Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)

Counterfactual Fairness: Unidentification, Bound and Algorithm

Yongkai Wu , **Lu Zhang** and **Xintao Wu**
University of Arkansas
{yw009, lz006, xintaowu}@uark.edu

Abstract

Fairness-aware learning studies the problem of building machine learning models that are subject to fairness requirements. Counterfactual fairness is a notion of fairness derived from Pearl's causal model, which considers a model is fair if for a particular individual or group its prediction in the real world is the same as that in the counterfactual world where the individual(s) had belonged to a different demographic group. However, an inherent limitation of counterfactual fairness is that it cannot be uniquely quantified from the observational data in certain situations, due to the unidentifiability of the counterfactual quantity. In this paper, we address this limitation by mathematically bounding the unidentifiable counterfactual quantity, and develop a theoretically sound algorithm for constructing counterfactually fair classifiers. We evaluate our method in the experiments using both synthetic and real-world datasets, as well as compare with existing methods. The results validate our theory and show the effectiveness of our method.

1 Introduction

It is important to develop fairness-aware machine learning algorithms and models such that the decisions made with their assistance are subject to fairness requirements. In recent years, the research community has studied fairness-aware machine learning from the causal perspective [Zhang *et al.*, 2017b; Zhang *et al.*, 2017a; Zhang and Bareinboim, 2018; Nabi and Shpitser, 2018; Zhang *et al.*, 2018b; Zhang *et al.*, 2018a] using causal modeling [Pearl, 2009]. In these works, fairness is generally formulated and quantified as the average causal effect of the sensitive attribute on the decision attribute. The effect is evaluated by the intervention through the post-interventional distributions. Different from above works, [Kusner *et al.*, 2017] introduced counterfactual fairness, based on the counterfactual inference, which considers the causal effect within a particular individual/group specified by of observational profile attributes. The notion of counterfactual fairness is more general than the intervention-based notions where the set of profile attributes is empty. Consequently, the counterfactual inference is more challenging than the intervention. This is because measuring interventions only considers the post-interventional distributions, but counterfactual inference considers both the real world without the intervention and the counterfactual world with the intervention. Researchers have proved that the counterfactual quantity cannot be uniquely computed from the observational data in some situations, which are referred to as the unidentifiable situations [Pearl, 2009].

The unidentifiable situations are big barriers to the application of counterfactual fairness. In [Kusner *et al.*, 2017], the authors proposed three methods to evade the unidentifiability issue: 1) only non-descendants of the sensitive attribute are used in classification, 2) the non-deterministic substitutions of the hidden variables are postulated and inferred based on domain knowledge, or 3) the complete causal model is postulated and estimated, e.g., , being treated as the additive noise model then estimating the errors. However, the sensitive attribute is usually an inherent nature of data hence many attributes are its descendants. If all descendants are forbidden, very few attributes are allowed for classifier training, weakening the resultant fair classifier dramatically. Also, it is over-simplified to postulate the substitutions and their distributions, since the exogenous variables represent all possible sources of randomness; or presuppose that the causal model, which is supposed to represent the underlying mechanism of the world, is an additive model.

In this paper, we address the problem of learning counterfactually fair classifiers by mathematically bounding the unidentifiable counterfactual quantity. We leverage the counterfactual graph proposed in [Shpitser and Pearl, 2008] for depicting the independence relationships among variables in the real world and the counterfactual world which are of concern in the counterfactual quantity. Then, we adopt the c-component factorization to decompose the counterfactual quantity, and identify the terms that are the source of unidentification. We propose a graphical criterion for determining the identification of counterfactual fairness and develop the lower and upper bounds of counterfactual fairness in unidentifiable situations. Finally, we propose a post-processing method for reconstructing arbitrary classifiers in order to achieve counterfactual fairness. We formulate the reconstruction problem as a linear constrained optimization problem with the bounded counterfactual fairness criterion as the constraints.

In the experiments, we evaluate our methods and compare

them with existing ones using real-world datasets and synthetic datasets where the ground-truth of counterfactual fairness can be precisely quantified. The results show that our method correctly achieves counterfactual fairness as expected according to our theorem, while obtaining high accuracy of prediction. On the contrary, the methods proposed in [Kusner *et al.*, 2017] either fail to achieve counterfactual fairness or suffer from low accuracy due to simplified assumptions.

2 Preliminaries

2.1 Structural Causal Model and Intervention

Definition 1 (Structural Causal Model)**.** *A structural causal model $\mathcal{M}$ is represented by a triple $\langle \mathbf{U}, P(\mathbf{U}), \mathbf{V}, \mathbf{F} \rangle$ where*

1. *$\mathbf{U}$ is a set of exogenous variables of any types, i.e., discrete, continuous, or mixed. An arbitrary joint probability distribution $P(\mathbf{U})$ is defined over $\mathbf{U}$.*
2. *$\mathbf{V}$ is a set of endogenous variables that are determined by variables in $\mathbf{U} \cup \mathbf{V}$.*
3. *$\mathbf{F}$ is a set of functions mapping from $\mathbf{U} \cup \mathbf{V}$ to $\mathbf{V}$. Specifically, for each $X \in \mathbf{V}$, there is a function $f_X \in \mathbf{F}$ mapping from $\mathbf{U} \cup (\mathbf{V} \setminus X)$ to X, i.e., $X = f_X(\mathsf{Pa}(X), \mathbf{U}_X)$, where $\mathsf{Pa}(X) \subseteq \mathbf{V} \setminus X$ stands for the endogenous variables that directly determine the value of X, and $\mathbf{U}_X \subseteq \mathbf{U}$ represents all sources of randomness.*

A causal model is associated with a causal graph $\mathcal{G}$ where each node corresponds to a variable in $\mathbf{V}$ [1], and each edge, denoted by an arrow, points from a node X to another node Y if X is an input of f_Y. In this manuscript, we focus on the Markovian causal model where all exogenous variables are independent. Thus, the causal graph is simplified by omitting all exogenous variables and the edges associated with them. For any set of nodes $\mathbf{X}$, we use $\mathsf{Pa}(\mathbf{X})_{\mathcal{G}}$, $\mathsf{Ch}(\mathbf{X})_{\mathcal{G}}$, $\mathsf{An}(\mathbf{X})_{\mathcal{G}}$, and $\mathsf{De}(\mathbf{X})_{\mathcal{G}}$ to denote the sets of parents, children, ancestors, and descendants of $\mathbf{X}$ in $\mathcal{G}$.

