Revisiting the Berkeley Admissions data: Statistical Tests for Causal Hypotheses

Sourbh Bhadane[1] **Joris M. Mooij**[1] **Philip Boeken**[1] **Onno Zoeter**[2]

[1]Korteweg-de Vries Institute for Mathematics, University of Amsterdam
[2]Booking.com, Amsterdam, The Netherlands

Abstract

Reasoning about fairness through correlation-based notions is rife with pitfalls. The 1973 University of California, Berkeley graduate school admissions case from Bickel et al. [1975] is a classic example of one such pitfall, namely Simpson's paradox. The discrepancy in admission rates among male and female applicants, in the aggregate data over all departments, vanishes when admission rates per department are examined. We reason about the Berkeley graduate school admissions case through a causal lens. In the process, we introduce a statistical test for causal hypothesis testing based on Pearl's instrumental-variable inequalities [Pearl, 1995]. We compare different causal notions of fairness that are based on graphical, counterfactual and interventional queries on the causal model, and develop statistical tests for these notions that use only observational data. We study the logical relations between notions, and show that while notions may not be equivalent, their corresponding statistical tests coincide for the case at hand. We believe that a thorough case-based causal analysis helps develop a more principled understanding of both causal hypothesis testing and fairness.

1 INTRODUCTION

In the fall of 1973, the Graduate Division of the University of California, Berkeley, made admission decisions for 12763 applicants to its 101 departments. The admission rate for 8442 male applicants was approximately 44.2% and for 4321 female applicants was approximately 34.6%. This disparity prompted Bickel et al. [1975] to investigate whether the Graduate Admissions Office discriminated on the basis of sex. The authors found that despite there being a statistically significant disparity in the aggregate data, when each department was examined, the per-department admission rates did not differ significantly between the sexes, thus making this case an instance of Simpson's paradox. The resolution was that the "proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into". The disparity was therefore attributed to societal biases and the authors concluded that there was "no pattern of discrimination on the part of the admissions committee."

In the fairness literature, the Berkeley graduate admissions case is a canonical example of Simpson's paradox, which illustrates the limitations of correlation-based fairness notions such as demographic parity, and therefore motivates the need for causal reasoning of fairness. Pearl [2009, Section 4.5.4] analyzes the Berkeley example and frames the conclusion of Bickel et al. [1975] as discerning the direct effect of sex on admissions outcome by conditioning on the mediator, namely department choice. Most works in the fairness literature that mention the Berkeley example follow this analysis, which is predicated on the assumption that the causal model includes no latent confounders while the causal graph is akin to a simple mediation graph with sex being the treatment, department choice being the mediator and the admissions decision being the outcome. However, in both Pearl [2009] and Pearl and Mackenzie [2018], Pearl notes that merely conditioning on department choice might not always be appropriate. In particular, he cites a fascinating exchange between William Kruskal and Peter Bickel in Fairley and Mosteller [1977] where Kruskal objects to the analysis in Bickel et al. [1975] by pointing out that controlling for department leads to erroneous conclusions if there is a confounder that affects department choice and admissions outcome. To the best of our knowledge, subsequent works that mention the Berkeley example, do not address the latent confounder issue, including Pearl [2009] where the analysis assumes that the common causes are observed. Further, while there are multiple causal fairness notions proposed in the literature,[1] the issue of statistical testing of these fairness

[1]Some directly inspired by the Berkeley admissions case, for

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:271–295.

notions has received little attention.

In this work, we undertake a causal reasoning exercise centered around the Berkeley admissions case. We take the view that a causal analysis is predicated on causal modeling assumptions that define a family of causal models. A *fairness notion* is either an observational, interventional, counterfactual or a graphical query on a causal model which, as a result, defines a subset of the aforementioned family of causal models, i.e., a fairness notion defines a *causal hypothesis*. Given that we usually have only observational data at hand the question of fairness boils down to statistical testing of a causal hypothesis. Indeed, the Berkeley admissions data can be thought of as sampled from the joint distribution of the sex, department choice and admissions outcome.[2]

For the Berkeley admissions case, we consider multiple fairness notions based on graphical, counterfactual and interventional queries to the family of causal models defined by our causal modeling assumptions, which allows latent confounding between department choice and admissions outcome. For these notions, we develop new statistical tests. One of our key insights is that the graphical notion of fairness can be tested by using the instrumental-variable (IV) inequalities [Pearl, 1995], thus making our proposed statistical test a new test for the IV inequalities. Conversely, *any* statistical test for the IV inequalities can be used to test for fairness in settings that are analogous to the Berkeley case. In the process, we also prove a result of independent interest, namely the sharpness of the IV inequalities for the case where the instrument and the effect are binary, and the treatment takes any finite number of values. For the Berkeley example, while our proposed fairness notions correspond to different rungs of the causal hierarchy and are in general not equivalent, we show, rather surprisingly, that the tests are equivalent within the IV setting. Although our results are inspired by the Berkeley case, they can also be applied in other analogous settings, e.g. to investigate sex discrimination in awarding distinctions to PhD students [Bol, 2023].

1.1 RELATED WORK

The question of fairness in decision-making and predictive systems has received increased attention since the past few decades. See Hutchinson and Mitchell [2019], Barocas et al. [2023] for an excellent historical and technical overview, respectively. While attempts at formalizing fairness lead to correlation-based notions such as fairness through unawareness [Dwork et al., 2012], demographic parity, equality of odds [Hardt et al., 2016] etc., purely observational notions of fairness are at odds with each other Chouldechova [2017], Kleinberg et al. [2017] and are prone to erroneous conclusions. On the other hand, observational notions of fairness are readily translated to statistical tests.

Causal analysis tools such as counterfactuals and interventions provide a framework suitable for fairness. As a result, multiple general fairness notions based on counterfactuals were proposed. Kusner et al. [2017] defined a counterfactual fairness notion that required invariance of the distribution of the decision in a given context, with respect to hypothetical changes in the protected attribute. Nabi and Shpitser [2018] and Zhang et al. [2017] consider path-specific effects. Chiappa [2019] proposes a path-specific counterfactual fairness notion and a related notion appears in the appendix of Kusner et al. [2017]. Another separate line of work seeks to explain observed disparity through causal discrimination mechanisms [Zhang and Bareinboim, 2018, Plečko et al., 2024].

The Berkeley graduate admissions case makes an appearance in multiple papers to motivate the need for causal fairness notions. Kilbertus et al. [2017], Plečko et al. [2024], Kusner et al. [2017], Chiappa [2019], Berk et al. [2023] are a few among many works. In addition, the Berkeley example also serves as a motivation to introduce path-specific notions given the assumption that the direct effect of sex on admissions outcome is the only 'unfair' path. Also, see Barocas et al. [2023] for a critique of this common assumption. Pearl [2009] considers the Berkeley example at length and illustrates the objection to controlling for the mediator by positing an observed confounder.

Despite the fact that most causal fairness works mention the Berkeley example, to the best of our knowledge, no previous work gives a definitive answer to the question of fairness for the Berkeley dataset under unobserved confounding. Kilbertus et al. [2020] discusses the impact of unmeasured confounding under restrictive parametric assumptions. Zhang and Bareinboim [2018], Plečko et al. [2024] consider fairness models that allow for specific forms of unobserved confounding. Schröder et al. [2024] build on this by providing sensitivity analysis on fairness of prediction models. However, the kinds of unobserved confounding that they allow affects the sensitive attribute which is different from the kind we allow for in the Berkeley dataset.

2 PRELIMINARIES

We outline a few definitions that follow the formal setup of Bongers et al. [2021].

Definition 1 (Structural Causal Model (SCM))**.** *A **Structural Causal Model (SCM)** is a tuple $M = (V, W, \mathcal{X}, f, P)$ where a) V, W are disjoint, finite index sets of **endogenous** and **exogenous** random variables respectively, b) $\mathcal{X} = \prod_{i \in V \cup W} \mathcal{X}_i$ is the **domain** which is a product of standard measurable spaces $\mathcal{X}_i$, c) for every $v \in V$, $f_v : \mathcal{X} \mapsto \mathcal{X}_v$ is a measurable function, and $f = (f_v)_{v \in V}$*

example the path-dependent counterfactual fairness notion in Kusner et al. [2017, Appendix S4].

[2]Albeit possibly post-selection, which we don't address in this work.

*is called the **causal mechanism**; the equations $X_v = f_v(X)$ for every $v \in V$ are called the **structural equations**, and d) $P(\mathcal{X}_W) = \bigotimes_{w \in W} P(X_w)$ is the **exogenous distribution** which is a product of probability distributions $P(X_w)$ on $\mathcal{X}_w$.*

Definition 2 (Parent)**.** *Let $M = (V, W, \mathcal{X}, f, P)$ be an SCM. $k \in V \cup W$ is a **parent** of $v \in V$ if and only if it is not the case that for all $x_{V \setminus k}$, $f_v(x_V, X_W)$ is constant in x_k (if $k \in V$, resp. X_k if $k \in W$) $P(\mathcal{X}_W)$-a.s..*

Definition 3 (Causal Graph)**.** *Let $M = (V, W, \mathcal{X}, f, P)$ be an SCM. The **causal graph**, $G(M)$, is a directed mixed graph with nodes V, directed edges $u \longrightarrow v$ if and only if $u \in V$ is a parent of $v \in V$, and bidirected edges $u \leftrightarrow v$ if and only if $\exists w \in W$ that is a parent of both $u, v \in V$.*

For simplicity of exposition, we restrict attention to acyclic SCMs, i.e. SCMs whose causal graph is acyclic (contains no directed cycle $X \to \cdots \to Y \to X$).

