TECHNICAL REPORT
R-37-L
October, 2018

Equality of Opportunity in Classification: A Causal Approach

Junzhe Zhang
Purdue University, USA
zhang745@purdue.edu

Elias Bareinboim
Purdue University, USA
eb@purdue.edu

Abstract

The *Equalized Odds* (for short, EO) is one of the most popular measures of discrimination used in the supervised learning setting. It ascertains fairness through the balance of the misclassification rates (false positive and negative) across the protected groups – e.g., in the context of law enforcement, an African-American defendant who would not commit a future crime will have an *equal opportunity* of being released, compared to a non-recidivating Caucasian defendant. Despite this noble goal, it has been acknowledged in the literature that statistical tests based on the EO are oblivious to the underlying causal mechanisms that generated the disparity in the first place (Hardt et al. 2016). This leads to a critical disconnect between statistical measures readable from the data and the meaning of discrimination in the legal system, where compelling evidence that the observed disparity is tied to a specific causal process deemed unfair by society is required to characterize discrimination. The goal of this paper is to develop a principled approach to connect the statistical disparities characterized by the EO and the underlying, elusive, and frequently unobserved, causal mechanisms that generated such inequality. We start by introducing a new family of counterfactual measures that allows one to explain the misclassification disparities in terms of the underlying mechanisms in an arbitrary, non-parametric structural causal model. This will, in turn, allow legal and data analysts to interpret currently deployed classifiers through causal lens, linking the statistical disparities found in the data to the corresponding causal processes. Leveraging the new family of counterfactual measures, we develop a learning procedure to construct a classifier that is statistically efficient, interpretable, and compatible with the basic human intuition of fairness. We demonstrate our results through experiments in both real (COMPAS) and synthetic datasets.

1 Introduction

The goal of supervised learning is to provide a statistical basis upon which individuals with different group memberships can be reliably classified. For instance, a bank may want to learn a function from a set of background factors so as to determine whether a customer will repay her loan; a university may train a classifier to predict the future GPA of an applicant to decide whether to accept her into the program. The growing adoption of automated systems based on standard classification algorithms throughout society (including in law enforcement, education, and finance [14, 4, 9, 22, 1]) has raised concerns about potential issues due to unfairness and discrimination.

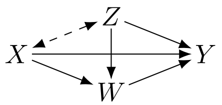
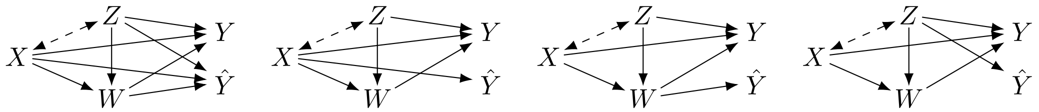


Figure 1: COMPAS

A recent high-profile example is a risk assessment tool called COMPAS, which has been widely used across the US to inform decisions in the criminal justice system. Fig. 1 graphically describes this setting – X represents the race (0 for Caucasian, 1 for African-American) of a defendant and Y stands for the recidivism outcome (0 for no, 1 otherwise), which are *mediated* by the prior convictions W, and *confounded* by other demographic information Z (e.g., age, gender) of the defendant. The COMPAS

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.



(a) $f(x, z, w)$ (b) $f_1(x) = x$ (c) $f_2(w) = w$ (d) $f_3(z) = z$

Figure 2: (a-d) Causal diagrams of classifiers f, f_1, f_2, f_3 in COMPAS. Nodes represent variables, directed arrows for functional relationships, and bi-directed arrows for unknown associations.

tool is a classifier $f(x, z, w)$ (shown in Fig. 2(a)) providing a prediction $\hat{Y}$ on whether the defendant is expected to commit a future crime. An analysis performed by the news organization ProPublica revealed that the odds of receiving a positive prediction ($\hat{Y} = 1$) for defendants who did not recidivate were on average higher among African-Americans than their Caucasians counterparts [1]. In words, the error rates of COMPAS disproportionately misclassified African-American defendants.

Many attempts have been made to model discrimination in the classification setting [27, 15, 12, 10, 16]. A recent, noteworthy framework comes under the rubric of *Equalized Odds* [8] (also referred to as *Error Rate Balance* [5]), which constrains the classification algorithm such that its disparate error rate $\textit{ER}_{x_0,x_1}(\hat{y}|y) = P(\hat{y}|x_1, y) - P(\hat{y}|x_0, y)$ is *equalized* (and equal to 0) across different demographics x_0, x_1, i.e., the odds of misclassification does not disproportionately affect any population sub-group. In the COMPAS example, the condition $\textit{ER}_{x_0,x_1}(\hat{Y} = 1|Y = 0) = 0$ implies that an African-American defendant who does not commit a future crime will have an *equal opportunity* of getting released, compared to non-recidivating Caucasian defendants. This notion of fairness is natural in many learning settings and, indeed, has been implemented in a number of algorithms [8, 7, 26, 24].

Unfortunately, the framework of equalized odds is not without its problems. To witness, consider a binary instance of Fig. 1 where the values of X and Z are determined such that $x = z$ and W is decided by the function $w \leftarrow x$. We are concerned with the ER disparity induced by different classifiers f_1, f_2, f_3 (Fig. 2(b-d)), where, for instance, $\hat{y} \leftarrow f_1(x) = x$ (i.e., f_1 takes only X as input, and ignores the other features). Remarkably, a simple analysis shows that $\textit{ER}_{x_0,x_1}(\hat{Y} = 1|Y = 0)$ is the same (and equal to 1) in all three classifiers, despite their fundamentally different mechanisms associating X and $\hat{Y}$. Note that f_1, f_2, f_3 corresponds to the direct path $X \to \hat{Y}$, the indirect path $X \to W \to \hat{Y}$, and the remaining spurious (non-causal) paths (e.g., $X \leftrightarrow Z \to \hat{Y}$), respectively.

This observation is not entirely new, and is part of a pattern noted by [8] – statistical tests based on the disparate ER are oblivious to the underlying causal mechanisms that generated the data. This realization has dramatic implications to the applicability of supervised learning in the real world since it seems to suggest that commonsense notions of discrimination, for example, the unequalized false positive rate *caused* by direct discrimination ($X \to \hat{Y}$), cannot be formally articulated, measured from data, and, therefore, controlled. More importantly, the legal frameworks of anti-discrimination laws in the US (e.g., Title VII) require that to establish a *prima facie* case of discrimination, the plaintiff must demonstrate *"a strong causal connection"* between the alleged discriminatory practice and the observed statistical disparity, otherwise the case will be dismissed (Texas Dept. of Housing and Community Affairs v. Inclusive Communities Project, Inc., 576 U.S. __ (2015)). Without a robust causal basis, an evidence of disparate ER on its own is not sufficient to lead to any legal liability.

More recently, the use of causal reasoning to help open the black-box of decision-making systems has attracted considerable interest in the community, leading to fine-grained explanations of observed statistical biases [12, 11, 26, 10]. One of the main tasks of causal inference is to explain "how nature works," or more technically, to decompose a composite statistical measure (e.g, the total variation $\textit{TV}_{x_0,x_1}(\hat{y}) = P(\hat{y}|x_1) - P(\hat{y}|x_0)$), into its most elementary and interpretable components [25, 18, 29]. In particular, [28] introduced the *causal explanation formula*, which allows fairness analysts to decompose TV into detailed counterfactual measures describing the effects along direct, indirect, and spurious paths from X to $\hat{Y}$. While [28] explains how the statistical inequality in the observed outcome is brought about, it is unclear how to apply such insight to correct the problematic behaviors of an alleged, discriminatory policy. Furthermore, the explanation formula allows the decomposition of marginal measures such as TV, but it's unable to explain disparities represented by conditional ones, such as the ER (e.g., non-recidivating African-American defendants).

This paper aims to overcome these challenges. We develop a causal framework to link the disparities realized through the ER and the (unobserved) causal mechanisms by which the protected attribute X

affects change in the prediction $\hat{Y}$. Specifically, (1) we introduce a family of counterfactual measures capable of describing the ER in terms of the direct, indirect, and spurious paths from X to $\hat{Y}$ on an arbitrary structural causal model (Defs. 1-3) and we prove different qualitative and quantitative properties of these measures (Thms. 1-2); (2) we derive adjustment-like formulas to estimate the counterfactual ERs from observational data (Thms. 3-4), which are accompanied with an efficient algorithm (Alg. 1, Thm. 5) to find the corresponding admissible sets; (3) we operationalize the proposed counterfactual estimands through a novel procedure to learn a fair classifier subject to constraints over the effect along the underlying causal mechanisms (Algs. 2-3, Thm. 6).

2 Preliminaries and Notations

We use capital letters to denote variables (X), and small letters for their values (x). We use the abbreviation $P(x)$ to represent the probabilities $P(X = x)$. For arbitrary sets $\boldsymbol{A}$ and $\boldsymbol{B}$, let $\boldsymbol{A} \backslash \boldsymbol{B}$ denote the set difference $\{x : x \in \boldsymbol{A} \text{ and } x \notin \boldsymbol{B}\}$, and let $|\boldsymbol{A}|$ be the dimension of set $\boldsymbol{A}$.

The basic semantical framework of our analysis rests on *structural causal models* (SCM) [17, Ch. 7]. A SCM is a tuple $\langle M, P(\boldsymbol{u})\rangle$, where M consists of a set of endogenous (observed) variables $\boldsymbol{V}$ and exogenous (unobserved) variables $\boldsymbol{U}$. The values of each $V_i \in \boldsymbol{V}$ are determined by a structural function f_{V_i} taking as arguments a combination of other endogenous and exogenous variables (i.e., $V_i \leftarrow f_{V_i}(PA_i, U_i), PA_i \subseteq \boldsymbol{V}, U_i \subseteq \boldsymbol{U}$)). Values of U are drawn from the distribution $P(\boldsymbol{u})$. Each SCM is associated with a directed acyclic graph (DAG) $G = \langle \boldsymbol{V}, \boldsymbol{E}\rangle$, termed a causal diagram, where nodes $\boldsymbol{V}$ represent endogenous variables and directed edges $\boldsymbol{E}$ stand for functional relations (e.g., see Fig. 1). By convention, $\boldsymbol{U}$ are not explicitly shown; a bi-directed arrow between V_i and V_j indicates the presence of an unobserved confounder (UC) U_k affecting both V_i, V_j, i.e., $V_i \leftarrow U_k \rightarrow V_j$.

A path is a sequence of edges where each pair of adjacent edges in the sequence share a node. We use d-separation and blocking interchangeably, following the convention in [17]. A path from a node X to a node $\hat{Y}$ consists exclusively of direct arrows pointing away from X is called causal; all the other non-causal paths are called spurious. The causal paths could be further categorized into the *direct* path $X \rightarrow \hat{Y}$ and the *indirect* paths, e.g., $X \rightarrow W \rightarrow \hat{Y}$ of Fig. 2(a). Let $(X \rightarrow \hat{Y})_G$, $(X \overset{i}{\rightarrow} \hat{Y})_G$ and $(X \overset{s}{\leftrightarrow} \hat{Y})_G$ denote, respectively, the direct, indirect and spurious paths between X and $\hat{Y}$ in a DAG G. A descendant of X is any node which X has a causal path to (including X itself). The descendant set of a set $\boldsymbol{X}$ is all descendants of any node in $\boldsymbol{X}$, which we denote by $De(\boldsymbol{X})_G$.

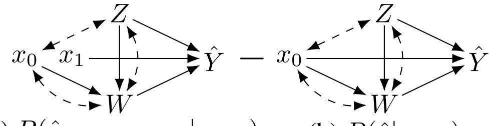
An intervention on a set of variables $X \subseteq \boldsymbol{V}$, denoted by $do(x)$, is an operation where values of X are set to constants x, regardless of how they were ordinarily determined (through the functions f_X). We denote by $\langle M_x, P(\boldsymbol{u})\rangle$ a sub-model of a SCM $\langle M, P(u)\rangle$ induced by $do(x)$. The potential response of $\hat{Y}$ to intervention $do(x)$, denoted by $\hat{Y}_x(\boldsymbol{u})$, is the solution of $\hat{Y}$ with $\boldsymbol{U} = \boldsymbol{u}$ in the sub-model M_x; it can be read as the counterfactual sentence "the value that $\hat{Y}$ would have obtained in situation $\boldsymbol{U} = \boldsymbol{u}$, had X been x." Statistically, averaging $\boldsymbol{U}$'s distribution ($P(\boldsymbol{u})$) leads to the counterfactual variable $\hat{Y}_x$. For a more detailed discussion on SCMs, please refer to [17, 2].

3 Counterfactual Analysis of Unequalized Classification Errors

In this section, we investigate the unequalized odds of misclassification observed in COMPAS by devising three simple thought experiments. These experiments could be generalized into a set of novel counterfactual measures, providing a fine-grained explanation of how the ER disparity of a classifier $f(\hat{\boldsymbol{pa}})$ is brought about. Throughout our analysis, we will let X be the protected attribute, $\hat{Y}$ be the prediction and Y be the true outcome; $\hat{\boldsymbol{PA}}$ is a set of (possible) input features of the predictor $\hat{Y}$. We will denote by value x_1 the disadvantaged group and x_0 the advantaged group. Given the space constraints, all proofs are included in Appendix A.

