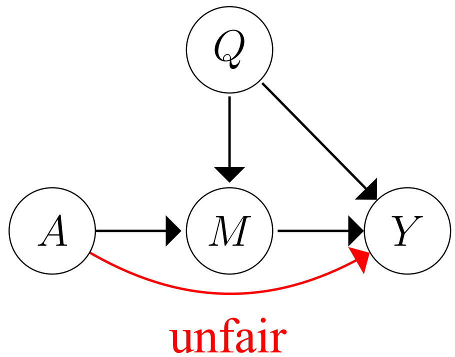

Figure 1: Causal graph representing a scenario of making hiring decisions for physically demanding jobs.

Such an unfair effect is represented as the *(natural) direct causal effect* along the pathway $A \to Y$ Petersen et al. [2006]. This is defined as the difference between the decision outcomes for each individual under the two *counterfactual situations* where gender A is changed differently (in one situation, the gender is changed to male, and in the other situation, to female) while physical strength M is set in the same way to the one when the gender is changed to female ($A = 0$), which is denoted by the variable $M(0)$. Such a decision outcome is represented by a random variable called a *potential outcome*. Let $Y(0)$ be the potential outcome where the gender is changed to $A = 0$ (female) and $Y(1, M(0))$ be the potential outcome where the gender is changed to $A = 1$ (male) while physical strength is set to $M(0)$. Then, the unfair effect is defined as $Y(1, M(0)) - Y(0)$, which quantifies how greatly the decision outcome for an individual is influenced just by changing gender from female to male.[1]

Although we can simply remove this direct effect by making decisions without A, such a naive strategy cannot be applied for a complex causal graph that contains multiple unfair pathways from A to Y (see, e.g., Figure 3 in Section 6.2).

Even in such cases, we can define unfair effects by introducing potential outcomes Shpitser [2013]. By letting potential outcomes be $Y_{[0]}$ and $Y_{[1]}$, the unfair effect for an individual is defined as $Y_{[1]} - Y_{[0]}$, which becomes zero if the decision is fair and non-zero otherwise.

For multiple individuals, the mean unfair effect is represented as the difference between marginal probabilities:

$$\mathbb{E}_{Y_{[0]},Y_{[1]}}[Y_{[1]} - Y_{[0]}] = \mathbb{E}_{Y_{[1]}}[Y_{[1]}] - \mathbb{E}_{Y_{[0]}}[Y_{[0]}] = \mathrm{P}(Y_{[1]} = 1) - \mathrm{P}(Y_{[0]} = 1) \tag{1}$$

where the first equality holds because of the linearity of the expectation, and the second holds since $Y_{[0]}, Y_{[1]} \in \{0, 1\}$.

[1] We can also measure the direct effect when changing the gender from male to female as $Y(1) - Y(0, M(1))$, where $M(1)$ denotes the physical strength when the gender is changed to male.

2.2 Marginal Potential Outcome Probabilities

Although it is impossible to observe potential outcome variables, the marginal potential outcome probabilities in (1) can be represented as the distributions of the observed data if several conditional independence relations hold between potential outcomes and observed variables Pearl [2009].

For instance, given the causal graph in Figure 1, we can use the marginal probability estimator presented in Huber [2014], which is formulated as

$$\begin{aligned} \mathrm{P}(Y_{[1]} = 1) &= \mathbb{E}_{M,Q}[w\,\mathrm{P}(Y = 1|A = 1, M, Q)] \\ \mathrm{P}(Y_{[0]} = 1) &= \mathbb{E}_{M,Q}[w\,\mathrm{P}(Y = 1|A = 0, M, Q)]. \end{aligned} \tag{2}$$

Here,

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

denotes a weight depending on conditional distributions $\mathrm{P}(A|M, Q)$ and $\mathrm{P}(A|Q)$, which can be estimated by fitting some parametric model (e.g., logistic regression) to the data. We can also express it as $w = \frac{\mathrm{P}(M|A=0,Q)}{\mathrm{P}(M|Q)}$ by using Bayes' theorem. Intuitively, with this weight, we can reduce the difference in physical strength M between men and women because the weight value becomes large for individuals whose M values are likely when $A = 0$ (female) and small for those whose M values are not likely when $A = 0$.

By computing marginal probabilities in this way, we can consistently estimate the mean unfair effect in (1), which is used as a measure of unfairness in the existing methods.

2.3 Weakness of Existing Methods

Since the mean unfair effect can be consistently estimated, several methods aim to learn fair predictive models by constraining the mean unfair effect to zero Zhang et al. [2017], Nabi and Shpitser [2018], Chiappa and Gillam [2019].

Unfortunately, just because the mean unfair effect is zero, unfair effects are not necessarily zero for all individuals. This is because their values might be largely positive for some individuals and largely negative for others, which indicates that the decisions are seriously discriminatory for these individuals.

To guarantee fairness for all individuals, instead of the mean unfair effect (1), we introduce a quantity called PIU, which is described in the next section.

3 Proposed Method

3.1 Quantifying Individual Unfairness

To quantify unfairness, we define the following notion:

Definition 1 (Probability of Individual Unfairness) *Let $Y_{[0]}, Y_{[1]} \in \{0, 1\}$ be binary potential outcome variables. We define the probability of individual unfairness (PIU) by $\mathrm{P}(Y_{[0]} \neq Y_{[1]})$.*

PIU is the probability that the potential outcomes (i.e., the decision outcomes in counterfactual situations for each individual) are different, which is given by using the joint distribution of $Y_{[0]}$ and $Y_{[1]}$ as described in Section 3.2.

We emphasize that PIU is different from the mean unfair effect because PIU is non-negative and takes zero only when the unfair effect is zero for all individuals. To illustrate this, we present a simple example with two individuals. Let the potential outcomes for one individual be $Y_{[0],1}$ and $Y_{[1],1}$ and those for the other be $Y_{[0],2}$ and $Y_{[1],2}$. Suppose that one has the potential outcomes $(Y_{[0],1},\, Y_{[1],1}) = (0,\, 1)$ and that the other has $(Y_{[0],2},\, Y_{[1],2}) = (1,\, 0)$, which is discriminatory for both. However, despite such unfairness, since the unfair effects (i.e., differences between potential outcomes) are 1 and -1 for the two individuals, respectively, the mean unfair effect is 0. In contrast, since the potential outcomes are different for both, the PIU value is 1, which indicates that the decisions are discriminatory for them. This demonstrates that PIU can appropriately quantify the unfairness.

3.2 Difficulty in Dealing with PIU

By using PIU, we aim to learn a classifier that is fair for all individuals. We consider a binary classification problem of deciding outcome $Y \in \{0, 1\}$ for individuals. Here, each individual has set of features $\mathbf{X}$ that includes a sensitive feature $A \in \{0, 1\}$. Let h_θ be a probabilistic classifier with parameter θ that outputs the conditional probability $\mathrm{P}(Y = 1|\mathbf{X})$, and let l be a loss function for measuring the accuracy of the classifier h_θ. As training data, we use records on the past decisions for n individuals, which are the observations of $(\mathbf{X}, Y)$, i.e., $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$.