Definition 4 (Observational Distribution)**.** *Given an acyclic SCM, $M = (V, W, \mathcal{X}, f, P)$, the exogenous distribution, P and the causal mechanism f induce a probability distribution over the endogenous variables which is called the **observational distribution**, $P_M(X_V)$.*

Definition 5 (Hard Intervention)**.** *Given an acyclic SCM, $M = (V, W, \mathcal{X}, f, P)$, an intervention target $T \subseteq V$, and an intervention value $x_T \in \mathcal{X}_T$, the **intervened SCM** is defined as $M_{do(X_T = x_T)} \triangleq \big(V, W, \mathcal{X}, (f_{V \setminus T}, x_T), P\big)$. Further, the observational distribution of the intervened SCM, $P_{M_{do(X_T = x_T)}}$, is called an **interventional distribution**, and denoted by $P_M(X_V \mid do(X_T = x_T))$.*

Definition 6 (Potential Outcome)**.** *Let $M = (V, W, \mathcal{X}, f, P)$ be an acyclic SCM, $C \subseteq V$ and $x_C \in \mathcal{X}_C$. A random variable $X^{do(x_C)}$ taking values in $\mathcal{X}_V \times \mathcal{X}_W$ is called a **potential outcome** of M under intervention x_C if a) its W-component has the exogenous distribution specified by M, i.e., $X_W^{do(x_C)} \sim P(\mathcal{X}_W)$, and b) it satisfies $X_C^{do(x_C)} = x_C$ a.s. and*

$$X_{V \setminus C}^{do(x_C)} = f_{V \setminus C}\Big(x_C, X_{V \setminus C}^{do(x_C)}, X_W^{do(x_C)}\Big) \text{ a.s..}$$

If $C = \emptyset$, we write X instead of $X^{do(x_\emptyset)}$. When dealing with multiple potential outcomes, we assume that they share the same W-component: for any n, for $C_1, C_2, \cdots, C_n \subseteq V$ and for $x_{C_1} \in \mathcal{X}_{C_1}, x_{C_2} \in \mathcal{X}_{C_2}, \cdots, x_{C_n} \in \mathcal{X}_{C_n}$,

$$X_W^{do(x_{C_1})} = X_W^{do(x_{C_2})} = \cdots = X_W^{do(x_{C_n})} = X_W \text{ a.s..}$$

Instrumental variables (IVs) are used to estimate the causal effect of a treatment X on the outcome Y in the presence of latent confounding.

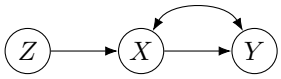


Figure 1: Causal graph of $M \in \mathbb{M}_{\text{IV}}$

Definition 7 (Instrumental Variable (IV) Model Class)**.** *The **instrumental variable model class**, $\mathbb{M}_{IV}$, is a collection of SCMs M such that $G(M)$ is a subgraph of Figure 1 where Z is the **instrument**, X is the **treatment**, and Y is the **outcome**.*

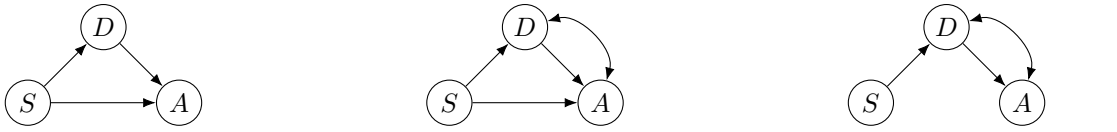
We make an explicit positivity assumption and define $\mathbb{M}_{\text{IV}+} \triangleq \{M \in \mathbb{M}_{\text{IV}} : \forall z, P_M(Z = z) > 0\}$. Causal graphs of SCMs in the IV model class rule out the directed edge $Z \to Y$ and any latent confounding between Z and Y as well. In Section D.4 we show that our results for the IV model class hold even when confounding is allowed between Z and X.

3 BERKELEY CASE: NO LATENT CONFOUNDING

Under the semantic framework of SCMs, we first make the same causal modeling assumptions that are commonplace in works that mention the Berkeley admissions case. We compare fairness notions that are tied to these modeling assumptions, with the view that modeling assumptions describe a family of SCMs and fairness notions define a subset of this family. We relate existing general notions of fairness in the literature to this viewpoint. While this is a re-examination of the various existing analyses of the Berkeley admissions case, in the next section, we relax the causal modeling assumptions and consider the more general family of models that allow for confounding between department choice and admissions outcome.

The set of endogenous variables consists of the protected attribute, namely sex of the applicant, S, the department they applied to, D, and the decision of the admissions committee, A. We assume that S, A are binary variables and D is a discrete-valued variable taking finite number of values, where $S = 0, 1$ corresponds to male, female applicants, respectively, and $A = 0, 1$ corresponds to reject and accept, respectively.[3] Given that, possibly, societal biases nudge applicants to departments at differing rates depending on their sex, we assume that S affects D. Since departments are the primary decision-making units and have different admission rates, we also assume that D affects A. The question of whether acceptance decisions discriminate against sex centers around the *direct* causal effect of S on A (w.r.t. the three variables S, D, A), and therefore we allow such an effect in the model. Indeed, if such a direct causal effect is absent, then the applicant's sex does not directly affect

[3]The assumption of binary sex is purely for mathematical simplicity.



(a) $M \in \mathbb{M}_{\text{no-cf}}$ (b) $M \in \mathbb{M}_{\text{cf}}$ (c) $M \in H^0_{\text{cf-graph+}}$ and $M \in \mathbb{M}_{\text{IV+}}$

Figure 2: Causal graphs, $G(M)$, assumed in various model classes.

the admission outcome (only indirectly through department choice). On the other hand, if such a direct causal effect is present, then we would consider this unfair, provided that we know that it is not mediated purely via latent unprotected attributes (e.g., undergraduate department, or undergraduate university chosen by the applicant) that could play a similar 'fair' mediating role as D. We will—for the time being—assume no latent unprotected attributes exist that mediate an effect of sex on admissions outcome, and come back to this point in Section 6.

In this section, we also assume the absence of any confounding between the variables (in addition to the absence of any selection bias). The structural equations are given by

$$\begin{aligned} S &= f_S(U_S), \\ D &= f_D(S, U_D), \\ A &= f_A(S, D, U_A), \end{aligned} \tag{1}$$

where U_S, U_D and U_A denote independent exogenous random variables. We denote the family of SCMs parameterized by the functions in (1) and the exogenous distribution as $\mathbb{M}_{\text{no-cf}}$. For $M \in \mathbb{M}_{\text{no-cf}}$, the causal graph $G(M)$ is a subgraph of the directed acyclic graph (DAG) in Figure 2a.

3.1 FAIRNESS NOTIONS

We define a fairness notion to be a certain condition that is required to be satisfied by a causal model to be deemed fair. These conditions can take the form of constraints expressed in terms of observational, interventional, counterfactual or graphical queries on the SCMs in the families of causal models defined by modeling assumptions, in our case, $\mathbb{M}_{\text{no-cf}}$. While the criteria for the fairness notions in Section 3 are phrased in terms of queries corresponding to different rungs of the causal ladder, in our case, any condition can only be tested using observational data.

The investigation of Berkeley's admission data was initiated on the observation that the well-known fairness notion of demographic parity $P_M(A = 1 \mid S = 0) = P_M(A = 1 \mid S = 1)$ did not hold. This fairness notion is based purely on observational data and we have already noted that it falls prey to Simpson's paradox. We now present another observational notion of fairness that can be seen as a conditional version of demographic parity.

Definition 8 (Observational Notion of Fairness)**.** *$M \in \mathbb{M}_{\text{no-cf}}$ is fair according to the observational notion of fairness if it belongs to*

$$\begin{aligned} H^0_{\text{no-cf-obs}} \triangleq \{M \in \mathbb{M}_{\text{no-cf}} : \forall d, s, P_M(D = d, S = s) > 0 \\ \implies P_M(A = 1 \mid S = s, D = d) = P_M(A = 1 \mid D = d)\}. \end{aligned}$$

Bickel et al. [1975] proposed this notion for the Berkeley data. It is equivalent to the conditional independence $A \perp\!\!\!\perp S \mid D$. Hence, any valid conditional independence test for $A \perp\!\!\!\perp S \mid D$ provides a statistical test for this notion. Indeed, the analysis of Bickel et al. [1975] shows that the data contain not enough evidence to reject the null hypothesis that this conditional independence holds, and therefore, concludes fairness.

From the causal graph of $\mathbb{M}_{\text{no-cf}}$ in Figure 2a, a natural subset of fair causal models is those without the edge $S \to A$.

Definition 9 (Graphical Notion of Fairness)**.** *$M \in \mathbb{M}_{\text{no-cf}}$ is fair according to the graphical notion of fairness if it belongs to the null hypothesis set $H^0_{\text{no-cf-graph}} \triangleq \{M \in \mathbb{M}_{\text{no-cf}} : S \to A \notin G(M)\}$.*

Pearl [2009, Section 4.5.3] discusses the direct effect in the context of the Berkeley admissions example, where he objects to conditioning on the department and instead proposes intervening on department choice, which corresponds to the controlled direct effect (CDE) [Pearl, 2001] of the 'treatment', S, on the outcome, A, for every value of the mediator, i.e., every department choice d.

Definition 10 (Interventional Notion of Fairness)**.** *$M \in \mathbb{M}_{\text{no-cf}}$ is fair according to the interventional notion of fairness if it belongs to the null hypothesis set*

$$\begin{aligned} H^0_{\text{no-cf-inter}} \triangleq \{M \in \mathbb{M}_{\text{no-cf}} : \forall d, s \\ P_M(A = 1 \mid do(S = s), do(D = d)) \\ = P_M(A = 1 \mid do(D = d))\}. \end{aligned}$$

Recent analyses of the Berkeley example emphasize counterfactual notions of fairness. In Pearl [2009, Section 4.5.4], Pearl and Mackenzie [2018], Pearl considers a counterfactual quantity, namely the natural direct effect (NDE) [Robins and Greenland, 1992, Pearl, 2001] by motivating a hypothetical experiment where "all female candidates retain their department preferences but change their gender [sex] identification (on the application form) from female to male".

Subsequent causal fairness works [Nabi and Shpitser, 2018, Chiappa, 2019] build on this and propose fairness notions based on known path-specific versions of NDE where the 'direct path' from S to A is viewed as 'unfair' as opposed to the 'fair' path $S \to D \to A$. For the Berkeley example, the NDE$(s \to s')$ is given by

$$P_M\Big(A^{\text{do}(S=s', D=D^{\text{do}(S=s)})} = 1\Big) - P_M\Big(A^{\text{do}(S=s)} = 1\Big)$$

for $s \neq s'$. Note that by Pearl's mediation formula [Pearl, 2001], the above is identified (assuming $\forall d, s, P_M(D = d, S = s) > 0$) as

$$\sum_d \left(P_M(A = 1 \mid D = d, S = s') - P_M(A = 1 \mid D = d, S = s)\right) P_M(D = d \mid S = s).$$

This implies that if the observational notion of fairness and positivity hold, the NDE is 0. However, the converse is not necessarily true. For example, if one department favors male applicants and another favors female applicants, then the NDE could be 0 while it is not necessary that the observational notion of fairness holds.

Other counterfactual notions of fairness include those by Kusner et al. [2017]. The authors define a counterfactual fairness notion that implies demographic parity (see Section E.1.1 for a proof) for the Berkeley example; we have already seen that this particular fairness notion falls prey to Simpson's paradox. In the appendix, however, they define a path-dependent notion of counterfactual fairness.[4] In Section E.1.2 we show that, in our setting, testing for the path-dependent counterfactual fairness notion is equivalent to testing for the conditional independence $A \perp\!\!\!\perp S \mid D$. We now propose an alternate counterfactual notion of fairness and later compare testing of the same.