We consider first the impact of the direct discrimination (i.e., the direct path $X \rightarrow \hat{Y}$) on the ER disparity observed in the COMPAS. We will devise a thought experiment concerning with a Caucasian defendant who does not recidivate (i.e., x_0, y). Imagine a hypothetical situation where this defendant were a non-recidivating African-American (x_1, y), while keeping the prior convictions W and other demographic information Z fixed at the level that the defendant x_0, y currently has. We then measure the prediction $\hat{Y}$ in this imagined world (counterfactually), compared to what the defendant currently receives from COMPAS (factually). If the prediction were different in these two situations, e.g., $\hat{Y}$

changes from 0 to 1, we could then say the path $X \to \hat{Y}$ is active, i.e., the direct discrimination against African-American defendants exists.



(a) $P(\hat{y}_{x_1,y,W_{x_0,y},Z}|x_0,y)$ (b) $P(\hat{y}|x_0,y)$

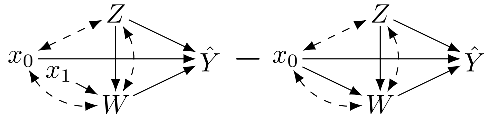
Figure 3: Graphical representation of the counterfactual direct ER in COMPAS.

Figs. 8(a-b) represent this thought experiment graphically. Fig. 8(b) shows the conditional SCM $\langle M, P(\boldsymbol{u}|x_0,y)\rangle$ of the non-recidivating Caucasian defendant (x_0,y): variables X, Z, W are correlated by conditioning on the collider Y [17, pp. 339]; we omit the true outcome Y for simplicity. Using this model as the baseline (i.e., what factually happened in reality), we change in Fig. 8(a) the input of X to the direct path $X \to \hat{Y}$ to x_1 (edges in G represent functional relations), while keeping the value of X to other variables (W, Z) fixed at the baseline level x_0, y. In this reality, variable $Z_{x_0,y} = Z$ since Z is a non-descendant node of X and Y [17, pp. 232]; the intervention on Y is omitted since Y does not directly affect the prediction $\hat{Y}$. Since the direct path $X \to \hat{Y}$ is the only difference between models of Figs. 8(a-b), the change in $\hat{Y}$ thus measure the influence of $X \to \hat{Y}$. Indeed, this hypothetical procedure could be generalized, applicable to any classifier in an arbitrary SCM, which we summarize as follows.

Definition 1 (Counterfactual Direct Error Rate). Given a SCM $\langle M, P(\boldsymbol{u})\rangle$ and a classifier $f(\hat{\boldsymbol{pa}})$, the counterfactual direct error rate for a sub-population x, y (with prediction $\hat{y} \neq y$) is defined as:

$$ER^{d}_{x_0,x_1}(\hat{y}|x,y) = P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}}|x,y) - P(\hat{y}_{x_0,y}|x,y) \tag{1}$$

In Eq. 1, $\hat{Y}_{x_1,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}}$ could be further simplified as $\hat{Y}_{x_1,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}}$ since Y is not an input of $f(\hat{\boldsymbol{pa}})$. The subscript $(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}$ is the solution of the input features (besides X) $(\hat{\boldsymbol{PA}}\backslash X)(\boldsymbol{u})$ in the sub-model $M_{x_0,y}$; values of $\boldsymbol{U}$ are drawn from the distribution $P(\boldsymbol{u})$ such that $X(\boldsymbol{u}) = x, Y(\boldsymbol{u}) = y$. The query of Eq. 1 could be read as: "For an individual with the protected attribute $X = x$ and the true outcome $Y = y$, how would the prediction $\hat{Y}$ change had X been x_1, while keeping all the other features $\hat{\boldsymbol{PA}}\backslash X$ at the level that they would attain had $X = x_0$ and $Y = y$, compared to the prediction $\hat{Y}$ she/he would receive had X been x_0 and Y been y?"



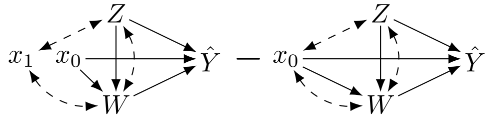
(a) $P(\hat{y}_{x_0,y,W_{x_1,y},Z}|x_0,y)$ (b) $P(\hat{y}|x_0,y)$

Figure 4: Graphical representations of the counterfactual indirect ER in COMPAS.

Similarly, we could devise a thought experiment to measure the effect of the indirect discrimination, mediated by the prior convictions W, i.e., the indirect path $X \to W \to \hat{Y}$. Consider again the non-recidivating Caucasian defendant x_0, y. We conceive a scenario where the prior convictions W of the defendant x_0, y changes to the level that it would have achieved had the defendant been a non-recidivating African-American x_1, y, while keeping the other features X, Z fixed at the level that they currently are. Fig. 4(a) describes this hypothetical scenario: we change only input value of edge $X \to W$ to x_1, while keeping all the other paths untouched (at the baseline). We then measure the prediction $\hat{Y}$ in both the counterfactual (Fig. 4(a)) and factual (Fig. 4(b)) world and compare their differences. The change in the prediction of these models thus represent the influence of indirect path $X \to W \to \hat{Y}$. We generalize this thought experiment and provide an estimand of the indirect paths for any SCM and classifier f, namely:

Definition 2 (Counterfactual Indirect Error Rate). Given a SCM $\langle M, P(\boldsymbol{u})\rangle$ and a classifier $f(\hat{\boldsymbol{pa}})$, the counterfactual indirect error rate for a sub-population x, y (with prediction $\hat{y} \neq y$) is defined as:

$$ER^{i}_{x_0,x_1}(\hat{y}|x,y) = P(\hat{y}_{x_0,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_1,y}}|x,y) - P(\hat{y}_{x_0,y}|x,y). \tag{2}$$



(a) $P(\hat{y}_{x_0,y}|x_1,y)$ (b) $P(\hat{y}_{x_0,y}|x_0,y)$

Figure 5: Graphical representations of the counterfactual spurious ER in COMPAS.

Finally, we introduce a hypothetical procedure measuring the influence of the spurious relations between the protected attribute X and prediction $\hat{Y}$ through the population attributes that are non-descendants of both X and $\hat{Y}$, e.g., the path $X \leftrightarrow Z \to \hat{Y}$ in Fig. 2(a). We consider a Caucasian x_0, y and an African-American x_1, y defendants who both would not recidivate. We measure the prediction $\hat{Y}$ these defendants would receive had they both been non-recidivating Caucasians (x_0, y). Figs. 5 (a-b) describes this experimental setup. Since the causal influence of X (on $\hat{Y}$) are fixed at x_0 in both models, the difference in $\hat{Y}$ must be due to the population characteristics that are not affected by X i.e., the spurious $X - \hat{Y}$ relationships.

Definition 3 (Counterfactual Spurious Error Rate)**.** Given a SCM $\langle M, P(\boldsymbol{u})\rangle$ and a classifier $f(\hat{\boldsymbol{pa}})$, the counterfactual spurious error rate for a sub-population x, y (with prediction $\hat{y} \neq y$) is defined as:

$$\textit{ER}^{s}_{x_0,x_1}(\hat{y}|y) = P(\hat{y}_{x_0,y}|x_1, y) - P(\hat{y}_{x_0,y}|x_0, y) \tag{3}$$

Def. 3 generalizes the thought experiment described above to an arbitrary SCM. In the above equation, the distribution $P(\hat{y}_{x_0,y}|x_0, y)$ coincides with $P(\hat{y}|x_0, y)$ since variable $\hat{Y}_{x_0,y} = \hat{Y}$ given that $X = x_0, Y = y$ (the composition axiom [17, Ch. 7.3]). Eq. 3 can be read as the counterfactual sentence: "For two demographics x_0, x_1 with the same true outcome $Y = y$, how would the prediction $\hat{Y}$ differ had they both been x_0, y?"

3.1 Properties of Counterfactual Error Rates

Theorem 1. *Given a SCM $\langle M, P(\boldsymbol{u})\rangle$ and a classifier $f(\hat{\boldsymbol{pa}})$, for any $x_0, x_1, x, \hat{y}, y$, the counterfactual ERs of Defs. 1-3 obey the following properties : (1) $(X \not\rightarrow Y)_{G_{|Y}} \Rightarrow \textit{ER}^{d}_{x_0,x_1}(\hat{y}|x, y) = 0$; (2) $|(X \overset{i}{\rightarrow} Y)_{G_{|Y}}| = 0 \Rightarrow \textit{ER}^{i}_{x_0,x_1}(\hat{y}|x, y) = 0$; (3) $|(X \overset{s}{\leftrightarrow} Y)_{G_{|Y}}| = 0 \Rightarrow \textit{ER}^{s}_{x_0,x_1}(\hat{y}|x, y) = 0$, where $G_{|Y}$ is the causal diagram of a conditional SCM $\langle M_y, P(\boldsymbol{u}|y)\rangle$.*

The conditional causal diagram $G_{|Y}$ is obtained from the original model G by (1) removing the node Y and (2) adding bi-directed arrows between nodes whose associated exogenous variables are correlated in $P(\boldsymbol{u}|y)$[1] (e.g., Fig. 8(b)). Thm. 1 says that Defs. 1-3 provide *prima facie* evidence for discrimination detection. For instance, $\textit{ER}^{d}_{x_0,x_1}(\hat{y}|x, y) \neq 0$ implies that the path $X \to \hat{Y}$ is active, i.e., the direct discrimination exists. It is expected that the proposed counterfactual measures capture the relative strength of different active pathways connecting node X and $\hat{Y}$ in the underlying SCM. We now derive how the counterfactual ERs are quantitatively related with the unequalized odds of misclassification induced by an arbitrary classifier.

Theorem 2 (Causal Explanation Formula of Equalized Odds)**.** *For any $x_0, x_1, \hat{y}, y$, $\textit{ER}_{x_0,x_1}(\hat{y}|x, y)$, $\textit{ER}^{d}_{x_0,x_1}(\hat{y}|x, y)$, $\textit{ER}^{i}_{x_0,x_1}(\hat{y}|x, y)$ and $\textit{ER}^{s}_{x_0,x_1}(\hat{y}|y)$ obey the following non-parametric relationship:*

$$\textit{ER}_{x_0,x_1}(\hat{y}|y) = \textit{ER}^{d}_{x_0,x_1}(\hat{y}|x_0, y) - \textit{ER}^{i}_{x_1,x_0}(\hat{y}|x_0, y) - \textit{ER}^{s}_{x_1,x_0}(\hat{y}|y). \tag{4}$$

Thm. 2 guarantees that the disparate ER with the transition from x_0 to x_1 is equal to the sum of the counterfactual direct ER with this transition, *minus* the indirect and spurious ER with *reverse* transition, from x_1 to x_0, on the sub-population x_0, y. Together with Thm. 1, each decomposing term in Eq. 4 thus estimates the adverse impact of its corresponding discriminatory mechanism on the total ER disparity. For instance, in COMPAS, $\textit{ER}^{d}_{x_0,x_1}(\hat{y}_1|x_0, y)$ explains how much the direct racial discrimination accounts for the unequalized false positive rate $\textit{ER}_{x_0,x_1}(\hat{y}_1|y_0)$ between non-recidivating African American (x_1, y) and Caucasian (x_0, y) defendants. Perhaps surprisingly, this result holds non-parametrically, which means that the counterfactual ERs decompose following Thm. 2 for any functional form of the classifier and the underlying causal models where the dataset was generated. Owed to their generality and ubiquity, we refer to this equation as the "Causal Explanation Formula" for the disparate ER in classification tasks.

Connections with Other Counterfactual Measures Defs. 1-3 can be seen as a generalization of the marginal counterfactual measures, including the counterfactual effects introduced in [28] and the natural effects in [18, 12, 16]. Unable to consider the additional evidence (in classification, the true outcome $Y = y$), the fairness analysis framework based on these marginal measures fails to provide a fine-grained quantitative explanation of the ER disparity (as in, Thm. 2). The counterfactual fairness [11] is another counterfactual measure. As noted in [28], however, it considers only the effects along the causal paths from the protected attribute X and the outcome $\hat{Y}$, thus unable to provide a full account of the $X - \hat{Y}$ associations, including the spurious relations. We provide in Appendix B a more detailed discussion about the relationships between our measures and the existing ones.

[1] $G_{|Y}$ explicitly represents the change of information flow due to conditioning on the true outcome Y: the information via arrows pointing away from Y is intercepted; measuring the collider Y makes its (marginally independent) common causes dependent, also known as the "explaining away" effect [17, pp. 339].

4 Estimating Counterfactual Error Rates

The Explanation Formula provides the precise relation between the counterfactual ERs, but it does not specify how they should be estimated from data. When the underlying SCM is provided, the counterfactual direct, indirect and spurious ERs (Defs. 1-3) are all well-defined and computable via the three-step algorithm of "predictions, interventions and counterfactuals" described in [17, Ch. 7.1].

However, the SCMs are not fully known in many applications, and one must estimate the proposed counterfactual measures from the passively-collected (observational) data. Let a classifier $f(\hat{\boldsymbol{pa}})$ be denoted by $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$, where $\hat{\boldsymbol{Z}} \subseteq \hat{\boldsymbol{PA}}$ are non-descendants of both X and Y and the subset of features $\hat{\boldsymbol{W}} = \hat{\boldsymbol{PA}} \backslash \hat{\boldsymbol{Z}}$. We first characterize a set of classifiers where such estimation is still feasible.