To formulate an optimization problem of training the classifier, let us assume that the true joint distribution of potential outcomes is given as P_θ, which is expressed with classifier parameter θ. Then we can find θ that minimizes the prediction loss and makes unfair effects to be (almost) zero for each individual by solving the following optimization problem:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^{n} l(h_\theta(\mathbf{x}_i), y_i) \\ \text{subject to} \quad & \mathrm{P}_\theta(Y_{[0]} \neq Y_{[1]}) \leq \delta \end{aligned} \tag{3}$$

where $\delta \in [0, 1]$ is the hyperparameter. The constraint requires the decisions made by h_θ to be fair in the sense that the PIU value $\mathrm{P}_\theta(Y_{[0]} \neq Y_{[1]})$ is at most δ.

However, we cannot directly solve the optimization problem (3) because the true joint distribution P_θ is unavailable. Specifically, to infer the joint distribution from data, we need the two decision outcomes for each individual: one is the outcome when the sensitive feature $A = 0$, and the other is the outcome when $A = 1$ Fan et al. [2017], Firpo and Ridder [2019]. However, the observed data contain one of the two decision outcomes but not both. For instance, in the case where A represents gender, if an individual is female, then we cannot obtain the decision outcome where the individual is male. Due to such a lack of decision outcome records, we cannot obtain a reliable estimator of the PIU value; if we use an unreliable estimator of the PIU value, we might incorrectly regard unfair decisions as fair.

To address this issue, we formulate a **stronger** constraint than the one in the optimization problem (3) such that if classifier parameter θ satisfies it, then θ also satisfies the constraint in (3) for **any** joint distribution of potential outcomes. To achieve this, we utilize an upper bound, which can be estimated from data, on the PIU value.

3.3 Constraining and Penalizing PIU

3.3.1 Upper bound formulation

To obtain a stronger fairness constraint than the one in (3), we derive an upper bound that can be estimated from data.

To this end, we use the independent distribution of potential outcomes $Y_{[0]}$ and $Y_{[1]}$, i.e., $\mathrm{P}^I_\theta(Y_{[0]}, Y_{[1]}) = \mathrm{P}_\theta(Y_{[0]})\,\mathrm{P}_\theta(Y_{[1]})$, where $\mathrm{P}_\theta(Y_{[0]})$ and $\mathrm{P}_\theta(Y_{[1]})$ denote marginal distributions.

Although using such a joint distribution implies ignoring the correlation between potential outcomes, the following theorem guarantees that this provides a reasonable upper bound on PIU:

Theorem 1 (Upper bound on PIU) *For any joint distribution P_θ, PIU in (3) is upper bounded by*

$$\mathrm{P}_\theta(\mathrm{Y}_{[0]} \neq \mathrm{Y}_{[1]}) \leq 2\,\mathrm{P}^I_\theta(\mathrm{Y}_{[0]} \neq \mathrm{Y}_{[1]}) \tag{4}$$

where $\mathrm{P}^I_\theta(\mathrm{Y}_{[0]} \neq \mathrm{Y}_{[1]})$ is PIU that is approximated with the independent joint distribution $\mathrm{P}^I_\theta = \mathrm{P}_\theta(Y_{[0]})\,\mathrm{P}_\theta(Y_{[1]})$.

The proof is detailed in Appendix A. Theorem 1 states that whatever joint distribution the potential outcomes $Y_{[0]}$ and $Y_{[1]}$ follow, the resulting PIU value is at most twice the PIU value that is approximated with P^I_θ.

Unlike PIU in (3), the upper bound in (4) can be consistently estimated. Let us use the notations $\hat{p}^{[0]}_\theta$ and $\hat{p}^{[1]}_\theta$ as short forms for the estimated marginal probabilities $\hat{\mathrm{P}}_\theta(Y_{[0]} = 1)$ and $\hat{\mathrm{P}}_\theta(Y_{[1]} = 1)$, respectively. Then since the approximated PIU value equals the probability that the potential outcomes are $(Y_{[0]}, Y_{[1]}) = (0,1)$ or $(1,0)$, the upper bound in (4) can be estimated as

$$2\hat{\mathrm{P}}^I_\theta(\mathrm{Y}_{[0]} \neq \mathrm{Y}_{[1]}) = 2(\hat{p}^{[1]}_\theta(1 - \hat{p}^{[0]}_\theta) + (1 - \hat{p}^{[1]}_\theta)\hat{p}^{[0]}_\theta). \tag{5}$$

Marginal probabilities $\hat{p}^{[0]}_\theta$ and $\hat{p}^{[1]}_\theta$ in (5) can be estimated from the outputs of classifier h_θ. For instance, when the causal graph in Figure 1 is given, where the features of each individual are $\mathbf{X} = \{A, M, Q\}$, by plugging classifier $h_\theta(\mathbf{X}) = \mathrm{P}(Y = 1|\mathbf{X})$ into (2) and approximating it with the empirical distribution, we can obtain the following weighted estimators:

$$\hat{p}^{[0]}_\theta = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i h_\theta(1, m_i, q_i)$$
$$\hat{p}^{[1]}_\theta = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i h_\theta(0, m_i, q_i),$$

where each weight is given as

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

We can estimate conditional distributions $\hat{\mathrm{P}}(A = 0|m_i, q_i)$ and $\hat{\mathrm{P}}(A = 0|q_i)$ by fitting a parametric model (e.g., logistic regression or neural network) to the data beforehand.

3.3.2 Proposed objective function

Since the PIU value is always smaller than its upper bound, by constraining the upper bound (5) to be at most δ, we can find classifier parameter θ that satisfies the constraint in (3). Specifically, we consider the following problem:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \frac{1}{n}\sum_{i=1}^{n} l(h_\theta(\mathbf{x}_i), y_i) \\ \text{subject to} \quad & 2(\hat{p}_\theta^{[1]}(1-\hat{p}_\theta^{[0]}) + (1-\hat{p}_\theta^{[1]})\hat{p}_\theta^{[0]}) \le \delta. \end{aligned} \tag{6}$$

To simplify (6), we add a penalty term with hyperparameter $\lambda \ge 0$ to the loss function, which gives the following unconstrained minimization problem:

$$\underset{\theta}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n} l(h_\theta(\mathbf{x}_i), y_i) + \lambda(\hat{p}_\theta^{[1]}(1-\hat{p}_\theta^{[0]}) + (1-\hat{p}_\theta^{[1]})\hat{p}_\theta^{[0]}). \tag{7}$$

To minimize the objective function (7), we used Adam Kingma and Ba [2015] in experiments (Section 6). We discuss the convergence guarantees when minimizing (7) in Appendix B.

3.4 Why Achieve Individual Fairness?

If we set $\lambda = \infty$ and solve the minimization problem (7), our method guarantees that the unfair effects become zero for all individuals. To demonstrate this, we present an example with two individuals.

In this example, an unfair effect (i.e., the difference in potential outcomes) is zero for both individuals only when $(Y_{[0],1}, Y_{[1],1}, Y_{[0],2}, Y_{[1],2}) = (0,0,0,0)$, $(1,1,1,1)$, $(0,0,1,1)$, and $(1,1,0,0)$, which implies that the marginal probabilities are $(\hat{p}_\theta^{[0]}, \hat{p}_\theta^{[1]}) = (0,0)$, $(1,1)$, $(0.5,0.5)$, and $(0.5,0.5)$, respectively.