Definition 11 (Counterfactual Notion of Fairness)**.** *$M \in \mathbb{M}_{\text{no-cf}}$ is fair according to the counterfactual notion of fairness if it belongs to the null hypothesis set*

$$H^0_{\text{no-cf-ctrf}} \triangleq \{M \in \mathbb{M}_{\text{no-cf}} : \forall d, s, P_M(A^{do(S=s, D=d)} = A^{do(D=d)}) = 1\Big\}$$

The alternative hypotheses are given by the complement of the null hypotheses w.r.t. $\mathbb{M}_{\text{no-cf}}$. We consider the interventional and counterfactual fairness notions as analogues of the absence of direct effect that are framed in terms of rung-2 and rung-3 notions, namely interventional and counterfactual distributions, respectively. Therefore, the interventional notion of fairness considers comparing interventional distributions that result from intervening on all the variables with intervening on all variables but the protected variable, and the counterfactual fairness notions compare potential outcomes resulting from the same set of interventions. Given that the notions are defined on different rungs of the causal hierarchy, it is perhaps not surprising that they are nested accordingly. The assumption of no confounding simplifies the relations as we can prove equivalence of a few notions under positivity. The proof is deferred to Section C.1.

Lemma 12.

$$H^0_{\text{no-cf-graph}} = H^0_{\text{no-cf-ctrf}} \subset H^0_{\text{no-cf-inter}} \subset H^0_{\text{no-cf-obs}}.$$

If for all s, d, $P_M(s, d) > 0$, then in addition, we have $H^0_{\text{no-cf-inter}} = H^0_{\text{no-cf-obs}}$.

Despite the nested nature of the fairness notions at different rungs of the causal hierarchy, we prove that the sets of observational distributions that these notions induce are identical. The proof is in Section C.2.

Theorem 13. *Let*

$$\mathcal{P}_{\text{no-cf-graph}} \triangleq \left\{P_M(D, A, S) : M \in H^0_{\text{no-cf-graph}}\right\},$$
$$\mathcal{P}_{\text{no-cf-ctrf}} \triangleq \left\{P_M(D, A, S) : M \in H^0_{\text{no-cf-ctrf}}\right\},$$
$$\mathcal{P}_{\text{no-cf-inter}} \triangleq \left\{P_M(D, A, S) : M \in H^0_{\text{no-cf-inter}}\right\},$$
$$\mathcal{P}_{\text{no-cf-obs}} \triangleq \left\{P_M(D, A, S) : M \in H^0_{\text{no-cf-obs}}\right\}.$$

Then $\mathcal{P}_{\text{no-cf-graph}} = \mathcal{P}_{\text{no-cf-ctrf}} = \mathcal{P}_{\text{no-cf-inter}} = \mathcal{P}_{\text{no-cf-obs}}$.

In summary, despite the fact that we analyze the Berkeley admissions case using multiple fairness notions, under the assumption of no confounding, with observational data, they can all be tested using a conditional independence test.

If the data contains enough evidence to reject conditional independence, then the data generating mechanism is unfair w.r.t. the observational notion of fairness (assuming no latent unprotected mediators). On the other hand, if the data does not contain enough evidence to reject conditional independence, then the data generating mechanism is fair w.r.t. the observational notion of fairness. However, this extrapolation of the outcome of the statistical test on the fairness implications does not hold for the interventional, counterfactual and graphical notions. The following example illustrates that for the graphical notion of fairness, an unfaithful causal model, where A is directly affected by S, could satisfy conditional independence.

Example 14. *Let $M \in \mathbb{M}_{\text{no-cf}}$ be defined as $U_S \sim Ber(\delta), U_A \sim Ber(\varepsilon), U_D \sim Ber(\frac{1}{2})$ where $\delta, \varepsilon \in [0, 1]$ and $S = U_S, D = S \oplus U_D, A = S \oplus D \oplus U_A$. Here, $A \perp\!\!\!\perp S \mid D$ but S is a parent of A, i.e., $M \in H^0_{\text{no-cf-obs}}$, but $M \notin H^0_{\text{no-cf-graph}}$.*

For the interventional notion of fairness, the following example illustrates that a causal model that violates positivity could satisfy conditional independence but not the interventional notion of fairness.

[4] This notion is specifically motivated by the Berkeley example.

Example 15. *Let $M \in \mathbb{M}_{\text{no-cf}}$ be defined as $U_S = 0, U_D = 0, U_A \sim Ber(\varepsilon)$ where $\varepsilon \in [0, \frac{1}{2})$, and $S = 0, D = 0, A = S \oplus U_A$. Here $A \perp\!\!\!\perp S \mid D$, but for all d, $P_M(A = 1 \mid do(S = 1), do(D = d)) = 1 - \varepsilon \neq P_M(A = 1 \mid do(D = d)) = \varepsilon$. Therefore, $M \in H^0_{\text{no-cf-obs}}$, but $M \notin H^0_{\text{no-cf-inter}}$.*

So, if the outcome of the test is that conditional independence cannot be rejected ($M \in H^0_{\text{no-cf-obs}}$), then due to the aforementioned observations, we cannot conclude that the underlying causal model belongs to the causal null hypothesis of the interventional or counterfactual or graphical fairness notions, i.e., our conclusion is that fairness is "undecidable" with respect to these notions. However, if the outcome of the statistical test is that there is enough evidence in the data to reject conditional independence ($M \notin H^0_{\text{no-cf-obs}}$), then we can conclude that the underlying causal model does not belong to the causal null hypothesis of *any* of the fairness notions. If we rule out the existence of any latent, unprotected mediator between S and A, we can conclude that the data generating mechanism is unfair.

In the next section, we enlarge the class of models to allow for confounding between D and A and perform a similar reasoning exercise.

4 BERKELEY CASE: WITH LATENT CONFOUNDING BETWEEN DEPARTMENT AND OUTCOME

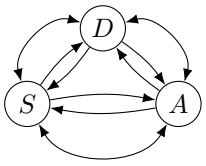


Figure 3: Causal graph of a model without assumptions

We now take a more careful causal modeling approach. Instead of starting from variables and reasoning about structural equations that we allow, we start with assuming that all structural equations exist.[5] For the Berkeley example, Figure 3 shows a causal graph of an SCM that we start with. We now provide rationale for ruling out few structural equations. Based on chronology of events, we rule out those where D directly affects S, where A directly affects D and where A directly affects S. We rule out unobserved common causes of S and D, and S and A since we model S to be sex at birth. While latent selection bias might introduce bidirected edges [Chen et al., 2024] that are incident on S, we assume for now that there is no selection bias in the dataset. The resulting class of SCMs has structural equations of the form $S = f_S(U_S), D = f_D(S, U, U_D), A = f_A(S, D, U, U_A)$ where U, U_S, U_D and U_A denote independent exogenous random variables. We define $\mathbb{M}_{\text{cf}}$ to be the family of models parameterized by the above structural equations and the exogenous distribution. Further, we define $\mathbb{M}_{\text{cf}+} = \{M \in \mathbb{M}_{\text{cf}} : \forall s, P_M(S = s) > 0\}$. For $M \in \mathbb{M}_{\text{cf}}$ (and $\mathbb{M}_{\text{cf}+}$), the causal graph is a subgraph of the one shown in Figure 2b.

Although we arrived at allowing confounding between department and outcome through a careful causal modeling approach, this is not a novel consideration. In particular, Kruskal [Fairley and Mosteller, 1977, Pg 128-129] demonstrated an example where the existence of a latent confounder, such as state of residence, can render Bickel et al. [1975]'s analysis incorrect. Other natural latent confounders include, for example, level of department-specific technical skills that influence both the department choice of an applicant and the admissions outcome.

Since our modeling assumptions expand the family of SCMs under consideration to $\mathbb{M}_{\text{cf}+}$, the fairness notions that we discussed in the previous section are modified accordingly to obtain null hypothesis sets $H^0_{\text{cf-graph}+}, H^0_{\text{cf-inter}+}$ and $H^0_{\text{cf-ctrf}+}$.

Definition 16 (Fairness Notions with Confounding)**.** *For $M \in \mathbb{M}_{cf+}$ the null hypothesis set corresponding to the interventional, counterfactual and graphical notion of fairness are*

$$
\begin{aligned}
H^0_{\textit{cf-inter}+} &\triangleq \{M \in \mathbb{M}_{cf+} : \forall d, s, \\
&\quad P_M(A = 1 \mid do(S = s), do(D = d)) \\
&\quad = P_M(A = 1 \mid do(D = d))\}, \\
H^0_{\textit{cf-ctrf}+} &\triangleq \{M \in \mathbb{M}_{cf+} : \forall d, s, \\
&\quad P_M(A^{do(S=s, D=d)} = A^{do(D=d)}) = 1\Big\}, \\
H^0_{\textit{cf-graph}+} &\triangleq \{M \in \mathbb{M}_{cf+} : S \to A \notin G(M)\}.
\end{aligned}
$$

While the above notions generalize straightforwardly from the no-confounder setting, this is no longer the case for the observational notion. In addition, while the statistical tests for the no-confounder model are straightforward, this is no longer the case for the aforementioned null hypotheses since $A \not\perp\!\!\!\perp S \mid D$ in general.

4.1 GRAPHICAL NOTION AND THE INSTRUMENTAL VARIABLE (IV) INEQUALITIES

In the presence of latent confounding, graphical queries, such as absence of edges, impose equality or inequality constraints [Evans, 2016, Wolfe et al., 2019] in addition to conditional independence constraints which are the only

[5] Since this allows for causal cycles, this would require using e.g. the framework of simple SCMs [Bongers et al., 2021].

constraints imposed by a DAG. For the Berkeley case with confounding, since the path $S \rightarrow D \leftrightarrow A$ is open when conditioned on D, we have $S \not\perp\!\!\!\perp A \mid D$ in general. Our test for the graphical notion of fairness for $\mathbb{M}_{\text{cf}+}$ stems from the observation that a model $M \in H^0_{\text{cf-graph}+}$ lies in the instrumental variable (IV) model class $\mathbb{M}_{\text{IV}+}$ where S is considered the instrument, D the treatment, and A the effect. If all modeled endogenous variables are discrete-valued, a necessary condition for the observational distribution[6] resulting from $M \in \mathbb{M}_{\text{IV}+}$ is to satisfy the IV inequalities [Pearl, 1995], which in the context of Figure 1 are given by

$$\max_x \sum_y \max_z P_M(X = x, Y = y \mid Z = z) \leq 1. \quad (2)$$

Since the IV inequalities are only necessary conditions, an arbitrary distribution on X, Y, Z that satisfies the IV inequality does not necessarily imply that it is an entailed distribution of a model from the IV model class. Bonet [2001] showed that for the binary instrument, treatment and effect case, the IV inequalities are also sufficient conditions. In Theorem 17, we show that for the case where the instrument and outcome are binary and the treatment is discrete-valued with finite support, any distribution that satisfies the IV inequality is also entailed by some causal model from the IV model class. To the best of our knowledge, Theorem 17 is a novel result. We defer the proof to Section D.1.