Definition 4 (Explanation Criterion)**.** Given a DAG G and a classifier $\hat{y} \leftarrow f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$, a set of covariates $\boldsymbol{C}$ satisfies the *explanation criterion* relative to f (called the explaining set) if and only if (1) $\hat{\boldsymbol{Z}} \subseteq \boldsymbol{C}$; (2) $\boldsymbol{C} \cap \mathit{Forb}(\{X, Y\}, \hat{\boldsymbol{W}} \backslash X) = \emptyset$ where $\mathit{Forb}(\{X, Y\}, \hat{\boldsymbol{W}} \backslash X)$ is a set of descendants $W_i \in \mathit{De}(W)_G$ for some $W \not\in \{X, Y\}$ on a proper causal path[2] from $\{X, Y\}$ to $\hat{\boldsymbol{W}} \backslash X$ in G; and (3) all spurious paths from $\{X, Y\}$ to $\hat{\boldsymbol{W}} \backslash X$ in G are blocked by $\boldsymbol{C}$. A classifier f is *counterfactually explainable* (ctf-explainable) if and only if it has an explaining set $\boldsymbol{C}$ satisfying Conditions 1-3.

Consider again the COMPAS model of Fig. 1. The classifier $f(x, w, z)$ has input features $\hat{\boldsymbol{W}} = \{X, W\}$ and $\hat{\boldsymbol{Z}} = \{Z\}$. The set $\boldsymbol{C} = \{Z\}$ does not satisfy the explanation criterion relative to f since it does not block the spurious path $Y \leftarrow W$. Indeed, one could show that there exists no set $\boldsymbol{C}$ satisfying Def. 4 relative to f, i.e., $f(x, w, z)$ is not ctf-explainable. However, if we remove the prior convictions W from the feature set, the new classifier $f(x, z)$ is ctf-explainable with $\boldsymbol{C} = \{Z\}$: $\hat{\boldsymbol{Z}} = \boldsymbol{C} = \{Z\}$ satisfies Condition 1; Conditions 2-3 follow immediately since $\hat{\boldsymbol{W}} \backslash X = \emptyset$.

Defs. 4 constitutes a sufficient condition upon which the counterfactual ERs could, at least in principle, be estimated from the observational data. This yields identification formulas as shown next:

Theorem 3. *Given a causal diagram G and a classifier $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$, if f is ctf-explainable (Def. 4) with an explaining set $\boldsymbol{C}$, $\mathit{ER}^d_{x_0,x_1}(\hat{y}|x, y)$, $\mathit{ER}^i_{x_0,x_1}(\hat{y}|x, y)$ and $\mathit{ER}^s_{x_0,x_1}(\hat{y}|y)$ can be estimated as follows:*

$$\mathit{ER}^d_{x_0,x_1}(\hat{y}|x, y) = \sum_{\hat{\boldsymbol{w}}, \boldsymbol{c}} (P(\hat{y}_{x_1, \hat{\boldsymbol{w}} \backslash x, \hat{\boldsymbol{z}}}) - P(\hat{y}_{x_0, \hat{\boldsymbol{w}} \backslash x, \hat{\boldsymbol{z}}})) P(\hat{\boldsymbol{w}} \backslash x | x_0, \boldsymbol{c}, y) P(\boldsymbol{c} | x, y), \tag{5}$$

$$\mathit{ER}^i_{x_0,x_1}(\hat{y}|x, y) = \sum_{\hat{\boldsymbol{w}}, \boldsymbol{c}} P(\hat{y}_{x_1, \hat{\boldsymbol{w}} \backslash x, \hat{\boldsymbol{z}}}) (P(\hat{\boldsymbol{w}} \backslash x | x_1, \boldsymbol{c}, y) - P(\hat{\boldsymbol{w}} \backslash x | x_0, \boldsymbol{c}, y)) P(\boldsymbol{c} | x, y), \tag{6}$$

$$\mathit{ER}^s_{x_0,x_1}(\hat{y}|y) = \sum_{\hat{\boldsymbol{w}}, \boldsymbol{c}} P(\hat{y}_{x_1, \hat{\boldsymbol{w}} \backslash x, \hat{\boldsymbol{z}}}) P(\hat{\boldsymbol{w}} \backslash x | x_1, \boldsymbol{c}, y) (P(\boldsymbol{c} | x_1, y) - P(\boldsymbol{c} | x_0, y)). \tag{7}$$

where $P(\hat{y}_{\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}}})$ is well-defined, computable from the classifier $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$[3].

In Eqs. 5-7, the conditional distributions $P(\boldsymbol{c}|x, y)$ and $P(\hat{\boldsymbol{w}} \backslash x | x_0, \boldsymbol{c}, y)$ do not involve any counterfactual variable, which means that they are readily estimable by any method from the observational data (e.g., through deep nets). Continuing from the COMPAS example, we could thus estimate the counterfactual ERs of $f(x, z)$ from the distribution $P(x, y, z, w)$ using Thm. 3 with $\boldsymbol{C} = \{Z\}$.

Inverse Propensity Weighting Estimators Eqs. 5-7 involve summing over all possible values of $\hat{\boldsymbol{W}}, \boldsymbol{C}$, which may present computational and sample complexity challenges as the cardinalities of $\hat{\boldsymbol{W}}, \boldsymbol{C}$ grow very rapidly. There exist robust statistical estimation techniques, known as the inverse propensity weighting (IPW) [13, 19], to circumvent such issues. Given the observed data $\boldsymbol{\mathcal{D}} = \{Y_i, \hat{\boldsymbol{W}}_i, \boldsymbol{C}_i\}_{i=1}^n$, we propose the IPW estimator for $\mathit{ER}^d_{x_0,x_1}(\hat{y}|x, y)$ as follows:

$$\hat{\mathit{ER}}^d_{x_0,x_1}(\hat{y}|x, y) = \frac{1}{n} \sum_{i=1}^{n} (P(\hat{y}_{x_1, \hat{\boldsymbol{W}}_i \backslash X_i, \hat{\boldsymbol{Z}}_i}) - P(\hat{y}_{x_0, \hat{\boldsymbol{W}}_i \backslash X_i, \hat{\boldsymbol{Z}}_i})) \frac{\hat{P}(x | \boldsymbol{C}_i, y) I_{\{X_i = x_0, Y_i = y\}}}{\hat{P}(x_0 | \boldsymbol{C}_i, y) \hat{P}(x, y)}, \tag{8}$$

where $I_{\{\cdot\}}$ is an indicator function and $\hat{P}(x, y)$ is the sample mean estimator of $P(x, y)$ (X, Y are finite). $\hat{P}(x|\boldsymbol{c}, y)$ is a reliable estimator of the conditional distributions $P(x|\boldsymbol{c}, y)$ and, in practice, could be estimated by assuming some parametric models such as logistic regression.

[2] A causal path from $\{X, Y\}$ to $\hat{\boldsymbol{W}} \backslash X$ is proper if it does not intersect $\{X, Y\}$ except at the end point [21].

[3] For a deterministic $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$, the probabilities $P(\hat{y}_{\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}}}) = I_{\{\hat{y} = f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})\}}$ where $I_{\{\cdot\}}$ is an indicator function.

Algorithm 1: FindExpSet

```
Input: Feature set $\{\hat{\boldsymbol{W}}, \hat{\boldsymbol{Z}}\}$, DAG $G = \langle \boldsymbol{V}, \boldsymbol{E} \rangle$
Output: Explaining set $\boldsymbol{C}$ (Def. 4) relative to $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$ in $G$, or $\perp$ if $f$ is not ctf-explainable.
Apply FindSep [23] to find a set $\boldsymbol{C}$ with $\hat{\boldsymbol{Z}} \subseteq \boldsymbol{C} \subseteq \boldsymbol{V} \backslash Forb(\{X, Y\}, \hat{\boldsymbol{W}} \backslash X)$ such that it d-separates $\{X, Y\}$ and $\hat{\boldsymbol{W}} \backslash X$ in $G^{pbd}_{\{X,Y\}, \hat{\boldsymbol{W}} \backslash X}$.
return $\boldsymbol{C}$
```

Algorithm 3: Ctf-FairLearning

```
Input: Samples $\boldsymbol{\mathcal{D}}$, DAG $G$, $\epsilon_d, \epsilon_i, \epsilon_s > 0$
Output: A fair classifier $f$
Let $\boldsymbol{\mathcal{F}} = C\text{-}SFFS(\boldsymbol{\mathcal{D}}, G)$.
Obtain a fair classifier $f$ from $\boldsymbol{\mathcal{F}}$ by solving Eq. 9 subject to $|ER^d| \leq \epsilon_d$, $|ER^i| \leq \epsilon_i$, $|ER^s| \leq \epsilon_s$.
```

Algorithm 2: Causal-SFFS

```
Input: Samples $\boldsymbol{\mathcal{D}} = \{Y_i, \boldsymbol{V}_i\}_{i=1}^n$, a causal diagram $G$
Output: A family of ctf-explainable classifiers $\boldsymbol{\mathcal{F}}$
Initialization: $\hat{\boldsymbol{PA}}_0 = \emptyset$, $k = 0$.
while $k < |\boldsymbol{V}|$ do
    Let subset $\hat{\boldsymbol{V}}_k$ be defined as
    $\{v_i \in \boldsymbol{V} \backslash \hat{\boldsymbol{PA}}_k : FindExpSet(\hat{\boldsymbol{PA}}_k \cup v_i, G) \neq \perp\}$.
    Let $v_{k+1} = \arg\max_{v_i \in \hat{\boldsymbol{V}}_k} J(\hat{\boldsymbol{PA}}_k \cup \{v_i\})$.
    Let $\hat{\boldsymbol{PA}}_{k+1} = \hat{\boldsymbol{PA}}_k \cup v_{k+1}$; $k = k + 1$.
    Continue with the conditional exclusion of [20, Step 2-3] and update the counter $k$.
end while
return $\boldsymbol{\mathcal{F}} = \{\forall f : \hat{\boldsymbol{PA}}_k \to \hat{Y}\}$.
```

Theorem 4. *For a ctf-explainable classifier $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$, $\hat{ER}^d_{x_0,x_1}(\hat{y}|x, y)$ (Eq. 8) is a consistent estimator for $ER^d_{x_0,x_1}(\hat{y}|x, y)$ (Eq. 5) if the model for $P(x|\boldsymbol{c}, y)$ is correctly specified.*

We provide IPW estimators for counterfactual indirect and spurious ERs in Appendix A.

4.1 Finding Adjustment Set for Explainable Classifiers

A few natural questions arise here is (1) how to systematically test whether a classifier f is ctf-explainable, and (2) if so, to find a set $\boldsymbol{C}$ satisfying the explanation criterion so that the counterfactual ERs could be identified. In this section, we will develop an efficient method to answer these questions.

Given a DAG G, by $G^{pbd}_{\{X,Y\}, \hat{\boldsymbol{W}} \backslash X}$ we denote the proper backdoor graph obtained from G by removing the first edge of every proper causal path from $\{X, Y\}$ to $\hat{\boldsymbol{W}} \backslash X$ [23]. We formulate next in graphical terms a set of identification conditions equivalent to the explanation criterion defined in Def. 4.

Definition 5 (Constructive Explanation Criterion)**.** Given a DAG G and a classifier $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$, covariates $\boldsymbol{C}$ satisfy the *constructive explanation criterion* relative to f if and only if (1) $\hat{\boldsymbol{Z}} \subseteq \boldsymbol{C} \subseteq \boldsymbol{V} \backslash Forb(\{X, Y\}, \hat{\boldsymbol{W}} \backslash X)$, where $Forb(\{X, Y\}, \hat{\boldsymbol{W}} \backslash X)$ is a set of nodes forbidden by Def. 4; (2) $\boldsymbol{C}$ d-separates $\{X, Y\}$ and $\hat{\boldsymbol{W}} \backslash X$ in the proper backdoor graph $G^{pbd}_{\{X,Y\}, \hat{\boldsymbol{W}} \backslash X}$.

Theorem 5. *Given a causal diagram G and a classifier f, covariates $\boldsymbol{C}$ satisfies the explanation criterion (Def. 4) to f if and only if it satisfies the constructive explanation criterion (Def. 5) to f.*

Thm. 5 allows us to use the algorithmic framework developed by [23] for constructing d-separating sets in DAGs. We summarize this procedure as *FindExpSet*, in Alg. 1. Specifically, the sub-routine *FindSep* find a covariates set $\boldsymbol{C}$ with $\hat{\boldsymbol{Z}} \subseteq \boldsymbol{C} \subseteq \boldsymbol{V} \backslash Forb(\{X, Y\}, \hat{\boldsymbol{W}} \backslash X)$, such that $\boldsymbol{C}$ d-separates all paths between $\{X, Y\}$ and $\hat{\boldsymbol{W}} \backslash X$ in $G^{pbd}_{\{X,Y\}, \hat{\boldsymbol{W}} \backslash X}$, i.e., the explaining set relative to classifier $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$ (Def. 4). This algorithm can be solved in $\mathcal{O}(n + m)$ runtime where n is the number of nodes and m is the number of edges in the proper backdoor graph $G^{pbd}_{\{X,Y\}, \hat{\boldsymbol{W}} \backslash X}$.

5 Achieving Equalized Counterfactual Error Rates

So far we have focused on analyzing the unequalized counterfactual ERs of an existing predictor in the environment. A more interesting problem is how to obtain an optimal classifier such that its induced counterfactual ERs along with a specific discriminatory mechanism are equalized.