In the causal model, the quantitative measure of causal effects is facilitated by interventions through *do*-calculus [Pearl, 2009], which simulates the physical interventions that force some variable X to take certain values x. Formally, the intervention that fixes the value of X to x is denoted by $do(x)$. The mathematical meaning of $do(x)$ in a causal model $\mathcal{M}$ is defined as the substitution of equation $X = f_X(\mathsf{Pa}(X)_{\mathcal{G}}, \mathbf{U}_X)$ with $X = x$. The causal model after performing $do(x)$ is called a submodel denoted by $\mathcal{M}_x$. For another endogenous variable Y which is affected by the intervention, its interventional variant in submodel $\mathcal{M}_x$ is denoted by Y_x. The distribution of Y_x, called the post-intervention distribution of Y under $do(x)$, is denoted by $P(y_x)$. For simplicity, we rewrite $P(y_x)$ as $P_x(y)$, meaning the distribution of (the variant of) Y in submodel $\mathcal{M}_x$. Similarly, we can rewrite the condition distribution of Y_x given Z_x, i.e., $P(y_x|z_x)$, as $P_x(y|z)$. Pearl [Pearl, 2009] proposed three rules of *do*-calculus to infer post-intervention distributions from observational data, by converting post-intervention distributions to observational distributions.

[1] An uppercase letter denotes an attribute and a lowercase letter denotes an attribute value. Similarly, a bold uppercase letter denotes a set of attributes and a bold lowercase letter denotes a set of values.

2.2 Counterfactual Inference and Unidentification

In Section 2.1, the causal effect is estimated using intervention where the post-intervention distribution concerns the counterfactual world represented by submodel $\mathcal{M}_x$ only. If we infer the post-intervention distribution while conditioning on certain individuals or groups specified by a subset of endogenous variables, the inferred quantity will involve two worlds simultaneously, the real world represented by causal model $\mathcal{M}$, and the counterfactual world $\mathcal{M}_x$, hence cannot be resolved by *do*-calculus directly. Such causal inference problem is called the counterfactual inference, and the distribution of Y_x conditioning on the real world observation $\mathbf{O} = \mathbf{o}$ is denoted by $P(y_x|\mathbf{o})$. Note that Y_x is a variable in submodel $\mathcal{M}_x$, while $\mathbf{O}$ are variables in original causal model $\mathcal{M}$.

Apparently, inferring $P(y_x|\mathbf{o})$ requires to know the connection between the real world and the counterfactual world. This can be done if we have complete knowledge of the causal model. According to [Pearl, 2009], the counterfactual inference can be exactly performed using three steps if the complete model, including all the structural equations, is known: **1. Abduction**: Update $P(\mathbf{u})$ by observation $\mathbf{O} = \mathbf{o}$ to obtain $P(\mathbf{u}|\mathbf{o})$. **2. Action**: Modify $\mathcal{M}$ by intervention $do(x)$ to obtain the submodel $\mathcal{M}_x$. **3. Prediction**: Use modified submodel $\langle \mathcal{M}_x, P(\mathbf{u}|\mathbf{o}) \rangle$ to compute the probability of Y_x,, i.e., the consequence of the counterfactual inference.

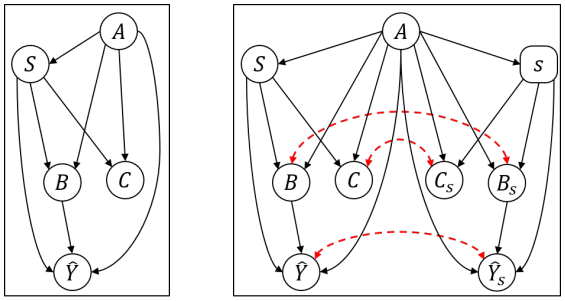
The above method is usually infeasible in practice due to the lack of the complete knowledge of the causal model. If we only have the causal graph and observational data, which is a common scenario in the literature, the counterfactual quantity might be evaluated by using the **IDC*** algorithm developed in [Shpitser and Pearl, 2008]. However, in certain situations where the **IDC*** algorithm fails, the corresponding counterfactual quantity cannot be uniquely computed from the observational data in theory. These situations are referred to as the unidentifiable situations. One typical unidentifiable situation [Shpitser and Pearl, 2008] is shown in Lemma 1.

Lemma 1. *Let X, Y be two variables such that Y is a parent of X, then $P(Y = y, Y_x = y')$ is unidentifiable if $y \neq y'$.*

3 Quantifying and Bounding Counterfactual Fairness

Fairness-aware learning is widely studied using causal modeling to capture the causal connection between the sensitive attribute and the challenged decision [Kilbertus *et al.*, 2017; Li *et al.*, 2017; Zhang *et al.*, 2017b; Zhang and Bareinboim, 2018; Nabi and Shpitser, 2018; Chiappa, 2019; Kusner *et al.*, 2017]. We adopt the notion of counterfactual fairness proposed in [Kusner *et al.*, 2017], which formulates fairness as the equivalence of two counterfactual quantities. Although this notion captures the true intuition behind fairness, it faces significant computational challenges due to the unidentifiability of counterfactual inference. In this section, we first give the formal definition of counterfactual fairness for predictive models and explain its physical meaning. Then, we show how to address above challenges by mathematically bounding the unidentifiable counterfactual quantity.