Theorem 17. *Let X, Y, Z be discrete random variables defined on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively, with $|\mathcal{X}| = n \geq 2, |\mathcal{Y}| = 2, |\mathcal{Z}| = 2$. Define $\mathcal{P}_{\mathbb{M}_{IV+}} \triangleq \{P_M(X, Y, Z) : M \in \mathbb{M}_{IV+}\}$ and the set of joint distributions that satisfy the IV inequalities as*

$$\mathcal{P}_{IV+} \triangleq \{P(X, Y, Z) : P(X, Y \mid Z) \text{ satisfies } (2) \text{ and } \forall z, P(Z = z) > 0\}.$$

Then $\mathcal{P}_{IV+} = \mathcal{P}_{\mathbb{M}_{IV+}}$.

For the Berkeley admissions case, the observational distribution satisfying the IV inequalities implies that there exists a causal explanation (model) where the directed edge $S \rightarrow A$ is absent, i.e., given that $P(A, D, S) \in \mathcal{P}_{\text{IV}+}$, there exists $M \in \mathbb{M}_{\text{IV}+}$ such that $P_M(A, D, S) = P(A, D, S)$. On the other hand, the observational distribution violating the IV inequalities does not necessarily imply that the edge $S \rightarrow A$ is present since the IV model class, $\mathbb{M}_{\text{IV}+}$, is only a subset of all the models that do not contain the edge $S \rightarrow A$ in the causal graph. For example, the existence of latent confounding between S and A in a model M may result in $M \notin \mathbb{M}_{\text{IV}+}$, even though $G(M)$ does not necessarily contain the directed edge $S \rightarrow A$. However, the causal modeling assumption that defined $\mathbb{M}_{\text{cf}+}$ rules out latent confounding between S and A. Therefore, given our modeling assumptions, $H^0_{\text{cf-graph}+} = \mathbb{M}_{\text{IV}+}$, and in turn, we conclude that violating the IV inequalities implies that $M \in \mathbb{M}_{\text{cf}+} \backslash H^0_{\text{cf-graph}+}$. As in the previous section, it is possible that causal models that lie outside $H^0_{\text{cf-graph}+}$ ("unfair" models) induce observational distributions that lie in $\mathcal{P}_{\text{IV}+}$, i.e., satisfy the IV inequalities. Therefore, satisfying the IV inequalities is not conclusive evidence that the data-generating mechanism is fair, i.e., our conclusion should be that fairness is undecidable. In Section 5 we introduce a Bayesian test for the IV inequalities.

[6] While we express the IV inequalities as a condition satisfied by the observational distribution, in Section B we reason that they are more appropriately expressed as conditions in terms of $P_M(X, Y \mid \text{do}(Z))$.

4.2 BOUNDS ON INTERVENTIONAL NOTION OF FAIRNESS

For $M \in \mathbb{M}_{\text{cf}+}$, the interventional notion of fairness is equivalent to a vanishing CDE, where the latter is not identifiable in our case. By a response-function parameterization [Balke, 1995, Balke and Pearl, 1997] of $M \in \mathbb{M}_{\text{cf}+}$, we can express the interventional distributions in Definition 16 as a linear function of response variables. Further, the observational distribution is also expressed as a linear function of the response variables. Using the symbolic linear programming approach of Balke [1995], we obtain upper and lower bounds in terms of the observational distribution, specifically, $P_M(A, D \mid S)$. Indeed, Cai et al. [2008] express the same bounds which we reproduce below. The CDE

$$P_M(A = 1 \mid \text{do}(S = 1), \text{do}(D = d)) - P_M(A = 1 \mid \text{do}(S = 0), \text{do}(D = d)),$$

lies in the interval

$$[\Pr(A = 1, d \mid S = 1) + \Pr(A = 0, d \mid S = 0) - 1, 1 - \Pr(A = 0, d \mid S = 1) - \Pr(A = 1, d \mid S = 0)].$$

For the interventional notion of fairness, the CDE must be 0 for all d. By setting the lower bound to be at most 0 and the upper bound to be at least 0, we recover the IV inequalities in (2). While Cai et al. [2008] do not point out the connection to the IV inequalities, they find it "remarkable that we [they] get such a simple formula, consisting of only one additive expression in the lower bound and one additive expression in the upper bound". In the next subsection, we show that the connection to the IV inequalities is not a coincidence.

4.3 A FAMILY OF EQUIVALENT TESTS

The graphical and interventional fairness notions end up imposing identical constraints on the observational distribution. However, note that $H^0_{\text{cf-inter}+} \supseteq H^0_{\text{cf-graph}+}$. In fact, we prove in Section D.2 that $H^0_{\text{cf-graph}+} = H^0_{\text{cf-ctrf}+} \subset H^0_{\text{cf-inter}+}$. Models in $H^0_{\text{cf-inter}+} \backslash H^0_{\text{cf-graph}+}$ (Example 14 with $\varepsilon = \frac{1}{2}$) are such that the edge $S \rightarrow A$ exists in the causal graph

and yet, the interventional queries in Definition 16 are equal. Given that the null hypothesis of the interventional fairness notion is a strict superset of that of the graphical fairness notion, we might expect the same relation to hold in the resulting set of observational distributions for these hypotheses, thus giving us potentially different tests. In contrast, like in Section 3, we show that the corresponding sets of observational distributions resulting from models in $H^0_{\text{cf-inter+}}$, $H^0_{\text{cf-ctrf+}}$, $H^0_{\text{cf-graph+}}$ are identical. Section D.3 contains the proof.

Theorem 18. *Let*

$$\mathcal{P}_{\textit{cf-graph}} \triangleq \left\{P_M(D,A,S) : M \in H^0_{\textit{cf-graph+}}\right\},$$
$$\mathcal{P}_{\textit{cf-inter}} \triangleq \left\{P_M(D,A,S) : M \in H^0_{\textit{cf-inter+}}\right\},$$
$$\mathcal{P}_{\textit{cf-ctrf}} \triangleq \left\{P_M(D,A,S) : M \in H^0_{\textit{cf-ctrf+}}\right\}.$$

Then $\mathcal{P}_{\textit{cf-inter}} = \mathcal{P}_{\textit{cf-ctrf}} = \mathcal{P}_{\textit{cf-graph}} = \mathcal{P}_{IV+}$, where $\mathcal{P}_{IV+}$ is defined in Theorem 17.

In summary, testing for the graphical, interventional and counterfactual notions of fairness, with confounding, all boil down to testing the IV inequalities.

4.4 COMPARISON WITH EXISTING FAIRNESS NOTIONS

The utility of considering statistical tests is that we can now compare different fairness notions for a particular case with respect to the same causal modeling assumptions. In this section, we consider the three existing counterfactual fairness notions, namely the NDE [Nabi and Shpitser, 2018, Chiappa, 2019], and the counterfactual and path-dependent counterfactual fairness notions in Kusner et al. [2017].

For NDE, Kaufman et al. [2005] obtain bounds for the all-binary setting. Using these bounds, we obtain a strictly weaker test than the IV inequalities.[7] We show in Section E.2.2 that the counterfactual notion of fairness of Kusner et al. [2017] implies demographic parity even when confounding is allowed. In Section E.2.3 we show that testing the path-dependent counterfactual fairness notion of Kusner et al. [2017] is equivalent to testing the IV inequalities.

5 BAYESIAN TESTING PROCEDURE

We start with proposing a Bayesian test for IV inequalities using finite data. Although Ramsahai [2008] proposed a frequentist test and derived the distribution of the likelihood ratio, providing a means to obtain a p-value for the case of a binary treatment, it is unclear what the test would be for our case where the "treatment" is not binary. In addition, Wang et al. [2017] provide a frequentist test that involves multiple one-sided independence tests. In contrast, the Bayesian test has straightforward extensions to other hypotheses that are expressed in terms of the observational distribution, including bounds on unidentifiable causal queries.

Consider discrete random variables X, Y, Z where $|\mathcal{X}| = n, |\mathcal{Y}| = 2, |\mathcal{Z}| = 2$. The observational distribution $P(X,Y,Z)$ lies in the $4n-1$ dimensional simplex, denoted by Δ, which is a subset of $\mathbb{R}^{4n}$. We consider a Bayesian model selection procedure where

$$\mathbb{M}_0 = \{\theta \in \Delta : \theta \text{ satisfies the IV inequalities}\},$$
$$\mathbb{M}_1 = \{\theta \in \Delta : \theta \text{ does not satisfy the IV inequalities}\}.$$

Note that $\mathbb{M}_0 \dot\cup \mathbb{M}_1 = \Delta$. Given a finite dataset of (X,Y,Z) tuples, denoted by $R_1, R_2, \cdots, R_m$, and choices of prior distributions for the models, $\pi(\theta \mid \mathbb{M}_0), \pi(\theta \mid \mathbb{M}_1)$, we report a confidence interval for the posterior probability of satisfying the IV inequalities, i.e., $P(\mathbb{M}_0 \mid R_1, R_2, \cdots, R_m) = \int_{\theta \in \mathbb{M}_0} P(\theta \mid R_1, R_2, \cdots, R_m) d\theta$. Given the posterior density $P(\theta \mid R_1, R_2, \cdots, R_m)$, we estimate the posterior probability of $\mathbb{M}_0$ by IID sampling from the posterior density n times and counting how often the sample satisfies the IV inequalities, which we denote by N. Since N is a binomial random variable with parameters n and $P(\mathbb{M}_0 \mid R_1, R_2, \cdots, R_m)$, a confidence interval on $P(\mathbb{M}_0 \mid R_1, R_2, \cdots R_m)$ is readily obtained by the Clopper-Pearson method [Clopper and Pearson, 1934].