Given finite samples $\boldsymbol{\mathcal{D}} = \{Y_i, \boldsymbol{V}_i\}_{i=1}^n$ drawn from $P(y, \boldsymbol{v})$ (where the protected attribute $X \in \boldsymbol{V}$), the associated causal diagram G, and a set of candidate ctf-explainable classifiers $\boldsymbol{\mathcal{F}}$, the goal of the supervised learning is to obtain an optimal classifier $f^*(\hat{\boldsymbol{pa}})$ from $\boldsymbol{\mathcal{F}}$ such that a loss function $L(\boldsymbol{\mathcal{D}}, f)$ measuring the distance between the prediction $\hat{Y}$ and the true outcome Y is minimized. We will elaborate later about how to construct the ctf-explainable set $\boldsymbol{\mathcal{F}}$. Among the quantities evolved by Thm. 3, the counterfactual distribution $P(\hat{y}_{x, \hat{\boldsymbol{w}} \backslash x, \hat{\boldsymbol{z}}})$ is defined from the classifier f and the other conditional distributions (e.g., $P(\boldsymbol{c}|x, y)$) are estimable from the data $\boldsymbol{\mathcal{D}}$. We could thus represent a counterfactual ER (e.g., direct) of a classifier $f \in \boldsymbol{\mathcal{F}}$ as a function $g(\boldsymbol{\mathcal{D}}, f)$ (e.g., Eq. 8). A fair classifier is obtained by minimizing $L(\boldsymbol{\mathcal{D}}, f)$ subject to a box constraint over $g(\boldsymbol{\mathcal{D}}, f)$, namely,

$$\min_{f \in \boldsymbol{\mathcal{F}}} L(\boldsymbol{\mathcal{D}}, f) \;\text{ s.t. } |g(\boldsymbol{\mathcal{D}}, f)| \leq \epsilon, \tag{9}$$

where $\epsilon \in \mathbb{R}^+$ and the smaller ϵ is, the fairer the learned classifier would be. In general, the constraints $|g(\boldsymbol{\mathcal{D}}, f)| \leq \epsilon$ are non-convex and solving the problem of Eq. 9 seems to be difficult. However, this optimization problem could be significantly simpler in certain cases, solvable using standard convex optimization methods [3]. We provide two canonical settings that fit this requirement.

First, we assume that the features $\boldsymbol{V}$ are discrete, and let $\theta_{\hat{y},x,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}}$ denote the probabilities $P(\hat{y}_{x,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}})$. The counterfactual constraints $|g(\boldsymbol{\mathcal{D}}, f)| \leq \epsilon$ are thus reducible to a set of linear inequalities on the parameter space $\{\theta\}$. Second, consider a classifier making decision based on a decision boundary $\tilde{Y} = \theta^\intercal \phi(x, \hat{\boldsymbol{w}}\backslash x, \hat{\boldsymbol{z}})$ (e.g., logistic regression), where $\phi(\cdot)$ is the basis function. The boundary $\tilde{Y}$ acts as a proxy to the prediction $\hat{Y}$. For instance, the condition $\textit{ER}^d_{x_0,x_1}(\tilde{y}|x, y) = 0$ implies $\textit{ER}^d_{x_0,x_1}(\hat{y}|x, y) = 0$. The same reasoning applies to the counterfactual indirect and spurious ERs. We will employ the techniques in [26] and approximate the constraints $|g(\boldsymbol{\mathcal{D}}, f)| \leq \epsilon$ using the counterfactual ERs of X on the boundary $\tilde{Y}$. Assume that we are interested in the mean effect and replace the quantities $P(\hat{y}_{x,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}})$ in Thm. 3 with $\theta^\intercal \phi(x, \hat{\boldsymbol{w}}\backslash x, \hat{\boldsymbol{z}})$. Given the convexity of $L(\boldsymbol{\mathcal{D}}, f)$, Eq. 9 is a convex optimization problem and can thus be efficiently solved using standard methods.

5.1 Constructing Counterfactually Explainable Classifiers

The counterfactual explainability (Def. 4) of a classifier f relies on its input feature $\hat{\boldsymbol{PA}}$: the smaller the set $\hat{\boldsymbol{PA}}$ is, the easier it would be to find a explaining set $\boldsymbol{C}$ relative to $f(\hat{\boldsymbol{pa}})$. In practice, some features contain critical information about the prediction task, which means that their exclusion could lead to poorer performance. This observation suggests a novel feature selection problem in the fairness-aware classification task: we would like to find a subset $\hat{\boldsymbol{PA}}$ from the available features $\boldsymbol{V}$ such that each classifier in the candidate set $\boldsymbol{\mathcal{F}} = \{\forall f : \hat{\boldsymbol{PA}} \rightarrow \hat{Y}\}$ is ctf-explainable, without significant loss of prediction accuracy.

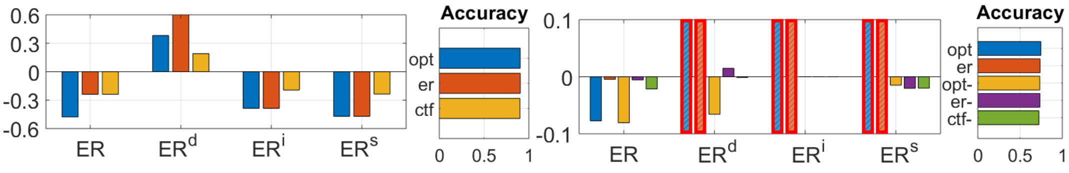
Our solution builds on the procedure *FindExpSet* (Alg. 1) and the classic method of Sequential Floating Forward Selection (*SFFS*) [20]. Let $\hat{\boldsymbol{PA}}_k$ be the set of k features. The score function $J(\hat{\boldsymbol{pa}}_k)$ evaluates the candidate subset $\hat{\boldsymbol{PA}}_k$ and returns a measure of its "goodness". In practice, this score could be obtained by computing the statistical measures of dependence, or by evaluating the best in-class predictive accuracy for classifiers in $\{\forall f : \hat{\boldsymbol{PA}}_k \rightarrow \hat{Y}\}$ on the validation data. We denote our method by Causal *SFFS* (*C-SFFS*) and summarize it in Alg. 2. Starting with a subset $\hat{\boldsymbol{PA}}_k$, *C-SFFS* (Step 2-3) adds one feature which gives the highest score J. *FindExpSet* ensures that the resulting subset $\hat{\boldsymbol{PA}}_{k+1}$ induces a ctf-explainable classifier $f(\hat{\boldsymbol{pa}}_{k+1})$. Step 5 repeatedly removes the least significant feature v_d from the newly-formed $\hat{\boldsymbol{PA}}_k$ until no feature could be excluded to improve the score J. During the exclusion phase, we do not apply *FindExpSet*, since removing features from a ctf-explainable classifier does not violate the explanation criterion (Def. 4). It follows immediately from the soundness of *FindExpSet* that *C-SFFS* always returns a ctf-explainable set $\boldsymbol{\mathcal{F}}$.

Theorem 6. *For $\boldsymbol{\mathcal{F}} =$ C-SFFS$(\boldsymbol{\mathcal{D}}, G)$, each classifier $f \in \boldsymbol{\mathcal{F}}$ is ctf-explainable.*

We summarize in Alg. 3 the procedure of training an optimal classifier satisfying the fairness constraints over the counterfactual ERs. $\textit{ER}^d$, $\textit{ER}^i$, and $\textit{ER}^s$ stand for the counterfactual quantities $\textit{ER}^d_{x_0,x_1}(\hat{y}|x_0, y)$, $\textit{ER}^i_{x_1,x_0}(\hat{y}|x_0, y)$, and $\textit{ER}^s_{x_1,x_0}(\hat{y}|y)$, respectively. We use *C-SFFS* (Alg. 2) to obtain a candidate set $\boldsymbol{\mathcal{F}}$ such that each $f \in \boldsymbol{\mathcal{F}}$ is ctf-explainable. The fair classifier is computed by solving the optimization problem in Eq. 9 subject to the box constraints over $\textit{ER}^d$, $\textit{ER}^i$, and $\textit{ER}^s$.

6 Simulations and Experiments

In this section, we will illustrate our approach on both synthetic and real datasets. We focus on the *false positive rate* $\textit{ER}_{x_0,x_1}(\hat{y}_1|y_0)$ across demographics $x_0 = 0, x_1 = 1$, where $\hat{y}_1 = 1, y_0 = 0$, and the corresponding components $\textit{ER}^d_{x_0,x_1}(\hat{y}_1|x_0, y_0)$, $\textit{ER}^i_{x_1,x_0}(\hat{y}_1|x_0, y_0)$ and $\textit{ER}^s_{x_1,x_0}(\hat{y}_1|y_0)$ (following Thm. 2). We shorten the notation and write $\textit{ER}_{x_0,x_1}(\hat{y}_1|y_0) = \textit{ER}$, and similarly to $\textit{ER}^d$, $\textit{ER}^i$ and $\textit{ER}^s$. Details of the experiments are provided in Appendix C.



(a) Standard Prediction Model

(b) COMPAS

Figure 7: Results of Experiments 1-2. Measures that are not estimable via the explanation criterion are shaded and highlighted. *ER* stands for the false positive rate $ER_{x_0,x_1}(\hat{y}_1|y_0)$; ER^d, ER^i and ER^s represent the corresponding counterfactual direct, indirect, and spurious ERs (Thm. 2). Classifier f_{opt}, f_{er}, and f_{ctf} in Exp. 1 correspond to, respectively, color blue, orange, and yellow in Fig. (a); f_{opt}, f_{er}, $f_{opt\text{-}}$, $f_{er\text{-}}$, and $f_{ctf\text{-}}$ in Exp. 2 correspond to blue, orange, yellow, purple, and green in Fig. (b).

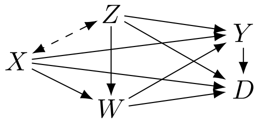


Figure 6: Standard fairness prediction model

Experiment 1: Standard Prediction Model We consider a generalized COMPAS model containing the common descendant D, shown in Fig. 6, which we call here the *standard fairness prediction model* (for short, standard prediction model). We train two classifiers with the same feature set $\{X, W, Z, D\}$ where the first is obtained via the standard, unconstrained optimization, which we call f_{opt}, and the second one constrains the disparate *ER* to half of that of f_{opt}, which we call f_{er}. We further compute the counterfactual ERs (Defs. 1-3). The results are shown in Fig. 7(a). We first confirm that the procedure f_{er} is sound in the sense that f_{eo} (90.4%) achieves a comparable predictive accuracy to f_{opt} (90.4%) while reducing the disparate ER in half ($ER_{er} = -0.238$, $ER_{opt} = -0.476$). Second, ER^d is larger in f_{er} ($ER^d_{eo} = 0.620$) than in the unconstrained f_{opt} ($ER^d_{opt} = 0.381$). This materializes the concern acknowledged in [8], namely, that optimizing based on *ER* may not be enforcing any type of real-life fairness notion related to the underlying causal mechanism. To circumvent this issue, we train a classifier with the same feature set such that its counterfactual ERs are reduced to half of that of the unconstrained f_{opt}, called f_{ctf}. The results (Fig. 7(a)) support the counterfactual approach: f_{ctf} (90.1%) reports *ER* comparable to f_{er} ($ER_{ctf} = -0.238$), but a smaller significant direct, indirect and spurious ER disparities ($ER^d_{ctf} = 0.191$, $ER^d_{ctf} = -0.194$, $ER^d_{ctf} = -0.236$).

Experiment 2: COMPAS In the COMPAS model of Fig. 1, we are interested in predicting whether a defendant would recidivate, while avoiding the direct discrimination (the threshold $\epsilon = 0.01$). We compute a classifier f_{er} with a feature set $\{X, Z, W\}$ subject to $|ER_{er}| \leq \epsilon$. We also include an unconstrained classifier f_{opt} as the baseline. The results (Fig. 7(b)) reveal that f_{er} (73.7%) and f_{opt} (74.6%) are comparable in prediction accuracy while f_{er} has much smaller disparate ER ($ER_{er} = -0.005$, $ER_{opt} = -0.077$). Given that the underlying causal model is not fully known, we could only estimate the counterfactual direct ER from passively-collected samples. Since classifiers with feature set $\{X, W, Z\}$ are not ctf-explainable in the COMPAS model (Def.4), ER^d of f_{er} and f_{opt} cannot be identified via Thm. 3. Previous analysis (Experiment 1) implies that ER^d could be significant even when *ER* is small, which suggests one should be wary of the direct discrimination of f_{er} and f_{opt}. To overcome this issue, we remove W from the feature set and obtain $f_{opt\text{-}}$ and $f_{er\text{-}}$ following a similar procedure. We estimate their ER^d via Thm. 3 with covariates $\boldsymbol{C} = \{Z\}$. The results show that the direct discrimination are significant in both $f_{er\text{-}}$ and $f_{opt\text{-}}$ ($ER^d_{eo-} = 0.015$, $ER^d_{opt-} = -0.066$). To remove the direct discrimination, we train a classifier $f_{ctf\text{-}}$ following Alg. 3 with the features $\{X, Z\}$ and $\epsilon_d = \epsilon$. The results support the efficacy of Alg. 3: $f_{ctf\text{-}}$ performs slightly worse in prediction accuracy (72.1%) but ascertains no direct discrimination ($ER^d_{ctf-} = -0.001$).