In our notations, $S \in \{s^+, s^-\}$ denotes the sensitive attribute, $Y \in \{y^+, y^-\}$ denotes the decision, and $\mathbf{X}$ denotes



(a) (b)

Figure 1: (a) Causal Graph $\mathcal{G}$. (b) Counterfactual Graph $\mathcal{G}'$ for $P(\hat{y}_s|s', \mathbf{z})$.

the set of other attributes. The historical dataset $\mathcal{D}$ drawn from a distribution $P(\mathbf{X}, S, Y)$ is used to train a classifier $f : \mathbf{X}, S \to \hat{Y}$. The underlying mechanism that determines a distribution $P(\mathbf{X}, S, \hat{Y})$ is represented by a causal model $\mathcal{M}$. The causal graph associated with the causal model is denoted by $\mathcal{G}$. Then, counterfactual fairness is defined as follows.

Definition 2. *(**Counterfactual Fairness**) Given a set of attributes $\mathbf{Z} \subseteq \mathbf{X}$, a classifier $f : \mathbf{X}, S \to \hat{Y}$ is counterfactually fair w.r.t. $\mathbf{Z}$, if under any observational condition $\mathbf{Z} = \mathbf{z}$ we have*

$$P(\hat{y}_{s'}|s', \mathbf{z}) = P(\hat{y}_s|s', \mathbf{z}), \text{ where } s', s \in \{s^+, s^-\}.$$

Recall that a lowercase letter with a subscript represents a value assignment to the corresponding variable in the submodel, e.g., $\hat{y}_s$ is a value of $\hat{Y}_s$ in the submodel $\mathcal{M}_s$.

The physical meaning of counterfactual fairness can be interpreted as follows. Consider candidates are applying for a job and a predictive model is used to make the decision $\hat{Y}$. We concern an individual from disadvantage group s^- who is specified by a profile $\mathbf{z}$. Straightforwardly, the probability of the individual to get the positive decision is $P(\hat{y}|s^-, \mathbf{z})$, which is equivalent to $P(\hat{y}_{s^-}|s^-, \mathbf{z})$ since the intervention makes no change to S's value of that individual. Now assume the value of S for this very individual had been changed from s^- to s^+. The probability of this individual to get the positive decision after the hypothetical change is given by $P(\hat{y}_{s^+}|s^-, \mathbf{z})$. Therefore, if two probabilities $P(\hat{y}_{s^-}|s^-, \mathbf{z})$ and $P(\hat{y}_{s^+}|s^-, \mathbf{z})$ are identical, we can claim the individual is treated fairly as if he/she had been from the other group.

3.1 Identification of Counterfactual Quantity

In this section, we identify the source of unidentification for the counterfactual quantity and give a graphical criterion determining the identifiability of the counterfactual quantity. Our method is inspired by the **IDC*** algorithm and we further extend it to bound the unidentifiable quantity.

The analysis of $P(\hat{y}_s|s', \mathbf{z})$ concerns the connection between two causal models, $\mathcal{M}$ and $\mathcal{M}_s$. Thus, we apply the **make-cg** algorithm [Shpitser and Pearl, 2008] to the causal graph $\mathcal{G}$ to construct a new graph $\mathcal{G}'$ that depicts the independence relationship among all variables in $\mathcal{M}$ and $\mathcal{M}_s$ that are of concern in the analysis. The **make-cg** algorithm first combines the two causal graphs and makes them share the same exogenous variables $\mathbf{U}$, corresponding to the shared causal context or background. Then, it removes the duplicated endogenous nodes which are also not affected by $do(s)$. The resultant graph is the so-called counterfactual graph. Next, we apply the c-component factorization [Tian and Pearl, 2002] to decompose counterfactual graph $\mathcal{G}'$ into disjoint subgraphs called the c-components, such that any two nodes in the same c-component are connected by a bi-directed path[2]. After that, the joint distribution of all variables in the counterfactual graph can be factorized as the product of the conditional distribution of each c-component. Our theoretical analysis will show that if certain c-component has the unidentifiability issue that cannot be resolved by summation, the corresponding counterfactual quantity is unidentifiable. Without loss of generality, we first use an example to illustrate our idea. Consider the causal graph $\mathcal{G}$ shown in Figure 1 (a) where there are five attributes $A, B, C, S, \hat{Y}$: S is the sensitive attribute; $\hat{Y}$ is the prediction of the decision attribute obtained by any classifier; A is the ancestor of $\hat{Y}$ but not the descendant of S; B is the intersection between the ancestor of Y and the descendant of S; and C is the descendant of S but not the ancestor of $\hat{Y}$. We aim to study the identifiability of $P(\hat{y}_s|s', \mathbf{z})$, where $\mathbf{Z}$ is an arbitrary subset of $\{A, B, C\}$.