In fact, to make an unfair effect to be zero for both individuals, it is sufficient for the marginal probabilities to be $(\hat{p}_\theta^{[0]}, \hat{p}_\theta^{[1]}) = (0,0)$ or $(1,1)$. With these probabilities, all the potential outcomes always take the same value, i.e., $(Y_{[0],1}, Y_{[1],1}, Y_{[0],2}, Y_{[1],2}) = (0,0,0,0)$ or $(1,1,1,1)$, which indicates that unfair effects are zero for both individuals. Note that if we allow $(\hat{p}_\theta^{[0]}, \hat{p}_\theta^{[1]})$ to be $(0.5, 0.5)$, decisions can be unfair because the potential outcome values can be $(Y_{[0],1}, Y_{[1],1}, Y_{[0],2}, Y_{[1],2}) = (0,1,1,0)$ or $(1,0,0,1)$; in these cases the unfair effects are 1 for one individual and -1 for the other, which indicates that the decisions are discriminatory for both.

Therefore, to obtain classifier parameter θ such that unfair effects are zero for all individuals, we only have to limit the marginal probabilities to be $(\hat{p}_\theta^{[0]}, \hat{p}_\theta^{[1]}) = (0,0)$ or $(1,1)$, which is what our method does when using $\lambda = \infty$ in (7).

In contrast, as will be described in Section 4.1, the existing methods can only limit the marginal probabilities to the same value, i.e., $\hat{p}_\theta^{[0]} = \hat{p}_\theta^{[1]}$, which is insufficient to guarantee the individual fairness as illustrated above.

3.5 Extension for Dealing with Unobserved Variables

So far, to estimate the upper bound on PIU, we have assumed that the marginal potential outcome distributions can be inferred from the data distributions. However, this assumption does not hold particularly when observed variables are affected by unobserved variables called *latent confounders* Pearl [2009], which is possible in real-world scenarios.

Although inferring marginal probabilities in such scenarios is extremely challenging in the field of causal inference, existing studies have shown that we can infer the lower and upper bounds on marginal probabilities in some cases (e.g., Miles et al. [2017]).

If such bounds are available, we can obtain an upper bound on PIU even in the presence of unobserved variables. Let true marginal probabilities $p^{[0]}$ and $p^{[1]}$ satisfy $\hat{l}_\theta^{[0]} \leq p^{[0]} \leq \hat{u}_\theta^{[0]}$ and $\hat{l}_\theta^{[1]} \leq p^{[1]} \leq \hat{u}_\theta^{[1]}$, respectively; as an example, we present the formulations of $\hat{l}_\theta^{[0]}$, $\hat{u}_\theta^{[0]}$, $\hat{l}_\theta^{[1]}$, and $\hat{u}_\theta^{[1]}$ based on Miles et al. [2017] in Appendix C. Given such bounds, we can simply reformulate our upper bound constraint in (6) as

$$2(\hat{u}_\theta^{[1]}(1-\hat{l}_\theta^{[0]}) + (1-\hat{l}_\theta^{[1]})\hat{u}_\theta^{[0]}) \leq \delta. \tag{8}$$

By using a penalty term derived from (8), as with (7), we can formulate an unconstrained optimization problem. Note that if the penalty term is derived from the bounds of Miles et al. [2017], the optimization problem is not easy to solve because such a penalty term is not differentiable. Leveraging other bounds on marginal probabilities to formulate an objective function that is differentiable and easy to optimize constitutes our future work.

4 Comparison of Proposed Fairness Constraints with Existing Methods

To ensure that the classifier is fair for all individuals, we have proposed a novel fairness constraint.

In what follows, we compare the constraint with that of the existing methods Zhang et al. [2017], Nabi and Shpitser [2018]. Although our method does not directly impose the constraint but uses the penalty term derived from it, the comparison of the constraints enables us to clarify why our method is effective.

4.1 Formulation of Fairness Constraints

As shown in (6), letting $\hat{p}_\theta^{[0]}$ and $\hat{p}_\theta^{[1]}$ denote the marginal potential outcome probabilities, our constraint is expressed by

$$2(\hat{p}_\theta^{[1]}(1-\hat{p}_\theta^{[0]}) + (1-\hat{p}_\theta^{[1]})\hat{p}_\theta^{[0]}) \leq \delta \tag{9}$$

where $\delta \in [0,1]$ is a hyperparameter.

On the other hand, the constraint of the existing methods Zhang et al. [2017], Nabi and Shpitser [2018] limits the absolute value of the mean unfair effect (1) to be at most $\delta' \in [0,1]$, i.e., we have

$$|\hat{p}_\theta^{[1]} - \hat{p}_\theta^{[0]}| \leq \delta', \tag{10}$$

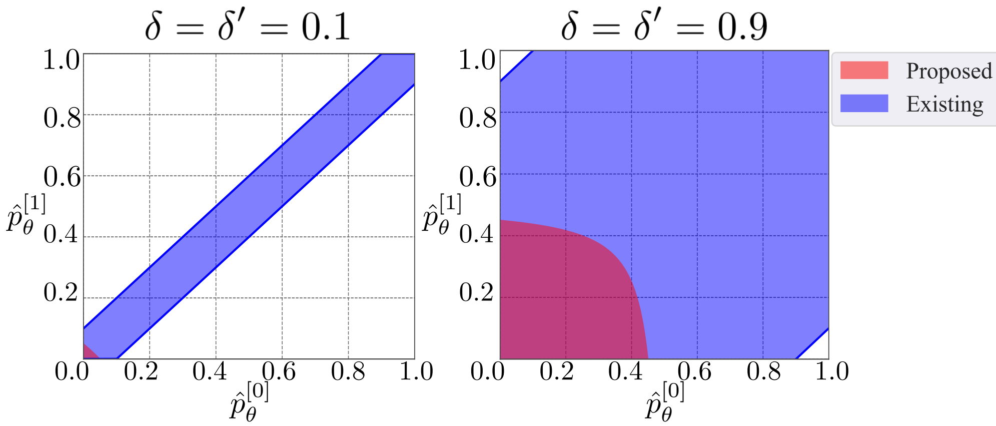

Figure 2: Feasible regions for our fairness constraint (red) and the existing methods (blue) when $\delta = \delta'$ $= 0.1, 0.9$.

which constrains the marginal probabilities to be $\hat{p}_\theta^{[0]} = \hat{p}_\theta^{[1]}$ when $\delta' = 0$.

4.2 Difference in Feasible Regions

Based on (9) and (10), we compare the feasible region of the constraint of our method (i.e., the set of possible marginal probabilities $\hat{p}_\theta^{[0]}$ and $\hat{p}_\theta^{[1]}$ that satisfy our constraint) with that of the existing methods. For simplicity, we set the hyperparameters in both constraints to the same value, i.e., $\delta = \delta'$; we will make a remark on this later.

By graphing the hyperbolic inequality (9) and the linear inequality (10), we show the feasible region for each fairness constraint in Figure 2. We can find that our constraint (9) is stronger than that of the existing methods (10); that is, the feasible region of the constraint of our method (red region) is a subset of that of the existing methods (blue region).

Especially when $\delta = \delta' = 1$, our method constrains the feasible regions to $0 \le \hat{p}_\theta^{[0]} \le 0.5$ and $0 \le \hat{p}_\theta^{[1]}$ ≤ 0.5 and to $0.5 \le \hat{p}_\theta^{[0]} \le 1$ and $0.5 \le \hat{p}_\theta^{[1]} \le 1$, while the existing methods regard any $(\hat{p}_\theta^{[0]}, \hat{p}_\theta^{[1]}) \in [0,1]^2$ as feasible.