7 Conclusions

We introduced a new family of counterfactual measures capable of explaining disparities in the misclassification rates (false positive and false negative) across different demographics in terms of the causal mechanisms underlying the specific prediction process. We then developed machinery based on these measures to allow data scientists (1) to diagnose whether a classifier is operating in a discriminatory fashion against specific groups, and (2) to learn a new classifier subject to fairness constraints in terms of fine-grained misclassification rates. In practice, this approach constitutes a formal solution to the notorious lack of interpretability of the equalized odds. We hope the causal machinery put forward here will help data scientists to analyze already deployed systems as well as to construct new classifiers that are fair even when the training data comes from an unfair world.

A Proofs

In this section, we provide proofs for the technical results presented in the main text. Our proofs build on the exclusion and independence restrictions rules of SCMs [17, pp. 232], and three axioms of structural counterfactuals: composition, effectiveness, and reversibility [17, Ch.7.3.1].

A.1 Proofs of Theorems 1-2

To prove Thm. 1, we first introduce the following three lemmas.

Lemma 1. *Given a SCM $\langle M, P(\boldsymbol{u})\rangle$ and a classifier $f(\hat{\boldsymbol{pa}})$, if node X has no direct path to node $\hat{Y}$ in the conditional causal diagram $G_{|Y}$, i.e., $(X \not\to Y)_{G_{|Y}}$, then for any $x_0, x_1, x, \hat{y}, y$, $ER^{d}_{x_0,x_1}(\hat{y}|x,y) = 0$ holds.*

Proof. To prove $ER^{d}_{x_0,x_1}(\hat{y}|x,y) = 0$, it suffices to show that for any $x_0, x_1, x, \hat{y}, y$,

$$P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}}|x,y) = P(\hat{y}_{x_0,y}|x,y). \tag{10}$$

Conditioned on $(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}$, $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}}|x,y)$ can be written as:

$$P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}}|x,y) = \sum_{\hat{\boldsymbol{pa}}\backslash x} P(\hat{y}_{x_1,\hat{\boldsymbol{pa}}\backslash x}|(\hat{\boldsymbol{pa}}\backslash x)_{x_0,y},x,y)P((\hat{\boldsymbol{pa}}\backslash x)_{x_0,y}|x,y). \tag{11}$$

The variable $\hat{Y}_{x_1,y,\hat{\boldsymbol{pa}}\backslash x} = \hat{Y}_{x_1,\hat{\boldsymbol{pa}}\backslash x}$ since Y is not an input feature to the classifier $f(\hat{\boldsymbol{pa}})$. Bacause X has no direct path to $\hat{Y}$, X is not a part of the input features $\hat{\boldsymbol{PA}}$. We could conclude that the subset $\hat{\boldsymbol{PA}}\backslash X = \hat{\boldsymbol{PA}}$, and for any $x', x, \hat{\boldsymbol{pa}}$,

$$\hat{Y}_{x',\hat{\boldsymbol{pa}}\backslash x} = \hat{Y}_{\hat{\boldsymbol{pa}}}.$$

We could further write Eq. 11 as:

$$\begin{aligned}P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}}|x,y) &= \sum_{\hat{\boldsymbol{pa}}\backslash x} P(\hat{y}_{\hat{\boldsymbol{pa}}}|(\hat{\boldsymbol{pa}}\backslash x)_{x_0,y},x,y)P((\hat{\boldsymbol{pa}}\backslash x)_{x_0,y}|x,y)\\ &= \sum_{\hat{\boldsymbol{pa}}\backslash x} P(\hat{y}_{x_0,y,\hat{\boldsymbol{pa}}\backslash x}|(\hat{\boldsymbol{pa}}\backslash x)_{x_0,y},x,y)P((\hat{\boldsymbol{pa}}\backslash x)_{x_0,y}|x,y).\end{aligned}$$

By the composition axiom, $(\hat{\boldsymbol{PA}}\backslash x)_{x_0,y} = \hat{\boldsymbol{pa}}\backslash x$ implies that $\hat{Y}_{x_0,y,\hat{\boldsymbol{pa}}\backslash x} = \hat{Y}_{x_0,y}$. This gives

$$\begin{aligned}P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}}|x,y) &= \sum_{\hat{\boldsymbol{pa}}\backslash x} P(\hat{y}_{x_0,y}|(\hat{\boldsymbol{pa}}\backslash x)_{x_0,y},x,y)P((\hat{\boldsymbol{pa}}\backslash x)_{x_0,y}|x,y)\\ &= P(\hat{y}_{x_0,y}|x,y).\end{aligned}$$

□

Lemma 2. *Given a SCM $\langle M, P(\boldsymbol{u})\rangle$ and a classifier $f(\hat{\boldsymbol{pa}})$, if there exists no indirect path from node X to $\hat{Y}$ in the conditional causal diagram $G_{|Y}$, i.e., $|X \overset{i}{\to} Y|_{G_{|Y}} = 0$, then for any $x_0, x_1, x, \hat{y}, y$, $ER^{i}_{x_0,x_1}(\hat{y}|x,y) = 0$ holds.*

Proof. Without loss of generality, we suppose $|\hat{\boldsymbol{PA}}| > 0$. To prove $ER^{i}_{x_0,x_1}(\hat{y}|x,y) = 0$, it suffices to show that for any $x_0, x_1, x, y, \boldsymbol{u}$,

$$\hat{Y}_{x_0,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_1,y}(\boldsymbol{u})}(\boldsymbol{u}) = \hat{Y}_{x_0,y}(\boldsymbol{u}). \tag{12}$$

We will first show that if $|X \overset{i}{\to} Y|_{G_{|Y}} = 0$, then for any $x_0, x_1, y, \hat{\boldsymbol{pa}}, \boldsymbol{u}$, one of the following equation must hold

$$\hat{Y}_{x_0,y,\hat{\boldsymbol{pa}}\backslash x}(\boldsymbol{u}) = \hat{Y}_{x_0,y}(\boldsymbol{u}), \tag{13}$$

$$(\hat{\boldsymbol{PA}}\backslash X)_{x_1,y}(u) = (\hat{\boldsymbol{PA}}\backslash X)_{y}(u). \tag{14}$$

Suppose that Eqs. 13 and 14 both fail, there must exist a unblocked causal path l_1 from X to $\hat{\boldsymbol{PA}}\backslash X$ and a unblocked causal path l_2 from $\hat{\boldsymbol{PA}}\backslash X$ to $\hat{Y}$ in the conditional causal diagram $G_{|Y}$ [6, Lem. 12]. We can then find an indirect path from X to $\hat{Y}$ in $G_{|Y}$ by concatenating l_1 and l_2, which is a contradiction to the assumption that $|X \overset{i}{\to} Y|_{G_{|Y}} = 0$. It is verifiable that Eq. 13 implies Eq. 12. By Eq. 14, we have:

$$\begin{aligned}\hat{Y}_{x_0,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_1,y}(\boldsymbol{u})}(\boldsymbol{u}) &= \hat{Y}_{x_0,y,(\hat{\boldsymbol{PA}}\backslash X)_{y}(\boldsymbol{u})}(\boldsymbol{u}) \\ &= \hat{Y}_{x_0,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}(\boldsymbol{u})}(\boldsymbol{u}).\end{aligned}$$

By the composition axiom, for any $\hat{\boldsymbol{pa}}\backslash x$, $(\hat{\boldsymbol{PA}}\backslash X)_y(\boldsymbol{u}) = \hat{\boldsymbol{pa}}\backslash x$ implies that $\hat{Y}_{x_0,y,\hat{\boldsymbol{pa}}\backslash x}(\boldsymbol{u}) = \hat{Y}_{x_0,y,}(\boldsymbol{u})$. Together with the above equation, we have:

$$\hat{Y}_{x_0,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_1,y}(\boldsymbol{u})}(\boldsymbol{u}) = \hat{Y}_{x_0,y}(\boldsymbol{u}). \qquad \square$$

Lemma 3. *Given a SCM $\langle M, P(\boldsymbol{u})\rangle$ and a classifier $f(\hat{\boldsymbol{pa}})$, if there exists no spurious path from node X to $\hat{Y}$ in the conditional causal diagram $G_{|Y}$, i.e., $|X \overset{s}{\leftrightarrow} Y|_{G_{|Y}} = 0$, then for any $x_0, x_1, \hat{y}, y$, $ER^s_{x_0,x_1}(\hat{y}|y) = 0$ holds.*

Proof. It suffices to prove that variables $\hat{Y}_{x_0,y}$ and X are independent given Y, i.e., for any $x_0, x, y, \hat{y}$,

$$P(\hat{y}_{x_0,y}, x|y) = P(\hat{y}_{x_0,y}|y)P(x|y). \tag{15}$$

Let $X', \hat{Y}'$ denote the protected attribute and the prediction in the conditional causal model $\langle M_y, P(\boldsymbol{u}|y)\rangle$, and let $P'(\cdot)$ denote the distributions induced by $\langle M_y, P(\boldsymbol{u}|y)\rangle$. By the backdoor criterion [17, Ch. 11.3.2], if there exists no spurious path between nodes X' and $\hat{Y}'$, then the factual X' and the counterfactual $\hat{Y}'_x$ are independent. We thus have:

$$P'(\hat{Y}'_{x_0} = \hat{y}, X = x') = P'(\hat{Y}'_{x_0} = \hat{y})P'(X = x').$$

Let $I_{\{\cdot\}}$ denote the indicator function. Expanding on $\boldsymbol{u}$ writes the above equation as:

$$\sum_{\boldsymbol{u}} I_{\{\hat{Y}'_{x_0}(\boldsymbol{u})=\hat{y}, X'(\boldsymbol{u})=x\}} P(\boldsymbol{u}|y) = \sum_{\boldsymbol{u}} I_{\{\hat{Y}'_{x_0}(\boldsymbol{u})=\hat{y}\}} P(\boldsymbol{u}|y) \cdot \sum_{\boldsymbol{u}} I_{\{X'(\boldsymbol{u})=x\}} P(\boldsymbol{u}|y) \tag{16}$$

By definition, given $\boldsymbol{U} = \boldsymbol{u}$, the solutions of endogenous variables in the conditional causal model $\langle M_y, P(\boldsymbol{u}|y)\rangle$ coincides with the potential response $V_y(u)$ in the original causal model $\langle M, P(u)\rangle$. Eq. 16 can thus be written as:

$$\sum_{\boldsymbol{u}} I_{\{\hat{Y}_{x_0,y}(\boldsymbol{u})=\hat{y}, X_y(\boldsymbol{u})=x\}} P(\boldsymbol{u}|y) = \sum_{\boldsymbol{u}} I_{\{\hat{Y}_{x_0,y}(\boldsymbol{u})=\hat{y}\}} P(\boldsymbol{u}|y) \cdot \sum_{\boldsymbol{u}} I_{\{X_y(\boldsymbol{u})=x\}} P(\boldsymbol{u}|y), \tag{17}$$

By definition, the counterfactual distribution $P(\hat{y}_{x_0,y}, x|y)$ is equal to:

$$P(\hat{y}_{x_0,y}, x|y) = \sum_{\boldsymbol{u}} I_{\{\hat{Y}_{x_0,y}(\boldsymbol{u})=\hat{y}, X_y(\boldsymbol{u})=x\}} P(\boldsymbol{u}|y). \tag{18}$$

Eqs. 17 and 18 combined give

$$P(\hat{y}_{x_0,y}, x|y) = P(\hat{y}_{x_0,y}|y)P(x|y). \qquad \square$$

Thm. 1 follows immediately from Lems. 1-3.

Proof of Theorem 1. By Lem. 1, we have $(X \not\to Y)_{G_{|Y}} \Rightarrow ER^d_{x_0,x_1}(\hat{y}|x,y) = 0$. Similarly, Properties (2-3) are proved, respectively, by applying Lems. 2-3. □

We next provide the generalized form of the casual explanation formula of the equalized odds (Thm. 2), including more decompositions of the disparate ER.

Theorem 7 (Generalized Causal Explanation Formula of Equalized Odds)**.** *For any $x_0, x_1, \hat{y}, y$, $ER_{x_0,x_1}(\hat{y}|x, y)$, $ER^d_{x_0,x_1}(\hat{y}|x, y)$, $ER^i_{x_0,x_1}(\hat{y}|x, y)$ and $ER^s_{x_0,x_1}(\hat{y}|y)$ obey the following non-parametric relationship:*

$$\begin{aligned} ER_{x_0,x_1}(\hat{y}|y) &= ER^d_{x_0,x_1}(\hat{y}|x_0, y) - ER^i_{x_1,x_0}(\hat{y}|x_0, y) - ER^s_{x_1,x_0}(\hat{y}|y), \\ &= ER^i_{x_0,x_1}(\hat{y}|x_0, y) - ER^d_{x_1,x_0}(\hat{y}|x_0, y) - ER^s_{x_1,x_0}(\hat{y}|y), \\ &= ER^s_{x_0,x_1}(\hat{y}|y) + ER^d_{x_0,x_1}(\hat{y}|x_1, y) - ER^i_{x_1,x_0}(\hat{y}|x_1, y), \\ &= ER^s_{x_0,x_1}(\hat{y}|y) + ER^i_{x_0,x_1}(\hat{y}|x_1, y) - ER^d_{x_1,x_0}(\hat{y}|x_1, y). \end{aligned} \tag{19}$$

Thm. 2 is implied by the first decomposition of the above equations. To prove Thm. 7, we first introduce the effect of treatment on the treated (ETT) [17, Ch. 8.2.5] of treatment X on $\hat{Y} = \hat{y}$ contingent on the additional evidence $Y = y$, defined as:

$$ETT_{x_0,x_1}(\hat{y}|y) = P(\hat{y}_{x_1,y}|x_0, y) - P(\hat{y}_{x_0,y}|x_0, y). \tag{20}$$

The following two lemmas characterizes the quantitative relationships between $ETT_{x_0,x_1}(\hat{y}|y)$ and the counterfactual ERs.