Results on Berkeley admission data: We use the `UCBAdmissions` [R Core Team, 2023] dataset from `R` that contains counts for each sex-department-admissions outcome tuple for the 6 largest departments. Therefore, $|\mathcal{X}| = 6, |\mathcal{Y}| = 2, |\mathcal{Z}| = 2$. Since the data satisfies the positivity for sex, we can use the Bayesian model selection procedure. For parameters $\theta = (P(d,a,s) : s \in \mathcal{X}_S, d \in \mathcal{X}_D, a \in \mathcal{X}_A)$,we choose a flat Dirichlet prior over Δ giving us $\pi(\theta|\mathbb{M}_i) = c_i \text{Dir}(1,1,\cdots,1)\mathbf{1}[\theta \in \mathbb{M}_i]$ where c_i is a normalizing constant. The counts from the data are used to obtain the posterior, $P(\theta \mid R_1, R_2, \cdots, R_m)$ which is also a truncated Dirichlet distribution. Using $n = 10^6$ samples, we observe no violations of the IV inequalities. Therefore, the confidence interval for the posterior probability of the Berkeley data satisfying the IV inequalities is $\left[1 - 3.69 \times 10^{-6}, 1\right]$. As mentioned in Section 4.1 satisfying the IV inequalities implies that fairness is undecidable. In Section F, we carry out a sensitivity analysis by varying the chosen prior. We also report results on a different dataset from Bol [2023] that investigates sex-based discrimination in awarding cum-laude distinctions to graduate students.

In addition to the Bayesian test, the maximum likelihood (ML) estimator satisfies the IV inequalities, im-

[7] Intuitively, the reason is the same as in Section 3; the NDE averages over departments, and a positive bias in one department may cancel out against a similarly strong negative bias in another department. Hence, vanishing NDE does not imply that each department takes fair decisions.

plying that there isn't enough evidence to reject the null hypothesis when doing a likelihood ratio test. An implementation of Wang et al. [2017] for the Berkeley dataset also does not reject the null hypothesis (see Section F for details). The code can be found at `https://github.com/SourbhBh/BerkeleyCode`.

6 DISCUSSION

The Berkeley admissions case is a canonical example in the causal fairness literature. Bickel et al. [1975] reached the conclusion of there being no evidence to reject fairness, although under the unrealistic assumption of no unobserved confounding. When allowing for unobserved confounding, we arrived at a different conclusion: since there is very strong evidence that the data satisfies the IV inequalities, it is undecidable from the available data whether the admission procedure was fair or discriminated against sex.

While our analysis was centered around the Berkeley case, there are multiple aspects that generalize—a) $\mathbb{M}_{\text{cf}}$ can be thought of as a mediator with a confounder between mediator and outcome, which is common in mediation analysis. b) The approach of fairness notions being causal hypotheses, with respect to the class of models defined by modeling assumptions, that need to be translated into statistical tests to be useful in practice. c) The observation that for the case of inequality constraints on observational data, a straightforward Bayesian testing procedure is available.

Generalization: Non-binary variables: While the Berkeley dataset had a binary protected attribute and a binary decision outcome, our approach can be generalized to non-binary variables. The response-function parametrization of $\mathbb{M}_{\text{IV}+}$ can be used to characterize the set of induced observed distributions of X, Y and Z as a convex polyhedral set. Using computer algebra, this polyhedral set can be characterized by linear inequality constraints even when the alphabet sizes are arbitrary. However, the number of linear inequality constraints quickly explodes for non-binary Z (e.g. for binary X and Y, and for $|\mathcal{Z}| = 2, 3, 4$ and 5, we get $12, 48, 160$ and 420 inequalities). Nevertheless, these sets of inequality constraints can be tested using a Bayesian testing procedure akin to the one we outline in Section 5, thus giving a statistical test for all our fairness notions since the statement of Theorem 18 holds for any alphabet size.

Generalization: Relaxing unobserved confounding assumptions In Section 4 we made the assumption of no unobserved confounding between S and any other variable. In Section D.4 we show that our main theorems, Theorem 17 and Theorem 18 hold even in the case of allowing for confounding between S and D. If unobserved confounding is allowed between S and A, the set of observational distributions induced by causal models where the direct effect between sex and admissions outcome is absent is the entire simplex which is the same as the set of observational distributions induced by causal models where the direct effect between sex and admissions outcome is present. Therefore, from observational data, it is not possible to distinguish between the presence/absence of a direct effect from sex to admissions outcome if we allow for confounding between sex and admissions outcome.

Unmeasured Mediators: Our conclusions and fairness notions are with respect to the measured variables. The presence of unmeasured mediators could change the interpretation of the results. For example, an unmeasured mediator that is not a 'protected' variable, such as choice of undergraduate department, would result in a direct effect of sex on admissions outcome in the marginalized causal model, but might still be considered 'fair'. Therefore, even if the data does not satisfy the IV inequalities, we can only claim the existence of the direct effect from sex to admissions outcome and not whether there is unfairness. However, if the existence of latent unprotected mediators is ruled out, then we can consider the existence of a direct effect as 'unfair'.

Undecidability: Our analysis can be viewed as a Bayesian model comparison between causal models (i) without a direct effect of S on A, vs. (ii) with a direct effect of S on A. If we only have access to observational data then causal models in (ii) result in a saturated model in the observed distribution space, i.e., the set of induced observational distributions is the entire simplex, whereas for the causal models in (i), the IV inequalities hold. Therefore, satisfying the IV inequalities implies that fairness is 'undecidable' since causal models in both (i) and (ii) satisfy the IV inequalities. Violating the IV inequalities, however, would imply that there is a direct effect of sex on admissions outcome. Only with the additional modeling assumption of there being no unprotected mediators between S and A, could it then be concluded that the admissions process was 'unfair'.

Selection Bias: `UCBAdmissions` dataset only has data from the 6 largest departments as opposed to 85 in Bickel et al. [1975]. Also, the fraction of female students is significantly smaller than the fraction of male students. Hence, it is plausible that (latent) selection mechanisms alter the causal model resulting in violating the assumptions of, for instance, absence of a bidirected edge in the causal graph between S and A [Chen et al., 2024]. Since allowing for selection bias enlarges the model class $\mathbb{M}_{\text{cf}+}$, and given that the data satisfies the IV inequalities, we conclude that allowing for selection bias will not change our conclusion.

Acknowledgements

This work was supported by Booking.com. We thank an anonymous reviewer for their feedback which helped us modulate some of our conclusions.

References

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American statistical Association*, 92(439):1171–1176, 1997.

Alexander Abraham Balke. *Probabilistic counterfactuals: semantics, computation, and applications*. PhD thesis, University of California, Los Angeles, 1995. URL https://apps.dtic.mil/sti/tr/pdf/ADA332296.pdf.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Richard A Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Fair risk algorithms. *Annual Review of Statistics and Its Application*, 10(1):165–187, 2023.

Peter J Bickel, Eugene A Hammel, and William J O'Connell. Sex bias in graduate admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975.

Thijs Bol. Gender inequality in cum laude distinctions for PhD students. *Scientific Reports*, 13(1):20267, 2023.

Blai Bonet. Instrumentality tests revisited. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 48–55, 2001.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49 (5):2885–2915, 2021.

Zhihong Cai, Manabu Kuroki, Judea Pearl, and Jin Tian. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64(3):695–701, 2008.

Leihao Chen, Onno Zoeter, and Joris M Mooij. Modeling latent selection with structural causal models. *arXiv preprint arXiv:2401.06925*, 2024.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

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

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Robin J Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

W. Fairley and F. Mosteller. *Statistics and Public Policy*. Addison-Wesley Publishing Company, 1977. ISBN 9780201021851.

Patrick Forré and Joris M Mooij. A Mathematical Introduction to Causality. *Lecture Notes*, 2025. URL https://staff.fnwi.uva.nl/j.m.mooij/articles/causality_lecture_notes_2025.pdf.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331, 2016.

Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019.

Sol Kaufman, Jay S Kaufman, Richard F MacLehose, Sander Greenland, and Charles Poole. Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine*, 24(11):1683–1702, 2005.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 30, 2017.

Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in Artificial Intelligence*, pages 616–626. PMLR, 2020.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference*, 2017.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence*, pages 435–443, 1995.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence, 2001*, pages 411–420, 2001.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009. ISBN 9780521895606.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

Drago Plečko, Elias Bareinboim, et al. Causal fairness analysis: a causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/UCBAdmissions.html`.

Roland Ramsahai. *Causal inference with instruments and other supplementary variables*. PhD thesis, Oxford University, UK, 2008. URL `https://ora.ox.ac.uk/objects/uuid:3ae82165-aef2-4eb7-968d-0ea4523b5b81`.

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

Maresa Schröder, Dennis Frauen, and Stefan Feuerriegel. Causal fairness under unobserved confounding: A neural sensitivity framework. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=DqD59dQP37`.

Linbo Wang, James M Robins, and Thomas S Richardson. On falsification of the binary instrumental variable model. *Biometrika*, 104(1):229–236, 2017.

Elie Wolfe, Robert W Spekkens, and Tobias Fritz. The inflation technique for causal inference with latent variables. *Journal of Causal Inference*, 7(2):20170020, 2019.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3929–3935, 2017.

Revisiting the Berkeley Admissions data: Statistical Tests for Causal Hypotheses (Supplementary Material)

Sourbh Bhadane[1] **Joris M. Mooij**[1] **Philip Boeken**[1] **Onno Zoeter**[2]

[1]Korteweg-de Vries Institute for Mathematics, University of Amsterdam
[2]Booking.com, Amsterdam, The Netherlands

A ADDITIONAL PRELIMINARIES

Definition 19 (Twin SCM)**.** *Let $M = (V, W, \mathcal{X}, P, f)$ be an SCM. The twinning operation maps M to the **twin SCM***

$$M^{twin} \triangleq \left(V \cup V', W, \mathcal{X}_V \times \mathcal{X}_{V'} \times \mathcal{X}_W, P, \tilde{f}\right)$$

where $V' = \{v' : v \in V\}$ is a disjoint copy of V and the causal mechanism $\tilde{f} : \mathcal{X}_V \times \mathcal{X}_{V'} \times \mathcal{X}_W \mapsto \mathcal{X}_V \times \mathcal{X}_{V'}$ is given by $\tilde{f}(x_V, x_{V'}, x_W) = (f(x_V, x_W), f(x_{V'}, x_W))$.