A remark is needed regarding the values of the hyperparameters δ and δ'. Recall that the upper bound on PIU in (9) and the absolute value of the mean unfair effect in (10) are restricted to be at most δ and δ', respectively. It is reasonable to let $\delta' \in [0, 1]$ because the absolute value of the mean unfair effect is at most 1. In contrast, it may seem unreasonable to let $\delta \in [0, 1]$ by setting $\delta = \delta'$ because the upper bound on PIU can become up to 2, which is attained with $(\hat{p}_\theta^{[0]}, \hat{p}_\theta^{[1]}) = (0, 1)$ or $(1, 0)$.

Nevertheless, we do not consider the case when $\delta \in (1, 2]$. Recall that our purpose is to constrain the PIU value by limiting its upper bound to be at most δ. To achieve this, we need to use $\delta \in [0, 1]$ because $\delta \in (1, 2]$ imposes no constraint on the PIU value, which is at most 1 and is always less than $\delta \in (1, 2]$.

4.3 Advantages of Proposed Fairness Constraint

From the feasible regions in Figure 2, we can see that our constraint effectively excludes the *unfair regions.*

Our constraint never accepts the two points $(\hat{p}_\theta^{[0]}, \hat{p}_\theta^{[1]}) = (0, 1)$ or $(1, 0)$, where the unfair effect is always 1 or -1 for all individuals. Moreover, since our constraint is formulated by a hyperbolic inequality, for any $\delta \in [0, 1]$, it always rejects the region that is close to those two points, i.e., either $0 \leq \hat{p}_\theta^{[0]} \leq 0.5$ and $0.5 \leq \hat{p}_\theta^{[1]} \leq 1$, or $0.5 \leq \hat{p}_\theta^{[0]} \leq 1$ and $0 \leq \hat{p}_\theta^{[1]} \leq 0.5$, where the unfair effects are likely to be deviated from zero.

The above facts demonstrate that by formulating the fairness constraint as the hyperbolic inequality (9), our method can effectively find the value of θ that is fair for all individuals.

5 Related Work

Motivated by a recent development in inferring causal graphs from data Chikahara and Fujino [2018], Li et al. [2019], several causality-based approaches have been proposed.

These approaches can be divided into two types with respect to problem settings: one addresses cases where *all* the causal pathways from the sensitive feature to the decision outcome are unfair Kusner et al. [2017], Russell et al. [2017], and the other deals with scenarios where only *some of* the pathways are unfair Zhang et al. [2017], Kilbertus et al. [2017], Nabi and Shpitser [2018], Zhang et al. [2018], Zhang and Bareinboim [2018b,a], Chiappa and Gillam [2019], Wu et al. [2019].

Although the latter approaches can address complex real-world scenarios, none of them can guarantee that decisions are fair for all individuals. Several methods such as Wu et al. [2019] impose a strong constraint, which ensures that the mean unfair effect is zero for each *subgroup* of individuals who have the same value of the features $\mathbf{X}$. However, this approach is also insufficient to guarantee that unfair effect values are zero for all individuals if any of the subgroups consists of multiple individuals, which is often the case with real-world scenarios.

We propose a novel approach that guarantees individual fairness in the latter setting. To accomplish this, we derive an upper bound on PIU for solving an optimization problem that constrains it to zero.

Several studies have derived the upper bounds on the functionals of the joint distribution of potential outcomes Fan et al. [2017], Firpo and Ridder [2019], which are defined in a similar way to PIU. However, these bounds are different from ours because they are designed for continuous potential outcomes while ours is derived for binary ones.

Our upper bound is closely related to the *correlation gap* Agrawal et al. [2010]. When the joint distribution of random variables is uncertain because of, for example, small sample size, the correlation gap is used to see how closely we can approximate the expected value by using an independent joint distribution such as P_θ^I. This is defined as the worst-case ratio of the expected value with respect to the uncertain joint distribution to the one with respect to the independent joint distribution, which can be upper bounded by a constant in some cases.

When approximating PIU, Theorem 1 guarantees that the constant is 2 because PIU can be expressed as $\mathbb{E}_{Y_{[0]},Y_{[1]}}[\mathbf{1}(Y_{[0]} \neq Y_{[1]})]$, where $\mathbf{1}$ is an indicator function that takes 1 if $Y_{[0]} \neq Y_{[1]}$ and 0 otherwise. Although Rubinstein and Singla [2017] proposed an upper bound that can be used in more general cases, since their constant is 200, Theorem 1 provides a tighter upper bound on PIU.

By utilizing the tight upper bound, we can avoid imposing an unnecessarily severe constraint. This is because if the upper bound is loose (i.e., if the upper bound can be too large when the PIU value is small), the constraint on the upper bound, which we use to ensure that the PIU value is close to zero, needs to be excessively severe. In contrast, by using a tight upper bound, we can avoid imposing such a severe constraint, which prevents an unnecessary decrease in the prediction accuracy.

6 Experiments

We evaluated the performance of our method through synthetic and real-world data experiments, whose settings have been detailed in Appendix D.1.

6.1 Synthetic Data Experiments

We prepared synthetic data representing a scenario of making hiring decisions for physically demanding jobs, as detailed in Section 2. We sampled gender $A \in \{0, 1\}$, qualification Q, physical strength M, and hiring decision outcome $Y \in \{0, 1\}$ from the following model:

$$\begin{aligned} &A \sim \text{Bernoulli}(0.6), \\ &Q \sim \mathcal{N}(5, 2.5^2), \\ &M = 3A + \lfloor \varepsilon_M \rfloor, \quad \varepsilon_M \sim \mathcal{N}(3, 1.5^2), \quad \text{and} \\ &Y \sim \text{Bernoulli}(\sigma(-10 + 5A + M + \lfloor Q \rfloor)). \end{aligned} \tag{11}$$

Here, Bernoulli and $\mathcal{N}$ represent the Bernoulli and Gaussian distributions, respectively, $\lfloor \cdot \rfloor$ is the floor function, which returns an integer by removing the decimal places, and $\sigma(x) = 1/(1 + \exp(-x))$ is the standard sigmoid function. Note that in (11), the decision outcome Y is sampled such that the conditional expectation of Y given A, Q, and M is expressed by a logistic regression model.

Using these synthetic data, we evaluated the performance of our method. We used $5,000$ samples to train classifier h_θ and $1,000$ samples to evaluate the test accuracy.

To show that our method learns an individually fair classifier, regarding A as a sensitive feature, we evaluated PIU and its upper bound. In fact, with synthetic data, we can evaluate the PIU value because we can sample the potential outcomes by using the true data-generating model (11), which is called a *structural equation model* (SEM) Pearl [2009]. Note that this is not possible with real-world data because such an SEM is unavailable.

We evaluated the PIU values by sampling potential outcomes $Y_{[0]}$ and $Y_{[1]}$, which are represented as $Y_{[0]} = Y(0)$ and $Y_{[1]} = Y(1, M(0))$, respectively. To achieve this, following Pearl [2009],[2] we sampled $M(0)$ (i.e., the physical strength when gender A is changed to 0 (female)) from the model of M in

[2]It is necessary to sample the random variables (called mediators Pearl [2009]) that correspond to the descendant nodes of A because these are affected by a change in A.