The counterfactual graph denoted by $\mathcal{G}'$ is shown in Figure 1 (b), where the bi-directed dash edge implies that the two nodes share the same exogenous variables. Note that A and A_s are merged as A since they are duplicated. Next, we apply the c-component factorization. In Figure 1 (b), there are five c-components: $\langle A \rangle$, $\langle S \rangle$, $\langle B, B_s \rangle$, $\langle C, C_s \rangle$, and $\langle \hat{Y}, \hat{Y}_s \rangle$. We can factorize $P(\hat{y}_s, s', \mathbf{z})$ as

$$P(\hat{y}_s, s', \mathbf{z}) = \sum_{\mathbf{x}\setminus\mathbf{z}, \hat{y}, b', c'} R(a)R(s')R(c, c'_s)R(b, b'_s)R(\hat{y}', \hat{y}_s),$$

where $R(\mathbf{w}) = P\left(\mathbf{w}|\mathsf{Pa}(\mathbf{W})_{\mathcal{G}'}\right)$ for any node set $\mathbf{W}$, $\mathbf{x} = \{a, b, c\}$, and $\mathbf{z}$ is any subset of $\mathbf{x}$. Then, we can derive that

$$P(\hat{y}_s|s', \mathbf{z}) = \frac{\sum_{\mathbf{x}\setminus\mathbf{z}, \hat{y}, b', c'} \begin{bmatrix} P(a)P(s'|a)P(c,c'_s|s',a) \\ P(b,b'_s|s',a)P(\hat{y}', \hat{y}_s|a,b,b'_s) \end{bmatrix}}{P(s', \mathbf{z})}.$$

Note that c'_s in $P(c, c'_s|s', a)$ and $\hat{y}'$ in $P(\hat{y}, \hat{y}_s|a, b, b'_s)$ can be canceled out by summation. By applying the m-separation, we can remove b from $P(\hat{y}_s|a, b, b'_s)$, as B is d-separated from $\hat{Y}_s$ conditioning on A and B_s. Thus, we obtain

$$P(\hat{y}_s|s', \mathbf{z}) = \frac{\sum_{\mathbf{x}\setminus\mathbf{z}, b'} \begin{bmatrix} P(a)P(s'|a)P(c|s',a) \\ P(b,b'_s|s',a)P(\hat{y}_s|a,b'_s) \end{bmatrix}}{P(s', \mathbf{z})}. \quad (1)$$

To further analyze Eq. (1), we consider two cases below.

Case 1 ($B \notin \mathbf{Z}$): In this case, we have b under the Σ of Eq. (1), hence b in $P(b, b'_s|s', a)$ can be canceled out by summation, resulting in $P(b'_s|s', a)$. Then, we can remove s' from $P(b'_s|s', a)$ as B_s is d-separated from S conditioning on A, resulting in $P(b'_s|a)$. We can further rewrite $P(\hat{y}_s|a, b'_s)$ as $P_s(\hat{y}|a, b')$, and rewrite $P(b'_s|a)$ as $P(b'_s|a)$. At last, we invoke do-calculus Rule 2 [Pearl, 2009] to convert $P_s(\hat{y}|a, b')$

[2] A bi-directed path is a path consisting of bi-directed edges only.

to $P(\hat{y}|a,b',s)$, and $P_s(b'|a)$ to $P(b'|a,s)$. Finally, we obtain

$$P(\hat{y}_s|s',\mathbf{z}) = \frac{\sum_{\mathbf{x}\setminus\mathbf{z}\setminus\{b\},b'} P(s',a,c)P(b'|a,s)P(\hat{y}|a,b',s)}{P(s',\mathbf{z})} = \frac{\sum_{\mathbf{x}\setminus\mathbf{z}\setminus\{b\}} P(s',a,c)P(\hat{y}|a,s)}{P(s',\mathbf{z})}. \quad (2)$$

Case 2 ($B \in \mathbf{Z}$): In this case, since we don't have b under the Σ, term $P(b,b'_s|s',a)$ cannot be reduced, resulting in

$$P(\hat{y}_s|s',\mathbf{z}) = \frac{\sum_{\mathbf{x}\setminus\mathbf{z},b'} P(s',a,c)P(b,b'_s|a,s')P(\hat{y}|a,b',s)}{P(s',\mathbf{z})}. \quad (3)$$

From above two cases we see that, $P(\hat{y}_s|s',\mathbf{z})$ in Case 1 is identifiable as all terms in Eq. (2) can be read from observational data. One can verify that this result is consistent with the **IDC*** algorithm. However in Case 2, since $P(b,b'_s|s',a)$ in Eq. (3) is unidentifiable according to Lemma 1, $P(\hat{y}_s|s',\mathbf{z})$ is also unidentifiable. In this example, the identifiability of $P(\hat{y}_s|s',\mathbf{z})$ depends on whether node B, the intersection of S's descendants and $\hat{Y}$'s ancestors, is in set $\mathbf{Z}$ or not. We summarize this result as follows.

Proposition 1. *For the causal graph in Figure 1 (a), $P(\hat{y}_s|s',\mathbf{z})$ is unidentifiable if and only if $B \in \mathbf{Z}$.*

3.2 Bounding Unidentifiable Counterfactual Quantity

In Eq. (3), we identify the source of unidentifiability. Next, we derive the lower and upper bounds for $P(\hat{y}_s|s',\mathbf{z})$ as shown in the following proposition, which works for both identifiable and unidentifiable situations.

Proposition 2. *For the causal graph in Figure 1 (a) we have*

$$P(\hat{y}_s|s',\mathbf{z}) \leq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s',\mathbf{x}) \max_{\mathbf{m}'} \{P(\hat{y}|s,a,b')\}}{P(s',\mathbf{z})}, \quad (4)$$

$$P(\hat{y}_s|s',\mathbf{z}) \geq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s',\mathbf{x}) \min_{\mathbf{m}'} \{P(\hat{y}|s,a,b')\}}{P(s',\mathbf{z})}, \quad (5)$$

where $\mathbf{x} = \{a,b,c\}$, $\mathbf{z}$ is any subset of $\mathbf{x}$, and $\mathbf{M} = \{B\} \cap \mathbf{Z}$.