Lemma 4. *For any $x_0, x_1, \hat{y}, y$, $ER_{x_0,x_1}(\hat{y}|x, y)$, $ETT_{x_0,x_1}(\hat{y}|y)$ and $ER^s_{x_0,x_1}(\hat{y}|y)$ obey the following non-parametric relationship:*

$$ER_{x_0,x_1}(\hat{y}|y) = ETT_{x_0,x_1}(\hat{y}|y) - ER^s_{x_1,x_0}(\hat{y}|y), \tag{21}$$

$$= ER^s_{x_0,x_1}(\hat{y}|y) - ETT_{x_1,x_0}(\hat{y}|y). \tag{22}$$

Proof. By a simple application of telescoping sum, $ER_{x_0,x_1}(\hat{y}|y)$ can be written as:

$$\begin{aligned} ER_{x_0,x_1}(\hat{y}|y) &= P(\hat{y}|x_1, y) - P(\hat{y}|x_0, y) \\ &= P(\hat{y}|x_1, y) - P(\hat{y}_{x_1,y}|x_0, y) + P(\hat{y}_{x_1,y}|x_0, y) - P(\hat{y}|x_0, y). \end{aligned}$$

By the composition axiom, for any x, y, $X = x, Y = y$ implies that $\hat{Y}_{x,y} = \hat{Y}$. The above equation could thus be written as:

$$\begin{aligned} ER_{x_0,x_1}(\hat{y}|y) &= P(\hat{y}_{x_1,y}|x_1, y) - P(\hat{y}_{x_1,y}|x_0, y) + P(\hat{y}_{x_1,y}|x_0, y) - P(\hat{y}_{x_0,y}|x_0, y) \\ &= ETT_{x_0,x_1}(\hat{y}|y) - ER^s_{x_1,x_0}(\hat{y}|y). \end{aligned}$$

By replacing the decomposing term $P(\hat{y}_{x_1,y}|x_0, y)$ with $P(\hat{y}_{x_0,y}|x_1, y)$ in the above equation, we prove Eq. 22. □

Lemma 5. *For any $x_0, x_1, \hat{y}, y$, $ETT_{x_0,x_1}(\hat{y}|y)$, $ER^d_{x_0,x_1}(\hat{y}|x, y)$ and $ER^i_{x_0,x_1}(\hat{y}|x, y)$ obey the following non-parametric relationship:*

$$ETT_{x_0,x_1}(\hat{y}|y) = ER^d_{x_0,x_1}(\hat{y}|x_0, y) - ER^i_{x_1,x_0}(\hat{y}|x_0, y), \tag{23}$$

$$= ER^i_{x_0,x_1}(\hat{y}|x_0, y) - ER^d_{x_1,x_0}(\hat{y}|x_0, y). \tag{24}$$

Proof. By a simple application of telescoping sum, $ETT_{x_0,x_1}(\hat{y}|y)$ can be written as:

$$\begin{aligned} ETT_{x_0,x_1}(\hat{y}|y) &= P(\hat{y}_{x_1,y}|x_0, y) - P(\hat{y}|x_0, y) \\ &= P(\hat{y}_{x_1,y}|x_0, y) - P(\hat{y}_{x_0,y}|x_0, y) \\ &= P(\hat{y}_{x_1,y}|x_0, y) - P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\setminus X)_{x_0,y}}|x_0, y) \\ &\quad + P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\setminus X)_{x_0,y}}|x_0, y) - P(\hat{y}_{x_0,y}|x_0, y) \\ &= ER^d_{x_0,x_1}(\hat{y}|x_0, e) - ER^i_{x_1,x_0}(\hat{y}|x_0, e). \end{aligned}$$

By replacing the decomposing term $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\setminus X)_{x_0,y}}|x_0, y)$ with $P(\hat{y}_{x_0,y,(\hat{\boldsymbol{PA}}\setminus X)_{x_1,y}}|x_0, y)$ in the above equation, we prove Eq. 24. □

We are now ready to derive the generalized causal explanation formula for the disparate ER.

Proof of Theorem. 7. Lems. 4-5 combined give Eq. 19. □

A.2 Proofs of Theorems 3-4

We first introduce the following lemma providing the identification formula for the nested counterfactual quantity $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{PA}}\backslash X)_{x_0,y}}|x,y)$ induced by a ctf-explainable classifier.

Lemma 6. *Given a causal diagram G and a classifier $f(\hat{\boldsymbol{w}},\hat{\boldsymbol{z}})$, if f is ctf-explainable (Def. 4) with an explaining set $\boldsymbol{C}$, the counterfactual distribution $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$ can be estimated as follows:*

$$P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)=\sum_{\hat{\boldsymbol{w}},\boldsymbol{c}}P(\hat{y}_{x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}})P(\hat{\boldsymbol{w}}\backslash x|x_0,\boldsymbol{c},y)P(\boldsymbol{c}|x,y), \tag{25}$$

where $P(\hat{y}_{x,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}})$ is well-defined, computable from the classifier $f(\hat{\boldsymbol{w}},\hat{\boldsymbol{z}})$.

Proof. Values of the prediction $\hat{Y}$ is decided by the classifier $f(\hat{\boldsymbol{w}},\hat{\boldsymbol{z}})$. Since $\hat{\boldsymbol{Z}}$ are non-descendant nodes of both X and Y, variable $\hat{\boldsymbol{Z}}_{x_0,y}=\hat{\boldsymbol{Z}}$. $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$ could be simplified as:

$$P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)=P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X)_{x_0,y},\hat{\boldsymbol{Z}}}|x,y)$$

By definition (Def. 4), $\hat{\boldsymbol{Z}}\subseteq\boldsymbol{C}$. Expanding on the features $\hat{\boldsymbol{W}}$ and the explaining set $\boldsymbol{C}$ gives:

$$\begin{aligned}P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)&=\sum_{\hat{\boldsymbol{w}},\boldsymbol{c}}I_{\{f(x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}})=\hat{y}\}}P((\hat{\boldsymbol{w}}\backslash x)_{x_0,y}|x,y,\boldsymbol{c})P(\boldsymbol{c}|x,y)\\&=\sum_{\hat{\boldsymbol{w}},\boldsymbol{c}}P(\hat{y}_{x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}})P((\hat{\boldsymbol{w}}\backslash x)_{x_0,y}|x,y,\boldsymbol{c})P(\boldsymbol{c}|x,y).\end{aligned} \tag{26}$$

The last step holds since by definition, $P(\hat{y}_{x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}})=I_{\{f(x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}})=\hat{y}\}}$. By results in [21], the adjustment criterion (Conditions (2-3) of Def. 4) holds for the covariates set $\boldsymbol{C}$ relative to $(\{X,Y\},\hat{\boldsymbol{W}}\backslash X)$ in a causal diagram if and only if for any x,y, the counterfactual variable $(\hat{\boldsymbol{W}}\backslash X)_{x,y}$ is independent of variables X,Y given $\boldsymbol{C}$, i.e., $((\hat{\boldsymbol{W}}\backslash X)_{x,y}\perp\!\!\!\perp X,Y|Z)$. We could thus write the distribution $P((\hat{\boldsymbol{w}}\backslash x)_{x_0,y}|x,y,\boldsymbol{c})$ as:

$$\begin{aligned}P((\hat{\boldsymbol{w}}\backslash x)_{x_0,y}|x,y,\boldsymbol{c})=P((\hat{\boldsymbol{w}}\backslash x)_{x_0,y}|\boldsymbol{c})&=P((\hat{\boldsymbol{w}}\backslash x)_{x_0,y}|x_0,y,\boldsymbol{c})\\&=P((\hat{\boldsymbol{w}}\backslash x)|x,y,\boldsymbol{c}).\end{aligned} \tag{27}$$

The last step holds by the composition axiom: $X=x_0,Y=y$ implies that $\hat{Y}_{x_0,y}=\hat{Y}$. Replacing $P((\hat{\boldsymbol{w}}\backslash x)_{x_0,y}|x,y,\boldsymbol{c})$ with Eq. 27 in Eq. 26 gives:

$$P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X)_{x_0,y},\hat{\boldsymbol{Z}}}|x,y)=\sum_{\hat{\boldsymbol{w}},\boldsymbol{c}}P(\hat{y}_{x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}})P(\hat{\boldsymbol{w}}\backslash x|x_0,\boldsymbol{c},y)P(\boldsymbol{c}|x,y).$$

The above proof could be easily generalized for an stochastic classifier $f(\hat{\boldsymbol{w}},\hat{\boldsymbol{z}},\epsilon)$ where ϵ is an independent noise associated only with the predictor $\hat{Y}$. □

Proof of Theorem 3. The definitions of $\mathit{ER}^d_{x_0,x_1}(\hat{y}|x,y)$, $\mathit{ER}^i_{x_0,x_1}(\hat{y}|x,y)$ and $\mathit{ER}^s_{x_0,x_1}(\hat{y}|y)$ involve the counterfactual distributions of $P(\hat{y}_{x_1,y}|x_0,y)$ and $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$. We will first show that $P(\hat{y}_{x_1,y}|x_0,y)$ could be written as:

$$P(\hat{y}_{x_1,y}|x_0,y)=P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_1,y}}|x_0,y). \tag{28}$$

By expanding on $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{Z}}$, we have:

$$P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_1,y}}|x_0,y)=\sum_{\hat{\boldsymbol{w}},\hat{\boldsymbol{z}}}P(\hat{y}_{x_1,y,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}}|(\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}})_{x_1,y},x_0,y)P((\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}})_{x_1,y}|x,y).$$

By the composition axiom, $(\hat{\boldsymbol{W}}\backslash x)_{x_1,y}=\hat{\boldsymbol{w}}\backslash x$, $\hat{\boldsymbol{Z}}_{x_1,y}=\hat{\boldsymbol{z}}$ implies that $\hat{Y}_{x_1,y,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}}=\hat{Y}_{x_1,y}$. This gives:

$$\begin{aligned}P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_1,y}}|x_0,y)&=\sum_{\hat{\boldsymbol{w}},\hat{\boldsymbol{z}}}P(\hat{y}_{x_1,y}|(\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}})_{x_1,y},x_0,y)P((\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}})_{x_1,y}|x,y)\\&=P(\hat{y}_{x_1,y}|x_0,y),\end{aligned}$$

which proves Eq. 28. For any $x_0, x_1, x, \hat{y}, y$, the identification formula for the counterfactual distribution of the form $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$ is provided in Lem. 6. Applying Lem. 6, respectively, to $\boldsymbol{ER}^d_{x_0,x_1}(\hat{y}|x,y)$, $\boldsymbol{ER}^i_{x_0,x_1}(\hat{y}|x,y)$ and $\boldsymbol{ER}^s_{x_0,x_1}(\hat{y}|y)$ completes the proof. □

We next provide the IPW estimators for the counterfactual direct, indirect and spurious ERs (Defs. 1-3) induced by a ctf-explainable classifier $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$. Given the observed data $\boldsymbol{\mathcal{D}} = \{Y_i, \hat{\boldsymbol{W}}_i, \boldsymbol{C}_i\}_{i=1}^n$, we propose the IPW estimators for $\boldsymbol{ER}^d_{x_0,x_1}(\hat{y}|x,y)$, $\boldsymbol{ER}^i_{x_0,x_1}(\hat{y}|x,y)$ and $\boldsymbol{ER}^s_{x_0,x_1}(\hat{y}|y)$ as follows:

$$\begin{aligned}
\hat{\boldsymbol{ER}}^d_{x_0,x_1}(\hat{y}|x,y) \quad &= \frac{1}{n}\sum_{i=1}^n (P(\hat{y}_{x_1,\hat{\boldsymbol{W}}_i\backslash X_i,\hat{\boldsymbol{Z}}_i}) - P(\hat{y}_{x_0,\hat{\boldsymbol{W}}_i\backslash X_i,\hat{\boldsymbol{Z}}_i}))\frac{\hat{P}(x|\boldsymbol{C}_i,y)I_{\{X_i=x_0,Y_i=y\}}}{\hat{P}(x_0|\boldsymbol{C}_i,y)\hat{P}(x,y)},\\
\hat{\boldsymbol{ER}}^i_{x_0,x_1}(\hat{y}|x,y) \quad &= \frac{1}{n}\sum_{i=1}^n P(\hat{y}_{x_0,\hat{\boldsymbol{W}}_i\backslash X_i,\hat{\boldsymbol{Z}}_i})\frac{\hat{P}(x|\boldsymbol{C}_i,y)I_{\{X_i=x_1,Y_i=y\}}}{\hat{P}(x_0|\boldsymbol{C}_i,y)\hat{P}(x,y)}\\
&\quad - \frac{1}{n}\sum_{i=1}^n P(\hat{y}_{x_0,\hat{\boldsymbol{W}}_i\backslash X_i,\hat{\boldsymbol{Z}}_i})\frac{\hat{P}(x|\boldsymbol{C}_i,y)I_{\{X_i=x_0,Y_i=y\}}}{\hat{P}(x_1|\boldsymbol{C}_i,y)\hat{P}(x,y)},\\
\hat{\boldsymbol{ER}}^s_{x_0,x_1}(\hat{y}|x,y) \quad &= \frac{1}{n}\sum_{i=1}^n P(\hat{y}_{x_0,\hat{\boldsymbol{W}}_i\backslash X_i,\hat{\boldsymbol{Z}}_i})\Big(\frac{\hat{P}(x_1|\boldsymbol{C}_i,y)}{\hat{P}(x_1,y)} - \frac{\hat{P}(x_0|\boldsymbol{C}_i,y)}{\hat{P}(x_0,y)}\Big)\frac{I_{\{X_i=x_0,Y_i=y\}}}{\hat{P}(x_0|\boldsymbol{C}_i,y)}.
\end{aligned}$$

Among the above equations, $\hat{P}(x,y)$ is the sample mean estimator of $P(x,y)$ (X, Y are finite). $\hat{P}(x|\boldsymbol{c},y)$ is a reliable estimator of the conditional distributions $P(x|\boldsymbol{c},y)$ and, in practice, could be estimated by assuming some parametric models such as logistic regression.