(11) with $A = 0$; i.e., $M(0) \sim \varepsilon_M$. With such samples of $M(0)$, we sampled the potential outcomes as follows:[3]

$$Y_{[0]} \sim h_\theta(0, M(0), Q), \quad Y_{[1]} \sim h_\theta(1, M(0), Q).$$

Then using the n pairs of these samples $\{(y_{[0],1}, y_{[1],1}), \ldots, (y_{[0],n}, y_{[1],1})\}$, we evaluated the PIU value by

$$\hat{\mathrm{P}}_\theta(Y_{[0]} \neq Y_{[1]}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_{[0],i} \neq y_{[1],i}).$$

where $\mathbf{1}$ is an indicator function that takes 1 if $y_{[0],i} \neq y_{[1],i}$ $(i \in \{1, \ldots, n\})$ and 0 otherwise.

Table 1a presents the test accuracy, the PIU value, and its upper bound. We show the means and the standard deviations in 10 experiments with randomly generated training and test datasets. Here, we compared the performance of our method (**Proposed**) with an existing method called *fair inference on outcomes* (**FIO**) Nabi and Shpitser [2018].[4] As regards the classifiers, following their implementation, we used logistic regression for **FIO** while we used both logistic regression and neural network for **Proposed**. We used a grid search with 0.1 grid size to select the penalty parameter value of **Proposed** from $0 \leq \lambda \leq 10.0$ while we set the hyperparameter values of **FIO** following the original paper.

As expected, since **FIO** only constrains the mean unfair effect, even though the mean value was close to zero, the PIU value (0.164) was much larger than the value of 2.30×10^{-3} of our method (**Proposed** (logistic)). The PIU value of 0.164 indicates that the decision outcomes for each individual in the two counterfactual situations (where gender is changed differently and physical strength is given in the same way) become different with a probability of 0.164. This demonstrates that constraining only the mean unfair effect is insufficient to learn a classifier that is fair for all individuals.

In contrast, by using our method, both the mean unfair effect and the PIU value were close to zero, which indicates that the learned classifier is fair for all individuals. Furthermore, **Proposed** (logistic) achieved a test accuracy of 72.1%, which is comparable to that of **FIO** (73.4%). Moreover, thanks to the use of neural network, our **Proposed** (DNN) yielded a high test accuracy of 85.0% while achieving a small PIU value.

As discussed in Section 3.4, our method achieves a small PIU value because our constraint makes the marginal potential outcome probabilities ($\hat{p}_\theta^{[0]}$, $\hat{p}_\theta^{[1]}$) close to $(0, 0)$ or $(1, 1)$. We confirmed this in all of 10 experiments with different data. For instance, in one of these experiments, we obtained ($\hat{p}_\theta^{[0]}$, $\hat{p}_\theta^{[1]}$) $= (0.999, 0.999)$ with **Proposed** (logistic) and $(0.985, 0.984)$ with **Proposed** (DNN) while the probabilities were $(0.716, 0.711)$ with **FIO**.

6.2 Real-World Data Experiments

To test our method, we used the Adult dataset[6] from the UCI repository Bache and Lichman [2013]. This dataset is US census data that contain the features of individuals including gender, occupation,

[3]As with the existing methods (e.g., Zhang and Bareinboim [2018a]), we used classifier h_θ as an SEM for predicted outcomes.

[4]https://github.com/raziehna/fair-inference-on-outcomes.

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

(a) Synthetic data

	Test Accuracy (%)	Unfair Causal Effects Mean	PIU	Upper bound on PIU
FIO (logistic)	73.4 $\pm$ 3.6	$(-1.64 \pm 2.79) \times 10^{-3}$	$(1.64 \pm 0.43) \times 10^{-1}$	$(7.24 \pm 1.90) \times 10^{-1}$
Proposed (logistic)	72.1 $\pm$ 4.3	$(\mathbf{2.89 \pm 3.65}) \times \mathbf{10^{-4}}$	$(\mathbf{2.30 \pm 1.25}) \times \mathbf{10^{-3}}$	$(\mathbf{4.99 \pm 2.10}) \times \mathbf{10^{-3}}$
Proposed (DNN)	**85.0 $\pm$ 2.1**	$(8.52 \pm 2.27) \times 10^{-4}$	$(1.89 \pm 0.79) \times 10^{-2}$	$(4.29 \pm 2.48) \times 10^{-2}$

(b) Real-world data (Adult dataset)

	Test Accuracy (%)	Unfair Causal Effects Mean	Upper bound on PIU
FIO (logistic)	72.2	-3.52×10^{-2}	5.01×10^{-1}
Proposed (logistic)	70.2	$\mathbf{-2.64 \times 10^{-11}}$	$\mathbf{6.00 \times 10^{-9}}$
Proposed (DNN)	**76.6**	5.95×10^{-4}	1.11×10^{-4}

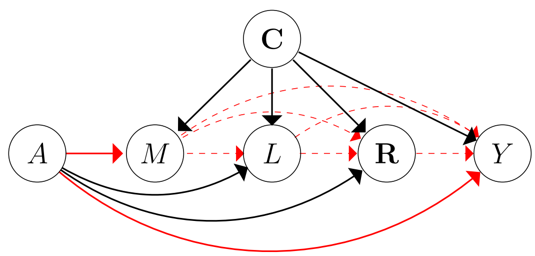

Figure 3: Causal graph for UCI Adult dataset. Unfair pathways are $A \to Y$ and $A \to M \to \cdots \to Y$.

and income.

Using this dataset, we evaluated the performance when predicting whether each individual has an annual income exceeding 50,000 US Dollars.

To quantify the fairness of the predicted decisions, following Nabi and Shpitser [2018], we used the causal graph shown in Figure 3, which includes gender A, marital status M, level of education L, occupation information (e.g., weekly working hours) $\mathbf{R}$, age and nationality $\mathbf{C}$, and income Y. Here A is regarded as a sensitive feature, and direct pathway $A \to Y$ and all the pathways from A to Y via M (i.e., $A \to M \to \cdots \to Y$) are considered as unfair. As with the causal graph in Figure 1, we defined the potential outcome variables and the estimators of their marginal probabilities, which have been detailed in Appendix D.2.

Unfortunately, unlike in the synthetic data experiments, we cannot measure the PIU values using the Adult dataset. Hence, we only evaluated the upper bound on PIU to quantify the fairness of the predicted decisions. When upper-bound values are close to zero, we can conclude that the PIU values

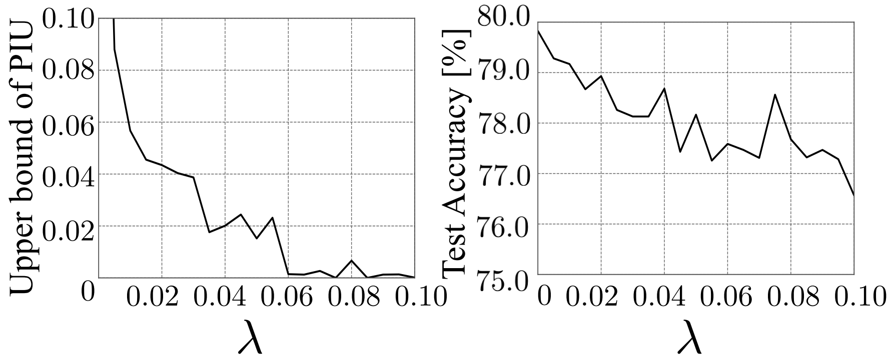

Figure 4: Variations in upper bound on PIU (left) and test accuracy (right) for the UCI Adult dataset with increasing penalty parameter value λ.

are also close to zero, although large upper-bound values do not always imply large PIU values.