Proof. Suppose $B \in \mathbf{Z}$, then $\mathbf{M} = \{B\}$. Obviously, we have

$$P(\hat{y}|s,a,b') \leq \max_{b'} \{P(\hat{y}|s,a,b')\}.$$

By applying this inequality to Eq. (3), we have

$$P(\hat{y}_s|s',\mathbf{z}) \leq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s',a,c) \max_{b'} \{P(\hat{y}|s,a,b')\} \sum_{b'} P(b,b'_s|s',a)}{P(s',\mathbf{z})} = \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s',a,c)P(b|s',a) \max_{b'} \{P(\hat{y}|s,a,b')\}}{P(s',\mathbf{z})} = \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s',\mathbf{x}) \max_{b'} \{P(\hat{y}|s,a,b')\}}{P(s',\mathbf{z})}.$$

The second step is due to the condition $\sum_{b'} P(b,b'_s|s',a) = P(b|s',a)$, and the third step is due to $B \perp C|A,S$. Similarly, we can replace max with min to obtain Eq. (5).

If $B \notin \mathbf{Z}$, $\mathbf{M} = \emptyset$. Then, we have $\max_\emptyset\{P(\hat{y}|s,a,b')\} = \min_\emptyset\{P(\hat{y}|s,a,b')\} = P(\hat{y}|s,a)$ and both Eq. (4) and Eq. (5) become $\frac{\sum_{\mathbf{x}\setminus\mathbf{z}\setminus\{b\}} P(s',a,c)P(\hat{y}|s,a)}{P(s',\mathbf{z})}$, which is consistent with the identifiable situations (i.e., Eq. (2)). □

3.3 Extending to General Case

Above results can be extended to the general case. Let $\mathbf{A}$ denote the ancestors of $\hat{Y}$ which are not the descendants of S, $\mathbf{B}$ denote the intersection between the ancestors of $\hat{Y}$ and the descendants of S, $\mathbf{C}$ denote the descendants of S which are not the ancestors of $\hat{Y}$, i.e.,

$$\mathbf{A} = \mathsf{An}(\hat{Y})_\mathcal{G} \setminus \mathsf{De}(S)_\mathcal{G}, \quad \mathbf{B} = \mathsf{An}(\hat{Y})_\mathcal{G} \cap \mathsf{De}(S)_\mathcal{G},$$
$$\mathbf{C} = \mathsf{De}(S)_\mathcal{G} \setminus \mathsf{An}(\hat{Y})_\mathcal{G}.$$

Note that $\mathbf{A},\mathbf{B},\mathbf{C}$ are disjoint and $\mathbf{X} = \mathbf{A} \cup \mathbf{B} \cup \mathbf{C}$. Now we are ready to extend Propositions 1 and 2 to the general case.

Theorem 1. ***(Identification of Counterfactual Quantity)*** *Given a causal graph $\mathcal{G}$ and the set of profile attributes $\mathbf{Z}$, the counterfactual quantity $P(\hat{y}_s|s',\mathbf{z})$ is unidentifiable if and only if $\mathbf{B} \cap \mathbf{Z} \neq \emptyset$.*

Theorem 2. ***(Bounds of Counterfactual Quantity)*** *Given a causal graph $\mathcal{G}$ and a set of profile attributes $\mathbf{Z}$, we have*

$$P(\hat{y}_s|s',\mathbf{z}) \leq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} \left[\max_{\mathbf{m}'}\left\{P^{(s',\mathbf{x})}(\hat{y}|s,\mathsf{pa}(\hat{\mathbf{Y}})_\mathcal{G} \cap \mathbf{m}', \mathsf{pa}(\hat{\mathbf{Y}})_\mathcal{G} \setminus \{s,\mathbf{m}'\}\right\}\right]}{P(s',\mathbf{z})},$$

$$P(\hat{y}_s|s',\mathbf{z}) \geq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} \left[\min_{\mathbf{m}'}\left\{P^{(s',\mathbf{x})}(\hat{y}|s,\mathsf{pa}(\hat{\mathbf{Y}})_\mathcal{G} \cap \mathbf{m}', \mathsf{pa}(\hat{\mathbf{Y}})_\mathcal{G} \setminus \{s,\mathbf{m}'\}\right\}\right]}{P(s',\mathbf{z})},$$

where we partition $\mathbf{B}$ to two disjoint sets: a set $\mathbf{M} \in \mathbf{Z}$ and a set $\mathbf{N} \notin \mathbf{Z}$ such that $\mathbf{M} = \mathbf{B} \cap \mathbf{Z}, \mathbf{N} = \mathbf{B} \setminus \mathbf{Z}$.

The proofs are similar to the previous ones.

4 Achieving Counterfactual Fairness in Classification

The derived bounds clear the path towards constructing counterfactually fair classifiers. In this section, we propose a post-processing method for reconstructing any classifier to achieve counterfactual fairness. To this end, we first give a relaxed quantitative criterion of fairness based on Definition 2.

Definition 3 (τ-Counterfactual Fairness)**.** *Given a profile attribute set $\mathbf{Z} \subseteq \mathbf{X}$ and a threshold τ, a classifier $f : \mathbf{X}, S \to \hat{Y}$ is counterfactually fair if under any condition $\mathbf{Z} = \mathbf{z}$,*

$$\left|DE(\hat{y}_{s^-\to s^+}|\mathbf{z})\right| \leq \tau,$$

where $DE(\hat{y}_{s^-\to s^+}|\mathbf{z}) = P(\hat{y}_{s^+}|s^-,\mathbf{z}) - P(\hat{y}_{s^-}|s^-,\mathbf{z})$.