Theorem 8. *For a ctf-explainable classifier $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$, $\hat{\boldsymbol{ER}}^d_{x_0,x_1}(\hat{y}|x,y)$, $\hat{\boldsymbol{ER}}^i_{x_0,x_1}(\hat{y}|x,y)$ and $\hat{\boldsymbol{ER}}^s_{x_0,x_1}(\hat{y}|x,y)$ are consistent estimators, respectively, for $\boldsymbol{ER}^d_{x_0,x_1}(\hat{y}|x,y)$, $\boldsymbol{ER}^i_{x_0,x_1}(\hat{y}|x,y)$ and $\boldsymbol{ER}^s_{x_0,x_1}(\hat{y}|y)$ if the model for $P(x|\boldsymbol{c},y)$ is correctly specified.*

To prove Thm. 8, we first introduce an IPW estimator for the nested counterfactual distribution $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$ induced by a ctf-explainable classifier.

Lemma 7. *Given a causal diagram G, a classifier $f(\hat{\boldsymbol{w}}, \hat{\boldsymbol{z}})$, an explaining set $\boldsymbol{C}$ relative to f and the observed data $\boldsymbol{\mathcal{D}} = \{Y_i, \hat{\boldsymbol{W}}_i, \boldsymbol{C}_i\}_{i=1}^n$, the IPW estimator for the distribution $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$ is defined as:*

$$\hat{P}(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y) = \frac{1}{n}\sum_{i=1}^n P(\hat{y}_{x_1,\hat{\boldsymbol{W}}_i\backslash X_i,\hat{\boldsymbol{Z}}_i})\frac{\hat{P}(x|\boldsymbol{C}_i,y)I_{\{X_i=x_0,Y_i=y\}}}{\hat{P}(x_0|\boldsymbol{C}_i,y)\hat{P}(x,y)}, \tag{29}$$

where $\hat{P}(x,y)$ is the empirical mean of $P(x,y)$. If the model for $P(x|\boldsymbol{c},y)$ is correctly specified, $\hat{P}(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$ is a consistent estimator for $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$.

Proof. By the law of large numbers, it suffices to prove that

$$P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y) = E\Big[P(\hat{y}_{x_1,\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}}})\frac{P(x|\boldsymbol{C},y)}{P(x_0|\boldsymbol{C},y)P(x,y)}I_{\{X=x_0,Y=y\}}\Big]$$

From Lem. 6, $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$ could be written as:

$$P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y) = \sum_{\hat{\boldsymbol{w}},\boldsymbol{c}} P(\hat{y}_{x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}})P(\hat{\boldsymbol{w}}\backslash x|x_0,\boldsymbol{c},y)P(\boldsymbol{c}|x,y).$$

By basic probabilistic operations, we could further write the above equation as:

$$\begin{aligned}
P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y) &= \sum_{\hat{\boldsymbol{w}},\boldsymbol{c}} P(\hat{y}_{x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}}) \frac{P(\hat{\boldsymbol{w}}\backslash x, x_0, \boldsymbol{c}, y)}{P(x_0,\boldsymbol{c},y)} \frac{P(x,\boldsymbol{c},y)}{P(x,y)} \\
&= \sum_{\hat{\boldsymbol{w}},\boldsymbol{c}} P(\hat{y}_{x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}}) \frac{P(\hat{\boldsymbol{w}}\backslash x, x_0, \boldsymbol{c}, y)}{P(x_0|\boldsymbol{c},y)P(\boldsymbol{c},y)} \frac{P(x|\boldsymbol{c},y)P(\boldsymbol{c},y)}{P(x,y)} \\
&= \sum_{\hat{\boldsymbol{w}},\boldsymbol{c}} P(\hat{y}_{x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}}) \frac{P(x|\boldsymbol{c},y)}{P(x_0|\boldsymbol{c},y)P(x,y)} P(\hat{\boldsymbol{w}}\backslash x, x_0, \boldsymbol{c}, y) \\
&= \sum_{\hat{\boldsymbol{w}},\boldsymbol{c},x',y'} P(\hat{y}_{x_1,\hat{\boldsymbol{w}}\backslash x,\hat{\boldsymbol{z}}}) \frac{P(x|\boldsymbol{c},y)}{P(x_0|\boldsymbol{c},y)P(x,y)} P(\hat{\boldsymbol{w}}\backslash x, x', \boldsymbol{c}, y') I_{\{x'=x_0,y'=y\}} \\
&= E\left[P(\hat{y}_{x_1,\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}}}) \frac{P(x|\boldsymbol{C},y)}{P(x_0|\boldsymbol{C},y)P(x,y)} I_{\{X=x_0,Y=y\}} \right] \qquad \square
\end{aligned}$$

We are finally ready to Thm. 8.

Proof of Theorem 8. From Eq. 28, we could obtain an IPW estimator for the distribution $P(\hat{y}_{x_1,y}|x_0,y)$ using Lem. 7. Recall that the definitions of $\mathit{ER}^d_{x_0,x_1}(\hat{y}|x,y)$, $\mathit{ER}^i_{x_0,x_1}(\hat{y}|x,y)$ and $\mathit{ER}^s_{x_0,x_1}(\hat{y}|y)$ involve only the counterfactual distributions $P(\hat{y}_{x_1,y}|x_0,y)$ and $P(\hat{y}_{x_1,y,(\hat{\boldsymbol{W}}\backslash X,\hat{\boldsymbol{Z}})_{x_0,y}}|x,y)$. Applying Lem. 7, respectively, to $\mathit{ER}^d_{x_0,x_1}(\hat{y}|x,y)$, $\mathit{ER}^i_{x_0,x_1}(\hat{y}|x,y)$ and $\mathit{ER}^s_{x_0,x_1}(\hat{y}|y)$ leads to Thm. 8. □

Proof of Theorem 4. Since Thm. 8 generalizes Thm. 4, the proof of Thm. 4 follows immediately. □

A.3 Proofs of Theorems 5-6

To prove Thm. 5, we first introduce the following two lemmas.

Lemma 8 (Def. 5 ⇒ Def. 4)**.** *Given a causal diagram G and a classifier f, if a set of covariates $\boldsymbol{C}$ satisfies the constructive explanation criterion (Def. 5) relative to f, it also satisfies the explanation criterion (Def. 4) to f.*

Proof. Cond. (1) of Def. 5 implies Conds. (1-2) of Def. 4. The proper backdoor graph $G^{pbd}_{\{X,Y\},\hat{\boldsymbol{W}}\backslash X}$ contains only the spurious paths from $\{X, Y\}$ to $\hat{\boldsymbol{W}}\backslash X$. Therefore, if $\boldsymbol{C}$ satisfies Cond. (2) of Def. 5, it will also satisfy Cond. (3) of Def. 4. □

Lemma 9 (Def. 4 ⇒ Def. 5)**.** *Given a causal diagram G and a classifier f, if a set of covariates $\boldsymbol{C}$ satisfies the explanation criterion (Def. 4) relative to f, it also satisfies the constructive explanation criterion (Def. 5) to f.*

Proof. Conds. (1-2) of Def. 4 implies Cond. (1) of Def. 5. By Cond. (3) of Def. 4, the covariates set $\boldsymbol{C}$ blocks all spurious paths from $\{X, Y\}$ to $\hat{\boldsymbol{W}}\backslash X$. This facts implies Cond. (2) of Def. 5. □

Proof of Theorem 5. It follows immediately from Lems. 8-9. □

We are now ready to prove the soundness of Causal-SFFS (Alg. 2).

Proof of Theorem 6. We will prove this theorem by contradiction. Let $\hat{\boldsymbol{PA}}_k$ denote the feature set returned by *C-SFFS*. Suppose any classifier f in $\mathcal{F} = \{\forall f : \hat{\boldsymbol{PA}}_k \to \hat{Y}\}$ is not ctf-explainable. Let $\hat{\boldsymbol{PA}}'_k$ denote the feature set containing $\hat{\boldsymbol{PA}}_k$ ($k' \geq k$) before the conditional exclusion (Step. 5). For a covariates set $\boldsymbol{C}$, if $\boldsymbol{C}$ is an explaining set relative to a classifier f with a feature set $\hat{\boldsymbol{PA}}'_k$, it must also be an explaining set relative to a classifier with the subset $\hat{\boldsymbol{PA}}_k$. We could thus conclude that for any f in $\mathcal{F} = \{\forall f : \hat{\boldsymbol{PA}}'_k \to \hat{Y}\}$ must also not be ctf-explainable. However, Steps 2-3 of *C-SFFS* guarantee that each feature set $\hat{\boldsymbol{PA}}_k^{'}$ before the conditional exclusion phase must induce a set of ctf-explainable classifiers (*FindExpSet*$(\hat{\boldsymbol{PA}}_k^{'}, G) \neq \perp$), which is a contradiction. □

B Connections with Other Counterfactual Measures

In this section, we will examine the relationships between the proposed counterfactual ERs and other frameworks of counterfactual fairness analysis. Specifically, we will compared the counterfactual ERs with the natural direct and indirect effects [18, 12, 16] and the counterfactual fairness condition [11] in the context of the COMPAS model (Fig. 1). By definitions of Defs. 1-3, the counterfactual direct, indirect and spurious ERs of a classifier $f(x, w, z)$ in the COMPAS model of Fig. 1 are written as:

$$\textit{ER}^d_{x_0,x_1}(\hat{y}|x, y) = P(\hat{y}_{x_1,W_{x_0},Z}|x, y) - P(\hat{y}_{x_0}|x, y) \tag{30}$$

$$\textit{ER}^i_{x_0,x_1}(\hat{y}|x, y) = P(\hat{y}_{x_0,W_{x_1},Z}|x, y) - P(\hat{y}_{x_0}|x, y) \tag{31}$$

$$\textit{ER}^s_{x_0,x_1}(\hat{y}|y) = P(\hat{y}_{x_0}|x_1, y) - p(\hat{y}_{x_0}|x_0, y) \tag{32}$$

In the above equations, we could ignore the effect of intervention *do*(y) on $\hat{Y}$ since the true outcome Y does not causally affect the prediction $\hat{Y}$.

B.1 Natural Direct and Indirect Effects

In the COMPAS model (Fig. 1), the natural direct (*NDE*) and indirect (*NIE*) effects [18] of treatment $X = x_1$ on $\hat{Y} = \hat{y}$ (with baseline $X = x_0$) are defined as:

$$\textit{NDE}_{x_0,x_1}(\hat{y}) = P(\hat{y}_{x_1,W_{x_0},Z}) - P(\hat{y}_{x_0}), \tag{33}$$

$$\textit{NIE}_{x_0,x_1}(\hat{y}) = P(\hat{y}_{x_0,W_{x_1},Z}) - P(\hat{y}_{x_0}). \tag{34}$$

We could observe that the counterfactual direct and indirect ERs of Eqs. 30-31 could be seen as the natural direct and indirect effects conditioned on the context $X = x, Y = y$, namely

Theorem 9. *Given the COMPAS model of Fig. 1 and a classifier $f(x, w, z)$, $\textit{ER}^d_{x_0,x_1}(\hat{y}|x, y)$, $\textit{ER}^i_{x_0,x_1}(\hat{y}|x, y)$, $\textit{NDE}_{x_0,x_1}(\hat{y})$ and $\textit{NIE}_{x_0,x_1}(\hat{y})$ obey the following relationships:*

$$\textit{NDE}_{x_0,x_1}(\hat{y}) = \sum_{x,y} \textit{ER}^d_{x_0,x_1}(\hat{y}|x, y)P(x, y), \tag{35}$$

$$\textit{NIE}_{x_0,x_1}(\hat{y}) = \sum_{x,y} \textit{ER}^i_{x_0,x_1}(\hat{y}|x, y)P(x, y). \tag{36}$$

Proof. By conditioning on x, y, we can write $\textit{NDE}_{x_0,x_1}(\hat{y})$ as:

$$\begin{aligned}\textit{NDE}_{x_0,x_1}(\hat{y}) &= \sum_{x,y}(P(\hat{y}_{x_1,W_{x_0},Z}|x, y) - P(\hat{y}_{x_0}|x, y))P(x, y) \\ &= \sum_{x,y} \textit{ER}^d_{x_0,x_1}(\hat{y}|x, y)P(x, y).\end{aligned}$$

Eq. 36 could be similarly proved. □

As a corollary of Thm. 9, it immediately follows that in the COMPAS model, the counterfactual direct and indirect ERs impose stronger constraints over the underlying mechanisms than *NDE* and *NIE*.