Table 1b shows the results. The upper bound on PIU of **Proposed** (logistic) was much closer to zero than that of **FIO**. This implies that the PIU value of **Proposed** (logistic) was also close to zero; hence, the learned classifier is fair for all individuals. Furthermore, the test accuracy of **Proposed** (logistic) was 70.2%, which is comparable to that of **FIO** (72.2%). This demonstrates that our method learns a classifier that is fair for all individuals at a slight cost of prediction accuracy.

In addition, we investigated the effect of penalty parameter λ in (7) for **Proposed** (DNN). Figure 4 shows the variations in the upper bound on PIU for the test data and the test accuracy for $\lambda = 0$, 0.005, 0.010, $\cdots$, 0.100. As expected, as λ increases, the upper bound value approaches (almost) zero while the test accuracy decreases from 79.8% to 76.6%.

7 Conclusion

We have addressed the problem of learning a classifier that is fair for all individuals in complex real-world scenarios.

To learn such a classifier, we have introduced a non-negative quantity called PIU and have proposed to solve an optimization problem using a penalty term that reduces the upper bound on PIU. Although it is impossible to estimate PIU from data, our upper bound can be consistently estimated.

We have discussed why our method guarantees individual fairness, which we experimentally confirmed. We also have shown the extension to real-world scenarios with unobserved variables.

In experiments, although the existing method made decisions that were unfair for some individuals, our method made decisions that were fair for all individuals. These results demonstrate the effectivity of utilizing the fairness constraint that is based on the upper bound on PIU.

References

S. Agrawal, Y. Ding, A. Saberi, and Y. Ye. Correlation robust stochastic optimization. In *SODA*, pages 1087–1096, 2010.

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

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.

S. Chiappa and T. P. Gillam. Path-specific counterfactual fairness. In *AAAI*, pages 7801–7808, 2019.

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

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.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268, 2015.

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

M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *NeurIPS*, pages 3315–3323, 2016.

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

N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *NeurIPS*, pages 656–666, 2017.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. arXiv:1412.6980.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *NeurIPS*, pages 4066–4076, 2017.

H. Li, V. Cabeli, N. Sella, and H. Isambert. Constraint-based causal structure learning with consistent separating sets. In *NeurIPS*, pages 14257–14266, 2019.

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.

R. Nabi and I. Shpitser. Fair inference on outcomes. In *AAAI*, pages 1931–1940, 2018.

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

M. L. Petersen, S. E. Sinisi, and M. J. van der Laan. Estimation of direct causal effects. *Epidemiology*, pages 276–284, 2006.

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

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.

I. Shpitser. Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science*, 37(6):1011–1035, 2013.

P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search.* MIT press, 2000.

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.

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

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

L. Zhang, Y. Wu, and X. Wu. A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI*, 2017.

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 Proof of Theorem 1

We first introduce several notations. Let the marginal potential outcome probabilities, $\mathrm{P}_\theta(Y_{[0]} = 1)$ and $\mathrm{P}_\theta(Y_{[1]} = 1)$, be α and β, and their joint probabilities, $(Y_{[0]}, Y_{[1]}) = (0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$, be p_{00}, p_{01}, p_{10}, and p_{11}.

Then we have

$$\begin{aligned} &p_{10} + p_{11} = \alpha, \\ &p_{00} + p_{01} = 1 - \alpha, \\ &p_{01} + p_{11} = \beta, \qquad \text{and} \\ &p_{10} + p_{00} = 1 - \beta. \end{aligned} \tag{A1}$$

As described in the main manuscript, the right-hand side in (4) 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

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

Here since all the joint probabilities in (A1) are non-negative, p_{01} and p_{10} become at most $\min\{\beta,\ 1-\alpha\}$ and $\min\{\alpha,\ 1-\beta\}$, respectively, which yields

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

Hence, it suffices to prove

$$\min\{\beta, 1-\alpha\} + \min\{\alpha, 1-\beta\} \leq 2\beta(1-\alpha) + 2\alpha(1-\beta). \tag{A2}$$

Since both sides in (A2) 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 $\min\{\alpha, 1-\beta\} = \alpha$, (A2) is reduced to

$$\begin{aligned} &\beta + \alpha \leq 2\beta(1-\alpha) + 2\alpha(1-\beta) \\ &\iff \alpha + \beta - 4\alpha\beta \geq 0. \end{aligned} \tag{A3}$$

Since we have $\alpha + \beta \leq 1$ in this case, which implies $\alpha + \beta - (\alpha+\beta)^2 \geq 0$, the inequality (A3) can be proven as follows:

$$\begin{aligned} &\alpha + \beta - 4\alpha\beta \\ &= \alpha + \beta - (\alpha+\beta)^2 + (\alpha-\beta)^2. \\ &\geq 0. \end{aligned} \tag{A4}$$

Thus, we prove Theorem 1.

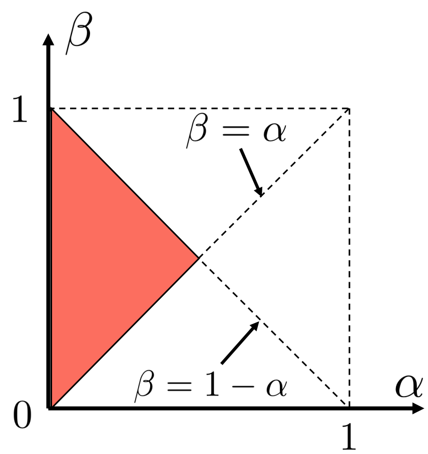

Figure A1: Red area represents region where marginal probability values α and β satisfy $\alpha \leq \beta \leq 1 - \alpha$.

B Convergence Guarantees for Minimizing (7)

To minimize the objective function (7) in the main manuscript, we used Adam Kingma and Ba [2015] in our experiments.

Although we empirically confirmed its convergence, stochastic first-order methods such as Adam are not theoretically guaranteed to converge if applied to (7) particularly when h_θ is a neural network classifier.

This is because the objective function (7) becomes nonconvex, and its gradient does not become *Lipschitz continuous*; that is, the maximum rate of change in the gradient is not bounded.

Fortunately, if we use only activation functions whose gradients are Lipschitz continuous (e.g., the sigmoid function), the gradient of the objective function becomes *locally Lipschitz continuous*, which can be optimized with convergence guarantees by using e.g., the gradient sampling method Burke et al. [2005].

C Upper Bound on PIU in the Presence of Unobserved Variables

In this section, we show how we can formulate the upper bound on PIU when there are unobserved variables in a given causal graph.

As described in Section 3.5 in the main manuscript, to obtain the upper bound on PIU when there are unobserved variables, we need the lower and upper bounds on the marginal potential outcome probabilities.

According to Miles et al. [2017], such upper and lower bounds can be obtained for some simple causal graphs.