In above definition, $\left|DE(\hat{y}_{s^-\to s^+}|\mathbf{z})\right|$ captures the amount of unfairness or discrimination of a classifier in terms of the difference in the positive decision rate for a certain group of individuals (specified by $\mathbf{z}$) between the counterfactual world (where they had been changed to s^+) and the real world (where they are actually in s^-). If the amount of unfairness of a classifier is smaller than τ, we claim this classifier is (counterfactually) fair. Note that the first term $P(\hat{y}_{s^+}|s^-,\mathbf{z})$ has the identification issue, but the second term $P(\hat{y}_{s^-}|s^-,\mathbf{z})$ simply equals to $P(\hat{y}|s^-,\mathbf{z})$ since the intervention $do(s^-)$ makes no change to the value of S for this group. By denoting the upper and lower bounds of $P(\hat{y}_{s^+}|s^-,\mathbf{z})$ obtained in Theorem 2 as $ub(P(\hat{y}_{s^+}|s^-,\mathbf{z}))$ and $lb(P(\hat{y}_{s^+}|s^-,\mathbf{z}))$ respectively, we obtain the bounds of $DE(\hat{y}_{s^-\to s^+}|\mathbf{z})$ as follows.

Corollary 1. ***(Bounds of Counterfactual Fairness)*** *The upper and lower bounds of counterfactual fairness $DE(\hat{y}_{s^-\to s^+}|\mathbf{z})$ are given by*

$$ub\left(DE(\hat{y}_{s^-\to s^+}|\mathbf{z})\right) = ub\left(P(\hat{y}_{s^+}|s^-,\mathbf{z})\right) - P(\hat{y}|s^-,\mathbf{z}), \quad (6)$$

$$lb\left(DE(\hat{y}_{s^-\to s^+}|\mathbf{z})\right) = lb\left(P(\hat{y}_{s^+}|s^-,\mathbf{z})\right) - P(\hat{y}|s^-,\mathbf{z}). \quad (7)$$

Corollary 1 can facilitate the detection of unfairness from observational data. Specifically, if we have $ub\left(DE(\hat{y}_{s^-\to s^+}|\mathbf{z})\right) \leq \tau$ and $lb\left(DE(\hat{y}_{s^-\to s^+}|\mathbf{z})\right) \geq -\tau$, then it is guaranteed that τ-counterfactual fairness is satisfied. If we have $ub\left(DE(\hat{y}_{s^-\to s^+}|\mathbf{z})\right) \leq -\tau$ or $lb\left(DE(\hat{y}_{s^-\to s^+}|\mathbf{z})\right) > \tau$, then it is guaranteed that τ-counterfactual fairness cannot be satisfied. Otherwise, it is uncertain and cannot be determined from data.

Based on Corollary 1, we then propose an efficient method for constructing counterfactually fair classifiers. Note that the bounds are consistent with identifiable situations, so the method works for both identifiable/unidentifiable situations.

We consider to construct a new decision variable $\tilde{Y}$ from $\hat{Y}$ in the causal model such that τ-counterfactual fairness regarding $\tilde{Y}$ is satisfied. The objective is to find an optimal probabilistic mapping function $P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$ that minimizes the difference between Y and $\tilde{Y}$, measured by the empirical loss $\mathbb{E}_{\mathcal{D}}[\ell(Y,\tilde{Y})]$, meanwhile, the new decisions are counterfactually fair. The formulation of this optimization problem is given below.

Problem Formulation 1. *Given a dataset $\mathcal{D}$ with prediction $\hat{Y}$ made by an arbitrary classifier, we aim to learn a post-processing mapping function $P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$ by solving the following optimization problem:*

$$\begin{aligned}
&\min \qquad \mathbb{E}_{\mathcal{D}}[\ell(Y,\tilde{Y})] \\
&\textit{s.t. for any } \mathbf{z}: \\
&ub\left(DE(\tilde{y}_{s^-\to s^+}|\mathbf{z})\right) \leq \tau, \quad lb\left(DE(\tilde{y}_{s^+\to s^-}|\mathbf{z})\right) \geq -\tau, \\
&\sum_{\tilde{y}} P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}}) = 1, \quad 0 \leq P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}}) \leq 1,
\end{aligned}$$

where $\ell(Y,\tilde{Y})$ is the 0-1 loss function.

It is easy to show that Problem Formulation 1 is a linear programming problem with $P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$ as variables. Note that distribution $P(\tilde{y}|\mathsf{pa}(\hat{Y})_{\mathcal{G}})$ can be obtained by $P(\tilde{y}|\mathsf{pa}(\hat{Y})_{\mathcal{G}}) = \sum_{\hat{y}} P(\hat{y}|\mathsf{pa}(\hat{Y})_{\mathcal{G}}) P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$. Thus, all constraints are linear w.r.t. $P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$. On the other hand, for the objective function we have

$$\mathbb{E}_{\mathcal{D}}[\ell(Y,\tilde{Y})] = \sum_{y,\tilde{y}\in\{y^+,y^-\}} \ell(y,\tilde{y}) P(\tilde{y},y) = 2P(\tilde{y}\neq y).$$

And we also have

$$\begin{aligned}
&P(\tilde{y}\neq y) = P(\hat{y}\neq y)P(\tilde{y}=\hat{y}) + P(\hat{y}=y)P(\tilde{y}\neq\hat{y}) \\
&= \sum_{\mathbf{x},s} P(\mathbf{x},s) \Bigg[P(\hat{y}\neq y|\mathbf{x},s) \left[\begin{smallmatrix} P(\tilde{y}=y^-|\hat{y}=y^-,\mathbf{x},s) \\ P(\hat{y}=y^-|\mathbf{x},s) \end{smallmatrix} + \begin{smallmatrix} P(\tilde{y}=y^+|\hat{y}=y^+,\mathbf{x},s) \\ P(\hat{y}=y^+|\mathbf{x},s) \end{smallmatrix} \right] \\
&+ P(\hat{y}=y|\mathbf{x},s) \left[\begin{smallmatrix} P(\tilde{y}=y^+|\hat{y}=y^-,\mathbf{x},s) \\ P(\hat{y}=y^-|\mathbf{x},s) \end{smallmatrix} + \begin{smallmatrix} P(\tilde{y}=y^-|\hat{y}=y^+,\mathbf{x},s) \\ P(\hat{y}=y^+|\mathbf{x},s) \end{smallmatrix} \right] \Bigg]
\end{aligned}$$