Definition 20 (Solution function)**.** *Let $M = (V, W, \mathcal{X}, f, P)$ be an acyclic SCM and $C \subseteq V$. A **solution function** of M with respect to C is a measurable mapping $g_C : \mathcal{X}_{V\setminus C} \times \mathcal{X}_W \mapsto \mathcal{X}_C$ that satisfies the structural equations for C, i.e., for all $x_{V\setminus C} \in \mathcal{X}_{V\setminus C}$, $P(X_W)$-a.a $x_W \in \mathcal{X}_W$,*

$$g_C\big(x_{V\setminus C}, x_W\big) = f_C\big(x_{V\setminus C}, g_C\big(x_{V\setminus C}, x_W\big), x_W\big).$$

Definition 21 (Markov kernels)**.** *Let $\mathcal{T}$ and $\mathcal{W}$ be measurable spaces. A Markov kernel is defined as a measurable map $K : \mathcal{T} \mapsto \mathcal{P}(\mathcal{W})$ where $\mathcal{P}(\mathcal{W})$ is defined as the space of probability measures on $\mathcal{W}$.*

B IV INEQUALITIES EXPRESSED AS MARKOV KERNELS

For (2) to be well defined, we required that $P_M(Z = z) > 0$ for all z for any $M \in \mathbb{M}_{\text{IV}+}$ (see Definition (7)). In this section, we relax this requirement by noting that, in fact, IV inequalities are more appropriately expressed in terms of $P_M(X, Y \mid \text{do}(Z))$.

Lemma 22. *Let $\mathbb{M}_{IV} \triangleq \{M : G(M)$ is a subgraph of Figure $1\}$. For any $M \in \mathbb{M}_{IV}$,*

$$\max_x \sum_y \max_z P_M(X = x, Y = y \mid do(Z = z)) \leq 1. \tag{3}$$

Proof. Since

$$\begin{aligned} P_M(X = x, Y = y \mid \text{do}(Z = z)) &= P_M(f_X(z, U) = x, f_Y(x, U) = y) \\ &\leq P(f_Y(x, U) = y) = P_M(Y = y \mid \text{do}(X = x)), \end{aligned}$$

$$\max_x \sum_y \max_z P_M(X = x, Y = y \mid \text{do}(Z = z)) \leq \max_x \sum_y P_M(Y = y \mid \text{do}(X = x)) = 1. \tag{4}$$

□

Note that (3) is defined even when $\exists z \in \mathcal{Z}$ such that $P(Z = z) = 0$. In contrast, positivity must be assumed in (2) for the terms to be well-defined. Further, if positivity is assumed, then $\mathbb{M}_{\text{IV}} = \mathbb{M}_{\text{IV}+}$ and (3) is identical to (2).

C PROOFS FOR SECTION 3

C.1 NESTED FAIRNESS NOTIONS: WITHOUT CONFOUNDING

Lemma 12.

$$H^0_{\textit{no-cf-graph}} = H^0_{\textit{no-cf-ctrf}} \subset H^0_{\textit{no-cf-inter}} \subset H^0_{\textit{no-cf-obs}}.$$

If for all s, d, $P_M(s,d) > 0$, then in addition, we have $H^0_{\textit{no-cf-inter}} = H^0_{\textit{no-cf-obs}}$.

Proof. $\boldsymbol{H^0_{\text{no-cf-graph}} = H^0_{\text{no-cf-ctrf}}}$: We first show that $H^0_{\text{no-cf-graph}} \subseteq H^0_{\text{no-cf-ctrf}}$. $M \in H^0_{\text{no-cf-graph}}$ implies $\forall d$, $f_A(s,d,U_A)$ is constant in s P-a.s. Therefore, for all d, s,

$$P_M(f_A(s,d,U_A) = f_A(S,d,U_A)) = 1. \tag{5}$$

Therefore, $M \in H^0_{\text{no-cf-ctrf}}$. For the converse, $M \in H^0_{\text{no-cf-ctrf}}$ implies (5). For $s \neq s'$, and all d,

$$P_M(f_A(s,d,U_A) = f_A(S,d,U_A)) = P_M(f_A(s,d,U_A) = f_A(s',d,U_A))P_M(S = s') + P_M(S = s).$$

From (5) if $P_M(S = s') > 0$, we conclude $P_M(f_A(s,d,U_A) = f_A(s',d,U_A)) = 1$. If $P_M(S = s') = 0$, since (5) holds for s', i.e., for all d, s', $P_M(f_A(s',d,U_A) = f_A(S,d,U_A)) = 1$, we have

$$P_M(f_A(s',d,U_A) = f_A(s,d,U_A)) = 1.$$

Therefore, $M \in H^0_{\text{no-cf-graph}}$.

$\boldsymbol{H^0_{\text{no-cf-ctrf}} \subset H^0_{\text{no-cf-inter}}}$: For $M \in H^0_{\text{no-cf-ctrf}}$, (5) implies $P_M(f_A(s,d,U_A)) = P_M(f_A(S,d,U_A))$ for all d, s, implying $M \in H^0_{\text{no-cf-inter}}$. Since Example 14 belongs to $H^0_{\text{no-cf-inter}} \backslash H^0_{\text{no-cf-ctrf}}$, the inclusion is strict.

$\boldsymbol{H^0_{\text{no-cf-inter}} \subset H^0_{\text{no-cf-obs}}}$: For $M \in H^0_{\text{no-cf-inter}}$, $P_M(A = 1 \mid \text{do}(D = d), \text{do}(S = s))$ is constant in s. Consider a pair s, d, such that, for $M \in H^0_{\text{no-cf-inter}}$, $P_M(s,d) > 0$. Then

$$P_M(A = 1 \mid \text{do}(D = d), \text{do}(S = s)) = P_M(A = 1 \mid D = d, S = s). \tag{6}$$

Note that, if, for $s' \neq s$, $P_M(s',d) = 0$, then $P_M(A = 1 \mid D = d, S = s) = P_M(A = 1 \mid D = d)$. If instead, $P_M(s',d) > 0$, then from (6) for $S = s'$, we have that

$$P_M(A = 1 \mid D = d, S = s) = P_M(A = 1 \mid D = d, S = s') = P_M(A = 1 \mid D = d).$$

Therefore, we conclude that $M \in H^0_{\text{no-cf-obs}}$ implying $H^0_{\text{no-cf-inter}} \subseteq H^0_{\text{no-cf-obs}}$. Since the SCM in Example 15 lies in $H^0_{\text{no-cf-obs}} \backslash H^0_{\text{no-cf-inter}}$, $H^0_{\text{no-cf-inter}} \subset H^0_{\text{no-cf-obs}}$.

If for all s, d, $P_M(s,d) > 0$, then for $M \in H^0_{\text{no-cf-obs}}$,

$$P_M(A = 1 \mid D = d, S = s) = P_M(A = 1 \mid \text{do}(D = d), \text{do}(S = s))$$

is constant in s and equal to $P_M(A = 1 \mid \text{do}(D = d))$. This implies $M \in H^0_{\text{no-cf-inter}}$. Therefore, if for all s, d, $P_M(s,d) > 0$, then $H^0_{\text{no-cf-inter}} = H^0_{\text{no-cf-obs}}$.

□

C.2 EQUIVALENCE OF TESTS WITHOUT CONFOUNDING

Theorem 13. *Let*

$$\begin{aligned}
\mathcal{P}_{\textit{no-cf-graph}} &\triangleq \left\{P_M(D,A,S) : M \in H^0_{\textit{no-cf-graph}}\right\}, \\
\mathcal{P}_{\textit{no-cf-ctrf}} &\triangleq \left\{P_M(D,A,S) : M \in H^0_{\textit{no-cf-ctrf}}\right\}, \\
\mathcal{P}_{\textit{no-cf-inter}} &\triangleq \left\{P_M(D,A,S) : M \in H^0_{\textit{no-cf-inter}}\right\}, \\
\mathcal{P}_{\textit{no-cf-obs}} &\triangleq \left\{P_M(D,A,S) : M \in H^0_{\textit{no-cf-obs}}\right\}.
\end{aligned}$$

Then $\mathcal{P}_{\textit{no-cf-graph}} = \mathcal{P}_{\textit{no-cf-ctrf}} = \mathcal{P}_{\textit{no-cf-inter}} = \mathcal{P}_{\textit{no-cf-obs}}$.

Proof. From Lemma 12, $\mathcal{P}_{\text{no-cf-graph}} = \mathcal{P}_{\text{no-cf-ctrf}} \subseteq \mathcal{P}_{\text{no-cf-inter}} \subseteq \mathcal{P}_{\text{no-cf-obs}}$. Therefore, it suffices to prove that $\mathcal{P}_{\text{no-cf-graph}} = \mathcal{P}_{\text{no-cf-obs}}$. For every $P_M \in \mathcal{P}_{\text{no-cf-obs}}$,

$$P_M(A, S, D) = P_M(S) \otimes P_M(D \mid S) \otimes P_M(A \mid D).$$

Hence, $\exists \tilde{M} \in H^0_{\text{no-cf-graph}}$ such that $P_M(A, D, S) = P_{\tilde{M}}(A, D, S)$. □

D PROOFS FOR SECTION 4

D.1 SHARPNESS OF IV INEQUALITIES

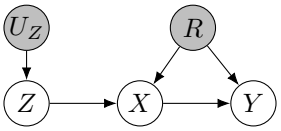


Figure 4: Response-function parameterization of $M \in \mathbb{M}_{\text{IV}+}$

Theorem 17. *Let X, Y, Z be discrete random variables defined on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively, with $|\mathcal{X}| = n \geq 2, |\mathcal{Y}| = 2, |\mathcal{Z}| = 2$. Define $\mathcal{P}_{\mathbb{M}_{IV+}} \triangleq \{P_M(X, Y, Z) : M \in \mathbb{M}_{IV+}\}$ and the set of joint distributions that satisfy the IV inequalities as*

$$\begin{aligned}\mathcal{P}_{IV+} \triangleq \{&P(X, Y, Z) : P(X, Y \mid Z) \textit{ satisfies } (2) \\ &\textit{and } \forall z, P(Z = z) > 0\}.\end{aligned}$$

Then $\mathcal{P}_{IV+} = \mathcal{P}_{\mathbb{M}_{IV+}}$.

Proof. We prove a more general statement that includes Theorem 17 as a special case.