Figure A2 presents an example of causal graph shown in Miles et al. [2017], where M is a discrete random variable, $Y \in \{0,1\}$ is a binary outcome, and R is a variable that can be continuous or discrete. As regards A, although Miles et al. [2017] dealt with the general case, where A can be continuous or discrete, for simplicity, we consider the binary case, i.e., $A \in \{0,1\}$. The causal graph in Figure A2 has an unobserved variable H, which affects the two variables R and Y.[5] To define the direct effect along $A \to Y$ in the causal graph in Figure A2, they introduced the potential outcome variables as $Y_{[0]} = Y(0)$ and $Y_{[1]} = Y(1, M(0))$.

Miles et al. [2017] showed that the marginal potential outcome probability $\mathrm{P}(Y_{[1]} = 1)$ follows $l^{[1]} \leq \mathrm{P}(Y_{[1]} = 1) \leq u^{[1]}$, where $l^{[1]}$ is the following lower bound:

$$l^{[1]} = \sum_m \min\{\mathrm{P}(M = m|A = 0), \sum_r \mathrm{P}(Y = 1|A = 1, m, r)\,\mathrm{P}(R = r|A = 1)\}$$

and $u^{[1]}$ is the following upper bound

$$u^{[1]} = \sum_m \max\{0, \mathrm{P}(M = m|A = 0) - 1 + \sum_r \mathrm{P}(Y = 1|A = 1, m, r)\,\mathrm{P}(R = r|A = 1)\}.$$

To represent these bounds as functions of classifier parameter θ, we only have to plug classifier output $h_\theta(1, m, r)$ into conditional distribution $\mathrm{P}(Y = 1|A = 1, m, r)$ and estimate the conditional probabilities $\mathrm{P}(M = m|A = 0)$ and $\mathrm{P}(R = r|A = 1)$ by fitting some regression models (e.g., logistic regression and neural networks) to the data beforehand, which yields

$$l_\theta^{[1]} = \sum_m \min\{\hat{\mathrm{P}}(M = m|A = 0), \sum_r h_\theta(1, m, r)\hat{\mathrm{P}}(R = r|A = 1)\} \tag{A5}$$

$$u_\theta^{[1]} = \sum_m \max\{0, \hat{\mathrm{P}}(M = m|A = 0) - 1 + \sum_r h_\theta(1, m, r)\hat{\mathrm{P}}(R = r|A = 1)\} \tag{A6}$$

where $\hat{\mathrm{P}}(M = m|A = 0)$ and $\hat{\mathrm{P}}(R = r|A = 1)$ denote the estimated conditional probabilities $\mathrm{P}(M = m|A = 0)$ and $\mathrm{P}(R = r|A = 1)$, respectively.

As with (A5) and (A6), we can formulate the lower- and upper-bounds on marginal probability $\mathrm{P}(Y_{[0]} = 1)$.

By using these expressions, as shown in (7) in the main manuscript, we can formulate the constraint on the upper bound on PIU and derive the penalty term from it.

Note that as described in Section 3.5 in the main manuscript, such a penalty term is not differentiable with respect to θ when using the bounds in (A5) and (A6) because these bounds are not differentiable. Utilizing other differentiable bounds on marginal potential outcome probabilities to formulate an objective function that is easy to optimize is left as our future work.

D Experimental Settings and Additional Experiments

D.1 Settings in Our Method

We describe the settings in our method.

[5] Such a variable is called an *unobserved confounder* in the causal inference community.

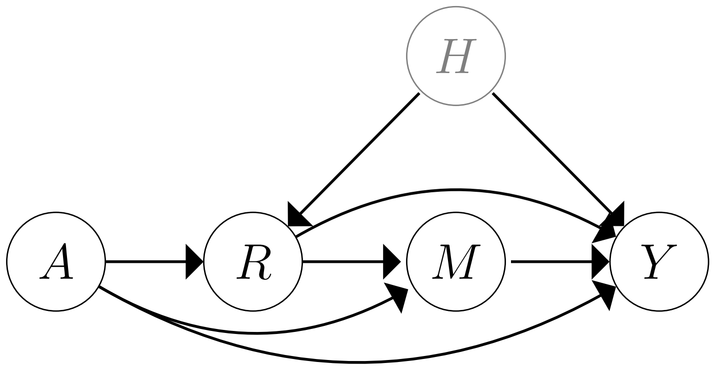

Figure A2: Example of causal graph containing an unobserved variable H, which affects both R and Y.

To evaluate the performance of our method, as classifier h_θ, we used the logistic regression for **Proposed** (logistic) and the two-layered feed-forward neural network for **Proposed** (DNN). As regards **Proposed** (DNN), we formulated the neural network by employing two linear layers with 100 and 50 hidden neurons and used the sigmoid functions as the activation functions.

To train the classifier, 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.

D.2 Formulations of Marginal Potential Outcome Probabilities in Adult Dataset

To evaluate the upper bound on PIU in the UCI Adult dataset, we formulated the marginal potential outcome probabilities as follows.

Let gender A represent 0 for females and 1 for males. To express the unfair effects, following Nabi and Shpitser [2018], we represent the potential outcomes as

$$\begin{aligned}
Y_{[0]} &= Y(0, M(0), L(0, M(0)), \mathbf{R}(0, M(0), L(0, M(0)))),\\
Y_{[1]} &= Y(1, M(1), L(0, M(1)), \mathbf{R}(0, M(1), L(0, M(1)))).
\end{aligned}$$

We estimated the marginal probabilities of these potential outcomes by

$$\begin{aligned}
\hat{p}_\theta^{[0]} &= \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i^{[0]} h_\theta(0, m_i, l_i, \mathbf{r}_i, \mathbf{c}_i),\\
\hat{p}_\theta^{[1]} &= \frac{1}{n}\sum_{i=1}^{n} \hat{w}_i^{[1]} h_\theta(1, m_i, l_i, \mathbf{r}_i, \mathbf{c}_i),
\end{aligned} \tag{A7}$$

where $\hat{w}^{[0]} > 0$ and $\hat{w}^{[1]} > 0$ are the estimated weights that are formulated as follows:

$$\hat{w}^{[0]} = \frac{\hat{\mathrm{P}}(A=0|m,l,\mathbf{r},\mathbf{c})}{\hat{\mathrm{P}}(A=0|\mathbf{c})}$$
$$\hat{w}^{[1]} = \frac{\hat{\mathrm{P}}(A=1|m,\mathbf{c})}{P(A=1|\mathbf{c})} \cdot \frac{\hat{\mathrm{P}}(A=0|m,l,\mathbf{r},\mathbf{c})}{\hat{\mathrm{P}}(A=0|m,\mathbf{c})}.$$

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

D.2.1 Proof of Eq. (A7)

In what follows, we follow a previous work Huber [2014] to prove that the estimators of marginal probabilities $\hat{p}_{\theta}^{[0]}$ and $\hat{p}_{\theta}^{[1]}$ can be expressed as (A7).