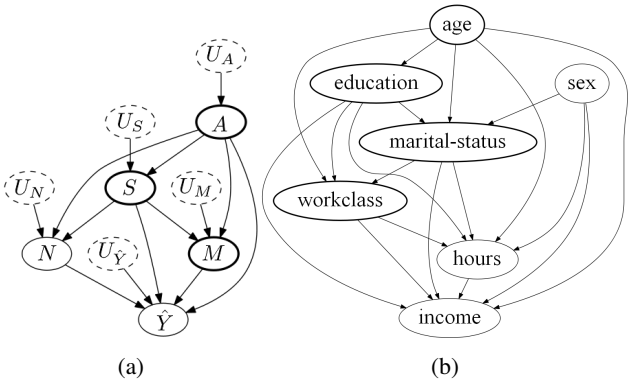


Figure 2: (a) Causal graph for the synthetic data. (b) Causal graph for the Adult data. Dashed nodes represent the exogenous variables. Bold nodes represent the profile attributes in $\mathbf{Z}$.

In the above expression, all probabilities except $P(\tilde{y}|\hat{y},\mathbf{x},s)$ are read from the training set $\mathcal{D}$, making it a linear expression of $P(\tilde{y}|\hat{y},\mathbf{x},s)$.

5 Experiments

We evaluate our method and compare it with previous methods on two datasets. To show the correctness of our method, we generate a synthetic dataset from a known causal model with complete knowledge in our evaluation. We also use the Adult dataset [Lichman, 2013] to evaluate these methods in a real-world environment. We evaluate four methods for constructing classifiers: (1) the original learning algorithm without fairness constraints as the baseline (denoted by **BL**); (2) two methods (denoted by **A1** and **A3**) from [Kusner *et al.*, 2017] where **A1** uses non-descendants of S only for building classifiers, and **A3** presuppose the additive noise model for estimating the noise terms, which are then used for building classifiers; (3) our method (denoted as **CF**). By default, the discrimination threshold τ is set as 0.05.

# of $\mathbf{z}$	$DE(\hat{y}_{s^-\to s^+}|\mathbf{z})$		
	ub	lb	$Truth$
1	0.399	0.105	0.328
2	0.471	0.177	0.467
3	0.147	-0.082	-0.038
4	0.374	0.145	0.145

Table 1: Bounds and ground truth of counterfactual fairness for all value combinations of $\mathbf{Z}$ using the synthetic data.

5.1 Datasets

Synthetic Data. We manually build a causal model (where all variables are discrete) with complete knowledge of the exogenous variables and the functions (i.e., the contingency table) using Tetrad [Scheines *et al.*, 1998]. The corresponding causal graph is shown in Figure 2a. This causal model consists of 5 endogenous variables, A, S, M, N, Y, and 5 independent exogenous variables, U_A, U_S, U_M, U_N, U_Y. For simplicity, all endogenous variables have two domain values and all exogenous variables have three domain values. The

# of $\mathbf{z}$	LR				SVM			
	BL	**A1**	**A3**	**CF**	**BL**	**A1**	**A3**	**CF**
1	0.000	0.000	**-0.233**	0.049	**0.114**	0.000	**0.174**	0.049
2	**1.000**	0.000	**1.000**	0.049	**0.762**	0.000	**0.648**	0.049
3	0.000	0.000	0.000	0.000	-0.021	0.000	-0.021	0.000
4	**1.000**	0.000	0.000	0.048	**1.000**	0.000	0.000	0.048

Table 2: Counterfactual fairness for prediction of the synthetic data. Values violating the threshold are highlighted in bold.

Accu. (%)	Data	**BL**	**A1**	**A3**	**CF**
LR	Train	60.103	55.760	59.433	61.987
	Test	60.421	56.563	59.713	62.512
SVM	Train	65.710	55.760	62.466	61.977
	Test	65.841	56.563	62.542	62.463

Table 3: Prediction accuracy for the synthetic data.

distributions of the exogenous variables and the deterministic functions of the endogenous variables are randomly assigned. Then, we generate 100,000 examples from this causal model and split the data into training and testing sets with a ratio of 80/20. We consider S as the sensitive attribute and Y as the decision attribute. The profile attribute set $\mathbf{Z}$ contains A, M.

Adult Data. This dataset consists of 65,123 records with 11 attributes including *education*, *sex*, *income* etc.. We select 7 attributes, binarize their domain values, and split the dataset into the training and testing sets, following the 80/20 ratio. We apply the PC algorithm implemented in Tetrad to build the causal graph while the significant threshold is set as 0.01 for conditional independence testing. We use three tiers in the partial order for temporal priority: *sex*, *age* in Tier 1, *education*, *marital-status* and *workclass* are defined in Tier 2, and *income* defined in Tier 3. The causal graph is shown in Figure 2b, where *sex* is considered as the sensitive attribute and *income* is the decision attribute. *age*, *education*, *marital-status*, and *workclass* are contained the profile attributes $\mathbf{Z}$.

5.2 Experiment on Synthetic Data

Quantifying Counterfactual Fairness. According to Theorem 1, the counterfactual fairness quantity is unidentifiable in this dataset. We evaluate the bounds of counterfactual fairness using Theorem 2. The ground truth (i.e., the exact values of all counterfactual quantities) is computed by applying the Abduction-Action-Prediction method. The results are shown in Table 1, where the first column indicates the indices of $\mathbf{z}$'s value combinations. As can be seen, the exact values of $DE(\hat{y}_{s^- \to s^+}|\mathbf{z})$ fall into the range of our bounds for all value combinations of $\mathbf{Z}$, which validates our theorem.