Lemma 23. *Let X, Y, Z be discrete random variables defined on $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ respectively, with $|\mathcal{X}| = n \geq 2, |\mathcal{Y}| = 2, |\mathcal{Z}| = 2$. Define $\mathcal{K}_{IV} \triangleq \{K(X, Y \mid Z) : K(X, Y \mid Z)$ satisfies (3)$\}$. Define $\mathcal{K}_{\mathbb{M}_{IV}} \triangleq \{P_M(X, Y \mid do(Z)) : M \in \mathbb{M}_{IV}\}$. Then $\mathcal{K}_{IV} = \mathcal{K}_{\mathbb{M}_{IV}}$.*

Note that $\mathcal{P}_{\text{IV}+} = \{P(Z) : \forall z, P(Z = z) > 0\} \otimes \mathcal{K}_{\text{IV}}$ since assuming positivity, (2) is identical to (3). Further, $\mathcal{P}_{\mathbb{M}_{\text{IV}+}} = \{P_M(Z) : M \in \mathbb{M}_{\text{IV}+}\} \otimes \mathcal{K}_{\mathbb{M}_{\text{IV}}}$ since assuming positivity, $\mathbb{M}_{\text{IV}} = \mathbb{M}_{\text{IV}+}$ and $P_M(X, Y \mid \text{do}(Z)) = P_M(X, Y \mid Z)$ for $M \in \mathbb{M}_{\text{IV}+}$. Since the first factors are identical, Theorem 17 follows from Lemma 23. □

Proof of Lemma 23. For $M \in \mathbb{M}_{\text{IV}}$, the response-function parameterization yields a counterfactually equivalent SCM [Forré and Mooij, 2025, Section 8.4] $\tilde{M} = (V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{f}, \tilde{P})$, where $V = \{Z, X, Y\}, \tilde{W} = \{R, U_Z\}, \tilde{\mathcal{X}} = \mathcal{X}_V \times \mathcal{X}_{\tilde{W}}, \tilde{f} = \left(\tilde{f}_Z, \tilde{f}_X, \tilde{f}_Y\right)$ where we define $\mathcal{X}_R, \tilde{f}, \tilde{P}$ through the function $\Phi : \mathcal{X}_W \mapsto \mathcal{X}_{\tilde{W}}$ where

$$\begin{aligned}\mathcal{X}_R &\triangleq \mathcal{X}^{\mathcal{Z}} \times \mathcal{Y}^{\mathcal{X}}, \\ \forall u_Z, u_X, u_Y, u, \Phi(u_Z, u_X, u_Y, u) &\triangleq ((z \mapsto f_X(z, u, u_X), x \mapsto f_Y(x, u, u_Y)), u_Z), \\ \forall u_Z, \tilde{f}_Z(u_Z) &\triangleq f_Z(u_Z), \\ \forall r, z, \tilde{f}_X(r, z) &\triangleq r_1(z), \\ \forall r, x, \tilde{f}_Y(r, x) &\triangleq r_2(x),\end{aligned}$$

where $r = (r_1, r_2)$ and $\tilde{P}$ is the push-forward distribution $\Phi_*(P)$. Note that $\mathcal{X}_R$ is a discrete space, R a discrete random variable, and $\tilde{P}(R)$ a discrete distribution over $\mathcal{X}_R$.

Under the response-function parameterization, only $\tilde{P}(R)$ is a parameter. We will consider $\tilde{P}(R)$ to be an element of $\mathbb{R}^{n_X^{n_Z} n_Y^{n_X}}$ where $\#\mathcal{X} = n_X, \#\mathcal{Z} = n_Z, \#\mathcal{Y} = n_Y$ and

$$\mathcal{K}_{\tilde{\mathbb{M}}_{\text{IV}+}} \triangleq \{P_{\tilde{M}}(X, Y \mid \text{do}(Z)) : M \in \mathbb{M}_{\text{IV}}\}$$

to be a subset of $\mathbb{R}^{n_X n_Y n_Z}$. Note that because of the counterfactual equivalence of the response-function parameterization, which in turn implies interventional equivalence, $\mathcal{K}_{\tilde{\mathbb{M}}_{\text{IV},r}} = \mathcal{K}_{\mathbb{M}_{\text{IV}}}$. From Lemma 22, $\mathcal{K}_{\mathbb{M}_{\text{IV}}} \subseteq \mathcal{K}_{\text{IV}}$.

To show the converse, we show that each extreme point of $\mathcal{K}_{\text{IV}}$ is obtained by a point in $\mathcal{K}_{\mathbb{M}_{\text{IV}}}$. We enumerate all extreme points of $\mathcal{K}_{\text{IV}}$ in Lemma 24. We show that each such extreme point is obtained by the following response-function. Choose $x, x' \in \mathcal{X}, y, y' \in \mathcal{Y}$ with $y = y'$ if $x = x'$. Then any response function satisfying

$$r_1(z) = \begin{cases} x & z = 0 \\ x' & z = 1 \end{cases}$$
$$r_2(\tilde{x}) = \begin{cases} y & \tilde{x} = x \\ y' & \tilde{x} = x' \\ \text{arbitrary} & \text{otherwise} \end{cases}$$

gives all extreme points of $\mathcal{K}_{\text{IV}}$. Therefore, $\mathcal{K}_{\mathbb{M}_{\text{IV}}} \supseteq \mathcal{K}_{\text{IV}}$, implying $\mathcal{K}_{\mathbb{M}_{\text{IV}}} = \mathcal{K}_{\text{IV}}$.

□

Lemma 24. *Consider the real vector space $\mathbb{R}^{n_X n_Y n_Z}$ spanned by the canonical basis vectors*

$$\left\{\delta_{x,y|z} : x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}\right\}$$

where $\delta_{x,y|z}$ denotes a unit vector of length $n_X n_Y n_Z$ where all entries except the one at (x, y, z) are zero. For $n_Y = n_Z = 2, n_X = n \geq 2$, $\mathcal{K}_{IV}$ considered as a subset of this vector space is a polyhedral set with extreme points

$$\mathbb{E} = \left\{\delta_{x,y|0} + \delta_{x',y'|1} : x, x' \in \mathcal{X}; y, y' \in \mathcal{Y} : x \neq x'\right\} \cup \left\{\delta_{x,y|0} + \delta_{x,y|1} : x \in \mathcal{X}, y \in \mathcal{Y}\right\}.$$

Proof. Consider $\mathcal{K}_{\text{IV}}$ to be a subset of $\mathbb{R}^{n_X n_Y n_Z}$ where each element of $\mathcal{K}_{\text{IV}}$ is represented as $\{K(X = x, Y = y \mid Z = z)\}_{x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}}$ and satisfies

$$\begin{aligned} \forall x \in \mathcal{X} : K(X = x, Y = 0 | Z = 0) + K(X = x, Y = 1 | Z = 1) \leq 1, \\ \forall x \in \mathcal{X} : K(X = x, Y = 0 | Z = 1) + K(X = x, Y = 1 | Z = 0) \leq 1, \\ \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall z \in \mathcal{Z} : K(X = x, Y = y | Z = z) \geq 0, \\ \forall z \in \mathcal{Z} : \sum_{x,y} K(X = x, Y = y | Z = z) = 1. \end{aligned}$$

An elementary result in convex geometry states that a point is an extreme point of a polyhedral set (defined by a set of linear (in)equality constraints) in $\mathbb{R}^m$ vector space if and only if it is a feasible point and there exists a subset A of m constraints that are active and linearly independent. For the case at hand, for a point to be an extreme point of $\mathcal{K}_{\text{IV}}$, it has to satisfy the above $6n + 2$ constraints (feasibility) and additionally, a subset of $n_X n_Y n_Z = 4n$ of them must be active and linearly independent. The two normalization constraints are equality constraints and hence must be active at any feasible point.

We first show that all points in $\mathbb{E}$ are extreme points of $\mathcal{K}_{\text{IV}}$. First, choose $x, x' \in \mathcal{X}$ with $x \neq x'$ and choose $y, y' \in \mathcal{Y}$. It can be verified that $\delta_{x,y|0} + \delta_{x',y'|1}$ satisfies the IV inequalities with 2 out of the $2n$ IV inequalities being active. Further, $4n - 2$ non-negativity constraints and both normalization constraints are active. For the subset A of active constraints we can take e.g. both active IV inequalities together with the $4n_X - 2$ active nonnegativity constraints; one can check that these are linearly independent. Second, choose $x \in \mathcal{X}, y \in \mathcal{Y}$. It can be verified that $\delta_{x,y|0} + \delta_{x,y|1}$ satisfies the IV inequalities with 2 out of the $2n$ IV inequalities being active. Further, $4n - 2$ non-negativity constraints and both normalization constraints are active. For the subset A of active constraints we can take e.g. both active IV inequalities together with the $4n_X - 2$ active nonnegativity constraints; one can check that these are linearly independent.

Finally, we check whether $\mathbb{E}$ exhausts all extreme points. Pick a feasible point $b \in \mathbb{R}^{n_X n_Y n_Z}$. We will refer to the indices i with $b_i \neq 0$ ($b_i = 0$) as the "non-zero (zero) entries of b", and to the indices j in active constraint $a^T b = 1$ with $a_j \neq 0$ as the "active entries of the constraint". For a set of entries of b and a set of active entries of one or more active constraints, we refer to their intersection as their "overlap". We also call the b_i corresponding to $K(X, Y | Z = z)$ a "stratum".

Suppose b is an extreme point of $\mathcal{K}_{\text{IV}}$. Then at least $4n - 2$ of the $6n$ inequality constraints must be active. Hence if exactly r IV inequalities are active at b, then we need at least $4n - 2 - r$ active nonnegativity constraints at b, or in other words,

b must have at least $4n-2-r$ zero entries. Because both strata of b need to be normalized, b can have at most $4n-2$ zero entries. The number of zero entries in b that can overlap with the active entries of the active IV inequalities is upper bounded by r; indeed, otherwise there would be an active IV inequality for which both its active entries are zero entries of b, contradicting that this IV inequality is active.

We will show that the only possibilities for b are the extreme points that we already identified, proceeding case by case.

1. $r > 2$. This yields a contradiction with the normalization constraints. Indeed, each entry of b appears in exactly one IV inequality, and each IV inequality contains exactly two active entries (one for each stratum). Hence, the sum of those entries of b that correspond with active entries of active IV constraints must be exactly r. However, by the normalization constraints, the sum over *all* entries of b must be 2. This implies that $r \leq 2$.
2. $r = 2$. Then b must contain at least $4n-4$ zero entries, of which at most two can overlap with the active entries of the active IV inequalities.
   (a) If there are no such overlaps, then all other entries of b must zero entries. This means that we need the $4n-4$ nonnegativity constraints corresponding to those other entries in the subset A, as well as the two normalization and inequality constraints; no other active constraints exist that could be added to A. But then the constraints in A would not be linearly independent. Indeed, the following linear dependence between the active constraints is obtained: adding the coefficient vectors of the two active IV inequalities and those of all active non-negativity constraints yields the same result as adding the coefficient vectors of the two normalization constraints. So we arrive at a contradiction.
   (b) If there is at least one overlap, then $b_j = 1$ with j the other active entry in the active IV inequality with the overlap. This implies $2n-1$ zeroes in the stratum of j. Consider now the other active IV inequality. Since $b_j = 1$, the active entry corresponding to that stratum must be a zero entry of b. Hence, $b_k = 1$ for k the other active entry in this IV inequality. This means that b must be of the form $\delta_{x_1,x_2|0} + \delta_{x_1',x_2'|1}$, with the two non-zero entries corresponding to active entries of two different IV inequalities, and hence we recover the extreme points already identified.
3. $r = 1$. Then b must contain at least $4n-3$ zero entries, of which at most one can overlap with the active entries of active IV inequalities.