To prove (A7), we use the following assumptions of the conditional independence relations between variables, which are called *sequential ignorability assumptions*:

Assumption 1 (Conditional Independence of A) *For all a, $a' \in \{0,1\}$, m in support of M, l in support of L, and $\mathbf{r}$ in support of $\mathbf{R}$, the gender A is conditionally independent of the variables $Y(a,m,l,\mathbf{r})$, $M(a)$, $L(a',m)$, and $\mathbf{R}(a',m,l)$ given age and nationality $\mathbf{C}$. In other words, we have*

$$\begin{aligned} Y(a,m,l,\mathbf{r}) &\perp\!\!\!\perp A \mid \mathbf{C} \\ M(a) &\perp\!\!\!\perp A \mid \mathbf{C} \\ L(a',m) &\perp\!\!\!\perp A \mid \mathbf{C} \\ \mathbf{R}(a',m,l) &\perp\!\!\!\perp A \mid \mathbf{C}. \end{aligned}$$

Assumption 2 (Conditional Independence) *For all a, $a' \in \{0,1\}$, and for m, l, and $\mathbf{r}$ in support of M, L, and $\mathbf{R}$, the following conditional relationships hold*

$$\begin{aligned} Y(a,m,l,\mathbf{r}) &\perp\!\!\!\perp M \mid A, \mathbf{C} \\ L(a',m) &\perp\!\!\!\perp M \mid A, \mathbf{C} \\ \mathbf{R}(a',m,l) &\perp\!\!\!\perp M \mid A, \mathbf{C} \\ Y(a,m,l,\mathbf{r}) &\perp\!\!\!\perp L \mid A, \mathbf{C} \\ \mathbf{R}(a',m,l) &\perp\!\!\!\perp L \mid A, \mathbf{C} \\ Y(a,m,l,\mathbf{r}) &\perp\!\!\!\perp \mathbf{R} \mid A, \mathbf{C}. \end{aligned}$$

In what follows, we only consider the estimator of $\mathrm{P}(Y^{[1]}=1)$ since we can estimate $\mathrm{P}(Y^{[0]}=1)$ in a similar manner.

Given the causal graph, marginal probability $\mathrm{P}(Y^{[1]}=1)$ can be represented as

$$\begin{aligned}&\mathrm{P}(Y^{[1]}=1)\\&=\iiiint \mathrm{P}(Y(1,m,l,\mathbf{r})=1|\mathbf{C}=\mathbf{c},M(1)=m,L(0,M(1))=l,\mathbf{R}(0,M(1),L(0,M(1)))=\mathbf{r})\\&\mathrm{d}F_{\mathbf{C}}(\mathbf{c})\mathrm{d}F_{M(1)|\mathbf{C}=\mathbf{c}}(m)\mathrm{d}F_{L(0,M(1))|M(1)=m,\mathbf{C}=\mathbf{c}}(l)\mathrm{d}F_{\mathbf{R}(0,M(1),L(0,M(1)))|L(0,M(1))=l,M(1)=m,\mathbf{C}=\mathbf{c})}(\mathbf{r}).\end{aligned}$$

With Assumptions 1 and 2, this can be rewritten as

$$\begin{aligned}&\mathrm{P}(Y^{[1]}=1)\\&=\iiiint \mathrm{P}(Y(1,m,l,\mathbf{r})|A=1,M=m,L=l,\mathbf{R}=\mathbf{r},\mathbf{C}=\mathbf{c})\\&\mathrm{d}F_{\mathbf{C}}(\mathbf{c})\mathrm{d}F_{M|\mathbf{C}=\mathbf{c}}(m)\mathrm{d}F_{L|M=m,\mathbf{C}=\mathbf{c}}(l)\mathrm{d}F_{\mathbf{R}|L=l,M=m,\mathbf{C}=\mathbf{c})}(\mathbf{r}),\end{aligned}$$

which can be expressed using Bayes' theorem as

$$\begin{aligned}&\mathrm{P}(Y^{[1]}=1)\\&=\iiiint \mathrm{P}(Y(1,m,l,\mathbf{r})|A=1,M=m,L=l,\mathbf{R}=\mathbf{r},\mathbf{C}=\mathbf{c})\frac{\mathrm{P}(A=0|\mathbf{r},l,m,\mathbf{c})}{\mathrm{P}(A=0|m,\mathbf{c})}\cdot\frac{\mathrm{P}(A=1|m,\mathbf{c})}{\mathrm{P}(A=1|\mathbf{c})}\\&\mathrm{d}F_{\mathbf{R}|L=l,M=m,\mathbf{C}=\mathbf{c})}(\mathbf{r})\ \mathrm{d}F_{L|M=m,\mathbf{C}=\mathbf{c}}(l)\mathrm{d}F_{M|\mathbf{C}=\mathbf{c}}(m)\mathrm{d}F_{\mathbf{C}}(\mathbf{c}).\end{aligned}\tag{A8}$$

Given the empirical distribution, by plugging classifier into conditional distribution $\mathrm{P}(A,M,L,\mathbf{R},\mathbf{C})$, the relation (A8) can be expressed as

$$\hat{p}_\theta^{[1]}=\frac{1}{n}\sum_{i=1}^{n}w_i^{(1)}h_\theta(1,m_i,l_i,\mathbf{r}_i,\mathbf{c}_i),\tag{A9}$$

where

$$w_i^{(1)}=\frac{\mathrm{P}(A=1|m_i,\mathbf{c}_i)}{\mathrm{P}(A=1|\mathbf{c}_i)}\cdot\frac{\mathrm{P}(A=0|m_i,l_i,\mathbf{r}_i,\mathbf{c}_i)}{\mathrm{P}(A=0|m_i,\mathbf{c}_i)}.$$

Thus, we prove (A7).

D.3 Computing Infrastructure

As an implementation of the optimization algorithm (i.e., Adam Kingma and Ba [2015]), we used PyTorch 1.1.0.

We performed synthetic data experiments on a 64-bit macOS machine with 2.7GHz Intel Core i7 CPUs and 16GB RAMs. Meanwhile, to conduct experiments with the large real-world dataset (the Adult dataset), we used a 64-bit CentOS machine with 2.6GHz Xeon E5-2697A-v4 (x2) CPUs and 512GB RAMs.

D.4 Additional Synthetic Data Experiments

In this section, to show the effectiveness of the proposed constraint, we provide additional results of synthetic data experiments. Specifically, by using synthetic data presented in Section 6.1, we compare our method with a naive strategy that learns the classifier without using gender A.

In fact, with the synthetic data, we can eliminate unfair effects by learning the classifier without using gender A. This is because in this case, the unfair effects are represented as the direct causal effects along the pathway $A \to Y$ (see, the causal graph in Figure 1). As described in Section 2.1, the direct causal effect represents how greatly the predicted decision for each individual is changed just by changing A values. However, if we learn the classifier without using A, the predicted decision (i.e., the classifier output) for each individual can never be affected by the change in A value because A is not included in the classifier input. Thus learning the classifier without A ensures that the unfair effect values are exactly zero for all individuals.

Compared with the above naive strategy, our method has the following difference. While our method can control the trade-off between fairness and accuracy by appropriately choosing the value of penalty parameter λ, the naive strategy cannot. Although making decisions without A makes the unfair effects exactly zero for all individuals, this corresponds to setting $\lambda = \infty$ in our method, which leads to a decrease in the prediction accuracy.

To confirm this, as with Section 6.1, using 10 randomly generated training and test datasets, we evaluated the mean test accuracy when learning the classifier without gender A.

As a result, the naive strategy with the logistic regression model and with the neural network yielded 67.9% and 84.5% test accuracies, respectively. In contrast, **Proposed** (logistic) and **Proposed** (DNN) achieved higher test accuracies (72.1% and 85.0%, respectively, as shown in Table 1a) while keeping the PIU values almost zero (2.30×10^{-3} and 1.89×10^{-2}, respectively). These results demonstrate that our method can strike a better balance between fairness and prediction accuracy than the naive strategy that learns the classifier without the sensitive feature A.