Building Counterfactually Fair Classifiers. We then evaluate the classifier learning methods. For the baseline method, we adopt the logistic regression (**LR**) and support vector machine (**SVM**). Then, we apply **A1**, **A3**, and **CF** on top of both classifiers. The counterfactual fairness is precisely evaluated and shown in Table 2 for all the methods using the Abduction-Action-Prediction method. The predictive accuracy is reported in Table 3. As expected, both **A1** and **CF** achieve fairness, but our method achieves higher accuracy than **A1**, implying that **A1** loses more information. On the other hand, we see that **BL** fails to achieve counterfactual fairness, because it ignores the fairness during the training. In addition, **A3** also fails to achieve counterfactual fairness. This implies that assuming additive model may produce biased results when the underlying causal model is non-linear.

	# of $\mathbf{z}$	**BL**		**A1**	**A3**		**CF**	
		ub	lb	val	ub	lb	ub	lb
LR	2	0.321	0.000	0.000	**-1.000**	**-1.000**	-0.007	-0.047
	4	0.523	0.000	0.000	**-1.000**	**-1.000**	0.038	-0.027
	13	**1.000**	**0.304**	0.000	0.000	-1.000	0.049	-0.016
	15	**1.000**	**0.398**	0.000	0.000	-1.000	0.050	-0.007
SVM	2	0.135	-0.186	0.000	**-0.186**	**-0.186**	-0.007	-0.047
	4	0.283	-0.240	0.000	**-0.240**	**-0.240**	0.038	-0.027
	13	**0.866**	**0.170**	0.000	**0.866**	**0.170**	0.049	-0.016
	15	**0.907**	**0.305**	0.000	**0.907**	**0.305**	0.050	-0.007

Table 4: Counterfactual fairness for prediction of the Adult data.

Accu. (%)	D	**BL**	**A1**	**A3**	**CF**
LR	Train	77.728	67.624	74.845	70.433
	Test	77.200	66.934	73.867	69.451
SVM	Train	78.071	67.624	77.845	70.413
	Test	77.449	66.934	77.166	69.438

Table 5: Prediction accuracy for the Adult data.

5.3 Experiment on Adult Dataset

We evaluate the fair classifier learning methods using the Adult dataset. Since we don't have the ground truth, we report bounds of counterfactual fairness for different methods. Table 4 shows that only **A1** and **CF** can achieve counterfactual fairness for all value combinations of $\mathbf{Z}$, but our **CF** consistently achieves higher accuracy than **A1** as shown in Table 5. This is as expected since **A1** is proved to be fair in [Kusner *et al.*, 2017] (and also identifiable according to Theorem 1), but will inevitably lead to lower accuracy as only S's non-descendants are used. For **BL** and **A3** in Table 4, either the lower bound is larger than τ or the upper bound is less than $-\tau$, indicating the τ-counterfactual fairness is not achieved.

6 Conclusion

We focus on the unidentifiability challenge when applying counterfactual fairness in practice. We decompose the counterfactual quantity and identify the source of unidentification by leveraging the counterfactual graph and c-component factorization from Pearl's framework. We then develop the criterion of identification and the upper/lower bounds for counterfactual fairness. Finally, we formulate counterfactually fair classification as a linear programming problem. Empirical evaluations show our method is guaranteed to achieve counterfactual fairness in classification, while previous approaches either cannot achieve counterfactual fairness or suffer bad performance due to over-simplified assumptions.

Acknowledgments

This work was supported in part by NSF 1646654.

References

[Chiappa, 2019] Silvia Chiappa. Path-Specific Counterfactual Fairness. *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[Kilbertus *et al.*, 2017] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding Discrimination through Causal Reasoning. *Neural Information Processing Systems*, 2017.

[Kusner *et al.*, 2017] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. *Neural Information Processing Systems*, 2017.

[Li *et al.*, 2017] Jiuyong Li, Jixue Liu, Lin Liu, Thuc Duy Le, Saisai Ma, and Yizhao Han. Discrimination detection by causal effect estimation. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1087–1094. IEEE, December 2017.

[Lichman, 2013] M Lichman. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, 2013.

[Nabi and Shpitser, 2018] Razieh Nabi and Ilya Shpitser. Fair Inference On Outcomes. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 1931–1940, 2018.

[Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[Scheines *et al.*, 1998] Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The TETRAD Project: Constraint Based Aids to Causal Model Specification. *Multivariate Behavioral Research*, 33(1):65–117, January 1998.

[Shpitser and Pearl, 2008] Ilya Shpitser and Judea Pearl. Complete Identification Methods for the Causal Hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

[Tian and Pearl, 2002] Jin Tian and Judea Pearl. A General Identification Condition for Causal Effects. *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence (IAAI 2002)*, (August):567–573, 2002.

[Zhang and Bareinboim, 2018] Junzhe Zhang and Elias Bareinboim. Fairness in Decision-Making – The Causal Explanation Formula. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[Zhang *et al.*, 2017a] Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving Non-Discrimination in Data Release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 1335–1344, New York, New York, USA, November 2017. ACM Press.

[Zhang *et al.*, 2017b] Lu Zhang, Yongkai Wu, and Xintao Wu. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3929–3935, California, August 2017. International Joint Conferences on Artificial Intelligence Organization.

[Zhang *et al.*, 2018a] 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*, pages 1–1, 2018.

[Zhang *et al.*, 2018b] Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving non-discrimination in prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 3097–3103, 2018.