Corollary 1. *Given the COMPAS model of Fig. 1 and a classifier $f(x, w, z)$, for any $x_0, x_1, x, \hat{y}, y$, $\textit{ER}^d_{x_0,x_1}(\hat{y}|x, y) = 0 \Rightarrow \textit{NDE}_{x_0,x_1}(\hat{y}) = 0$. Similarly, $\textit{ER}^i_{x_0,x_1}(\hat{y}|x, y) = 0 \Rightarrow \textit{NIE}_{x_0,x_1}(\hat{y}) = 0$.*

Proof. The proof follows immediately from Eqs. 35-36. □

B.2 Counterfactual Fairness

Following the note in [28], the counterfactual fairness measure [11] can be seen as the effect of treatment on the treated (ETT) [17, Ch. 8.2.5] contingent on additional evidence. In the COMPAS model of Fig. 1, the counterfactual fairness measure of X on the prediction $\hat{Y}$ given the context x, z, w, y is defined as:

$$\mathit{ETT}_{x_1,x_0}(\hat{y}|x,z,w,y) = P(\hat{y}_{x_1}|x,z,w,y) - p(\hat{y}_{x_0}|x,z,w,y). \tag{37}$$

It is verifiable that the difference of counterfactual direct and indirect ERs of Eqs. 30-31 equates to the weight sum of the counterfactual fairness measure over $P(z,w)$, namely,

Theorem 10. *Given the COMPAS model of Fig. 1 and a classifier* $f(x,w,z)$*,* $\mathit{ER}^d_{x_0,x_1}(\hat{y}|x,y)$*,* $\mathit{ER}^i_{x_0,x_1}(\hat{y}|x,y)$ *and* $\mathit{ETT}_{x_1,x_0}(\hat{y}|x,z,w,y)$ *obey the following relationships:*

$$\mathit{ER}^d_{x_0,x_1}(\hat{y}|x,y) - \mathit{ER}^i_{x_1,x_0}(\hat{y}|x,y) = \sum_{z,w} \mathit{ETT}_{x_1,x_0}(\hat{y}|x,z,w,y)P(z,w).$$

Proof. By basic probabilistic operations, the quantity $\mathit{ER}^d_{x_0,x_1}(\hat{y}|x,y) - \mathit{ER}^i_{x_1,x_0}(\hat{y}|x,y)$ can be written as:

$$\begin{aligned}
\mathit{ER}^d_{x_0,x_1}(\hat{y}|x,y) - \mathit{ER}^i_{x_1,x_0}(\hat{y}|x,y) &= P(\hat{y}_{x_1,W_{x_0},Z}|x,y) - P(\hat{y}_{x_0}|x,y) + P(\hat{y}_{x_1}|x,y) - P(\hat{y}_{x_1,W_{x_0},Z}|x,y) \\
&= P(\hat{y}_{x_1}|x,y) - P(\hat{y}_{x_0}|x,y) \\
&= \sum_{z,w}(P(\hat{y}_{x_1}|x,z,w,y) - p(\hat{y}_{x_0}|x,z,w,y))P(z,w) \\
&= \sum_{z,w} \mathit{ETT}_{x_1,x_0}(\hat{y}|x,z,w,y)P(z,w). \qquad \square
\end{aligned}$$

Corol. 2 follows immediately from Thm. 10, which describes the qualitative relationship between the counterfactual fairness measure and the counterfactual direct and indirect ERs.

Corollary 2. *Given the COMPAS model of Fig. 1 and a classifier* $f(x,w,z)$*, if* $\mathit{ETT}_{x_1,x_0}(\hat{y}|x,z,w,y) = 0$ *for any* z, y*, then* $\mathit{ER}^d_{x_0,x_1}(\hat{y}|x,y) - \mathit{ER}^i_{x_1,x_0}(\hat{y}|x,y) = 0$.

Note that Eq. 30 and 31 measure, respectively, the effects along the direct and indirect paths between node X and $\hat{Y}$ in Fig. 2(a). One could show that the counterfactual fairness corresponds to the cumulative effects of all causal paths (including direct and indirect paths) from X to $\hat{Y}$.

Proposition 1. *Given the COMPAS model of Fig. 1, the associated causal diagram G and a classifier* $f(x,w,z)$*, if there exists no causal path from node* X *to* $\hat{Y}$ *in the conditional causal diagram* $G_{|Y}$ *(Fig. 3(b)), i.e.,* $|X \overset{c}{\to} Y|_{G_{|Y}} = 0$*, then for any* $x_0, x_1, x, \hat{y}, z, w, y$*,* $\mathit{ETT}_{x_1,x_0}(\hat{y}|x,z,w,y) = 0$ *holds.*

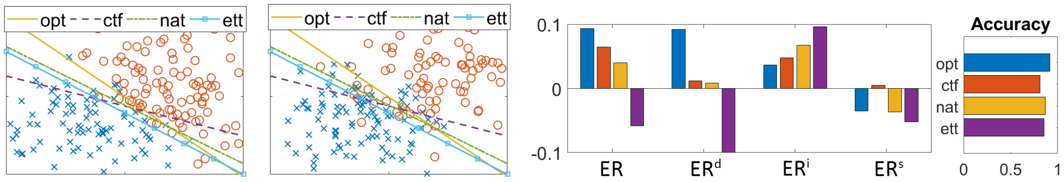
Proof. By [6, Lem. 12], if $|X \overset{c}{\to} Y|_{G_{|Y}} = 0$, then for any x, $\hat{Y}_x = \hat{Y}$. This implies

$$\begin{aligned}
\mathit{ETT}_{x_1,x_0}(\hat{y}|x,z,w,y) &= P(\hat{y}_{x_1}|x,z,w,y) - p(\hat{y}_{x_0}|x,z,w,y) \\
&= P(\hat{y}|x,z,w,y) - p(\hat{y}|x,z,w,y) = 0 \qquad \square
\end{aligned}$$

Thm. 10 and Prop. 1 together imply that the constraints over the counterfactual fairness measure does not necessarily apply to its decomposing counterfactual direct and indirect ERs. Indeed, it is easy to find a simple instance of Fig. 1 where the counterfactual fairness measure is controlled, but the discriminatory effects along the direct and indirect paths are significant, which we will show next.

B.3 A Simple Simulation

We will illustrate the results discussed in this section via simulations on a synthetic dataset. We focus on the *true positive rate* $\mathit{ER}_{x_0,x_1}(\hat{y}_1|y_1)$ where $\hat{y}_1 = y_1 = 1$ and the decomposing counterfactual ERs $\mathit{ER}^d_{x_0,x_1}(\hat{y}_1|x_0,y_1)$, $-\mathit{ER}^i_{x_1,x_0}(\hat{y}_1|x_0,y_1)$ and $-\mathit{ER}^s_{x_1,x_0}(\hat{y}_1|y_1)$ across demographics $x_0 = 0, x_1 = 1$. We shorten the notation and write $\mathit{ER}_{x_0,x_1}(\hat{y}_1|x_0,y_1) = \mathit{ER}$, and similarly to ER^d, ER^i and ER^s.



(a) (b) (c)

Figure 8: Results for simulations in Sec. B. (a-b) Decision boundaries of *opt*, *nat*, *ett* and *ctf* for $X = 0$ (a) and $X = 1$ (b). (c) *ER* stands for the disparate true positive rate $ER_{x_0,x_1}(\hat{y}_1|y_1)$ where $\hat{y}_1 = y_1 = 1$; ER^d, ER^i and ER^s correspond to its decomposing counterfactual effects following Thm. 2.

Consider an instance of the COMPAS model (Fig. 1) where U_Z, U_X, U_W, U_Y are independent exogenous variables, values of Z, X, W, Y are decided by functions

$$z = u_Z, \quad x = I_{\{z+u_X>0\}}, \quad w = 0.5x + 0.5z + u_W, \quad y = I_{\{2x+4w+4z+u_Y>0\}}.$$

U_Z and U_W follow the normal distribution of zero mean and unity variance. U_X, U_X are drawn from the logistic distribution,

$$P(U_X < u) = P(U_Y < u) \triangleq \frac{1}{1+e^{-u}}.$$

We train a logistic classifier taking as arguments values of X, Z, W while constraining its counterfactual ERs $|ER^d| \leq 0.01$ and $|ER^i| \leq 0.05$, labeled as *ctf*. For comparison, we impose the same constraints over the natural direct and indirect effects ($|NDE| \leq 0.01$ and $|NDE| \leq 0.05$) and label the trained classifier as *nat*. We also include an unrestricted classifier *opt* and a classifier *ett* with its counterfactual fairness measure $|ETT| \leq 0.06$. Fig. 8(a-b) show the decision boundaries of *opt*, *nat*, *ett* and *ctf* for $X = 0$ and $X = 1$ respectively. Our analysis (Fig. 8(c)) reveals that *ctf* satisfies the imposed constraints over the counterfactual direct and indirect ERs ($ER^d_{ctf} = 0.01$, $ER^i_{ctf} = 0.048$); *nat* coincides with *ctf* in the counterfactual direct ER ($ER^d_{ctf} = 0.009$) but is larger in the counterfactual indirect ER ($ER^i_{ctf} = 0.07$). The counterfactual fairness (*ett*) controls all causal paths, but effects along the direct and indirect path vary significantly ($ER^d_{ett} = -0.102$, $ER^i_{ett} = 0.096$). In summary, the counterfactual direct and indirect ERs are more specific than *NDE* and *NIE* since they focus on the population $Y = y$; the counterfactual fairness capture the effects along all causal paths, but not an individual direct or indirect path. Neither of these existing counterfactual measures provides a detailed explanation of the disparities in classification errors regarding the underlying mechanisms.

C Experiments Details

In this section, we provide details for simulations and experiments in this paper.

Experiment. 1: Discrete Domains We give the full parametrizations for the causal model used in Experiment. 1. Consider a standard prediction model $\langle M, P(u) \rangle$ of Fig. 6, where all variables (endogenous and exogenous) are binary in $\{0, 1\}$. Values of Z, X, W, Y, D are decided by, respectively, functions

$$z = u_Z, , \quad x = z \oplus u_X, \quad w = x \oplus z \oplus u_W,$$
$$y = x \oplus z \oplus w \oplus u_Y, \quad d = x \oplus z \oplus w \oplus y \oplus u_D.$$

where $\oplus$ stand for the *"xor"* operator. U_Z, U_X, U_W, U_Y, U_D are independent exogenous following the distributions $P(U_Z = 1) = 0.9$, $P(U_X = 1) = 0.5$, $P(U_W = 1) = 0.1$, $P(U_Y = 1) = 0.1$ and $P(U_D = 1) = 0.1$ respectively.

Experiment. 2: COMPAS Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS, is a risk assessment tool, created by the company Northpointe, that is being used across the US to determine whether to release or detain a defendant before his or her trial. Each pretrial defendant receives a COMPAS score based on factors including but not limited to demographics, criminal history, family hi story, and social status. Propublica [1] published two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida that contains scores for over 11000 people who were assessed at the pretrial stage and scored in 2013 and 2014. Besides the

COMPAS score, the data also includes records on defendant's age, gender, race, prior convictions, and whether or not recidivism occurred over a span of two years. We limited our attention to the group consisting of African-Americans and Caucasians. The causal model for this environment is described in Fig. 1.

Acknowledgments

This research is supported in parts by grants from IBM Research, Adobe Research, NSF IIS-1704352, and IIS-1750807 (CAREER).

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 23, 2016.

[2] E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345–7352, 2016.

[3] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] T. Brennan, W. Dieterich, and B. Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009.

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

[6] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

[7] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

[8] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[9] A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[10] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[11] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[12] X. W. Lu Zhang, Yongkai 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-17*, pages 3929–3935, 2017.

[13] J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

[14] J. F. Mahoney and J. M. Mohen. Method and system for loan origination and underwriting, Oct. 23 2007. US Patent 7,287,008.

[15] K. Mancuhan and C. Clifton. Combating discrimination using bayesian networks. *Artificial Intelligence and Law*, 22(2):211–238, Jun 2014.

[16] R. Nabi and I. Shpitser. Fair inference on outcomes. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

[17] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

[18] J. Pearl. Direct and indirect effects. In *Proc. of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 411–420. Morgan Kaufmann, CA, 2001.

[19] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

[20] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

[21] I. Shpitser, T. VanderWeele, and J. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 527–536. AUAI, Corvallis, OR, 2010.

[22] L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

[23] B. van der Zander, M. Liśkiewicz, and J. Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*. AUAI, 2014.

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

[25] S. Wright. The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215, 1934.

[26] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[27] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[28] J. Zhang and E. Bareinboim. Fairness in decision-making — the causal explanation formula. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 2037–2045, 2018.

[29] J. Zhang and E. Bareinboim. Non-parametric path analysis in structural causal models. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018.