   Because of the normalization constraints, we need at least one non-zero entry in both strata. This means that there must be a stratum x_3 with exactly one non-zero entry, and then b must be of the form $b = \delta_{x_1,x_2|x_3} + \gamma\delta_{x_1',x_2'|x_3'} + (1-\gamma)\delta_{x_1'',x_2''|x_3'}$ with $\gamma \in (0,1]$ and $x_3' \neq x_3$. This means that the active IV inequality must be the one that has active entry $(x_1, x_2|x_3)$. Its other active entry is then $(x_1, 1-x_2|x_3')$, and this must be an overlapping zero entry of b.

   If $\gamma = 1$ then this would also activate the IV inequality that contains active entry $(x_1', x_2'|x_3')$. Since the only active IV inequality must also contain active entry $(x_1, x_2|x_3)$, this gives a contradiction, as the two coefficients of b at the active entries sum to 2 instead of 1. Hence, $\gamma \in (0,1)$, and b contains exactly $4n-3$ zero entries.

   So the active constraints consist of 1 active IV inequality, 2 active normalization constraints and $4n-3$ active nonnegativity constraints. However, these are not linearly independent. Indeed: subtracting the coefficient vectors of the $2n-1$ nonnegativity constraints in stratum x_3 from the coefficient vector of the normalization constraint of that stratum, and adding the coefficient vector of the nonnegativity constraint corresponding to $(x_1, 1-x_2|x_3')$ gives a vector that is identical to the coefficient vector of the active IV inequality.

   So we have arrived at a contradiction.
4. $r = 0$. Then b must contain at least $4n-2$ zero entries. Because of the normalization constraints, b needs one non-zero entry in both strata, and its value must be 1. This will activate at least one IV inequality. Contradiction. □

D.2 NESTED FAIRNESS NOTIONS: WITH CONFOUNDING

Proposition 25.

$$H^0_{\textit{cf-graph}} = H^0_{\textit{cf-ctrf}} \subset H^0_{\textit{cf-inter}},$$
$$H^0_{\textit{cf-graph+}} = H^0_{\textit{cf-ctrf+}} \subset H^0_{\textit{cf-inter+}}.$$

Proof. If $M \in \mathbb{M}_{\text{cf}}$, then

$$A^{\text{do}(S=s,D=d)} = f_A(s, d, U_A, U), \qquad A^{\text{do}(D=d)} = f_A(S, d, U_A, U). \tag{7}$$

For $M \in H^0_{\text{cf-graph}}$, since S is not a parent of A, for all d, $f_A(s, d, U_A, U)$ is constant in s P-a.s. This implies that for all d, s, s',

$$P_M(f_A(s, d, U_A, U) = f_A(s', d, U_A, U)) = 1$$

Therefore, for all s, d,

$$P_M(f_A(s, d, U_A, U) = f_A(S, d, U_A, U)) = 1, \tag{8}$$

implying that $M \in H^0_{\text{cf-ctrf}}$. For the converse, $M \in H^0_{\text{cf-ctrf}}$ implies (8). For $s \neq s'$, and all d,

$$\begin{aligned} &P_M(f_A(s, d, U_A, U) = f_A(S, d, U_A, U)) \\ &= P_M(f_A(s, d, U_A, U) = f_A(s', d, U_A, U))P_M(S = s') + P_M(S = s). \end{aligned}$$

If $P_M(S = s') > 0$, we conclude $P_M(f_A(s, d, U_A, U) = f_A(s', d, U_A, U)) = 1$. If $P_M(S = s') = 0$, since (8) holds for s', we have $P_M(f_A(s, d, U_A, U) = f_A(s', d, U_A, U)) = 1$. Therefore, $M \in H^0_{\text{cf-graph}}$. Therefore, $H^0_{\text{cf-graph}} = H^0_{\text{cf-ctrf}}$. Further, $H^0_{\text{cf-graph+}} = H^0_{\text{cf-ctrf+}}$.

We now prove that $H^0_{\text{cf-ctrf}} \subseteq H^0_{\text{cf-inter}}$. For $M \in H^0_{\text{cf-ctrf}}$, (8) holds. Therefore, for all s, d,

$$P_M(f_A(s, d, U_A, U)) = P_M(f_A(S, d, U_A, U)). \tag{9}$$

Therefore, $H^0_{\text{cf-ctrf}} \subseteq H^0_{\text{cf-inter}}$ and subsequently $H^0_{\text{cf-ctrf+}} \subseteq H^0_{\text{cf-inter+}}$. Note that the Example 14 lies in $\mathbb{M}_{\text{cf}}$ (and in $\mathbb{M}_{\text{cf+}}$ since $P_M(S = s) > 0$ for all s) for any U that is independent of U_S, U_D, U_A. Further, $P_M(A = 1 \mid \text{do}(S = s), \text{do}(D = d)) = 0.5 = P_M(A = 1 \mid \text{do}(D = d))$; however, S is a parent of A. Therefore, $H^0_{\text{cf-ctrf}} \subset H^0_{\text{cf-inter}}$ and $H^0_{\text{cf-ctrf+}} \subset H^0_{\text{cf-inter+}}$.

□

D.3 EQUIVALENCE OF STATISTICAL TESTS

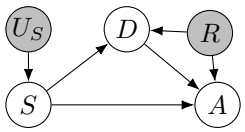


Figure 5: Response-function parameterization of $M \in \mathbb{M}_{\text{cf}}$

Theorem 18. *Let*

$$\begin{aligned} \mathcal{P}_{\textit{cf-graph}} &\triangleq \left\{P_M(D, A, S) : M \in H^0_{\textit{cf-graph+}}\right\}, \\ \mathcal{P}_{\textit{cf-inter}} &\triangleq \left\{P_M(D, A, S) : M \in H^0_{\textit{cf-inter+}}\right\}, \\ \mathcal{P}_{\textit{cf-ctrf}} &\triangleq \left\{P_M(D, A, S) : M \in H^0_{\textit{cf-ctrf+}}\right\}. \end{aligned}$$

Then $\mathcal{P}_{\textit{cf-inter}} = \mathcal{P}_{\textit{cf-ctrf}} = \mathcal{P}_{\textit{cf-graph}} = \mathcal{P}_{\textit{IV+}}$, where $\mathcal{P}_{\textit{IV+}}$ is defined in Theorem 17.

Proof. Like in Section D.1, we prove a more general statement, Lemma 26, that includes Theorem 18 as a special case. We define analogues of $\mathbb{M}_{\text{cf+}}, H^0_{\text{cf-notion+}}$ that remove the positivity assumption, $P_M(S = s) > 0$ for all s, as $\mathbb{M}_{\text{cf}}, H^0_{\text{cf-notion}}$, respectively (where we use 'notion' as a placeholder for 'graph', 'ctrf' and 'inter').

Lemma 26. *Let*

$$\begin{aligned} \mathcal{K}_{\textit{cf-graph}} &\triangleq \left\{P_M(D, A \mid do(S)) : M \in H^0_{\textit{cf-graph}}\right\}, \\ \mathcal{K}_{\textit{cf-inter}} &\triangleq \left\{P_M(D, A \mid do(S)) : M \in H^0_{\textit{cf-inter}}\right\}, \\ \mathcal{K}_{\textit{cf-ctrf}} &\triangleq \left\{P_M(D, A \mid do(S)) : M \in H^0_{\textit{cf-ctrf}}\right\}. \end{aligned}$$

Then $\mathcal{K}_{\textit{cf-inter}} = \mathcal{K}_{\textit{cf-ctrf}} = \mathcal{K}_{\textit{cf-graph}} = \mathcal{K}_{\textit{IV}}$, where $\mathcal{K}_{\textit{IV}}$ is defined in Lemma 23.

Note that $\mathcal{P}_{\text{IV+}} = \{P(Z) : \forall z, P(Z = z) > 0\} \otimes \mathcal{K}_{\text{IV}}$ since assuming positivity, (2) is identical to (3). Further, $\mathcal{P}_{\text{cf-notion}} = \left\{P_M(S) : M \in H^0_{\text{cf-notion+}}\right\} \otimes \mathcal{K}_{\text{cf-notion}}$ since assuming positivity, $\mathbb{M}_{\text{cf+}} = \mathbb{M}_{\text{cf}}$ and $P_M(D, A \mid \text{do}(S)) = P_M(D, A \mid S)$ for $M \in \mathbb{M}_{\text{cf+}}$. Since the first factors are identical, Theorem 18 follows from Lemma 26. □

Proof of Lemma 26. For $M \in \mathbb{M}_{\text{cf}}$, the response-function parameterization yields a counterfactually equivalent SCM, $\tilde{M}$ represented by the tuple $(V, \tilde{W}, \tilde{\mathcal{X}}, \tilde{f}, \tilde{P})$, where $V = \{S, D, A\}, \tilde{W} = \{R, U_S\}, \tilde{\mathcal{X}} = \mathcal{X}_V \times \mathcal{X}_{\tilde{W}}, \tilde{f} = \left(\tilde{f}_S, \tilde{f}_D, \tilde{f}_A\right)$ where we define $\mathcal{X}_R, \tilde{f}, \tilde{P}$ through the function $\Phi : \mathcal{X}_W \mapsto \mathcal{X}_{\tilde{W}}$ where

$$\begin{aligned}
\mathcal{X}_R &\triangleq \mathcal{X}_D^{\mathcal{X}_S} \times \mathcal{X}_A^{\mathcal{X}_S \times \mathcal{X}_D}, \\
\forall u_S, u_D, u_A, u, \Phi(u_S, u_D, u_A, u) &\triangleq ((s \mapsto f_D(s, u, u_D), (s, d) \mapsto f_A(s, d, u, u_A)), u_S), \\
\forall u_S, \tilde{f}_S(u_S) &\triangleq f_S(u_S), \\
\forall s, \tilde{f}_D(r, s) &\triangleq r_1(s), \\
\forall s, d, \tilde{f}_A(r, s, d) &\triangleq r_2(s, d),
\end{aligned}$$

where $r = (r_1, r_2)$ and $\tilde{P}$ is the push-forward distribution $\Phi_*(P)$. Note that $\mathcal{X}_R$ is a discrete space, R a discrete random variable, and $\tilde{P}(R)$ a discrete distribution over $\mathcal{X}_R$. Under the response-function parameterization, only $\tilde{P}(R)$ is a parameter and we will abuse notation and denote it as $\tilde{P}$ henceforth. Therefore, we can represent $H^0_{\text{cf-graph}}$ in the parameter space as

$$\bar{H}^0_{\text{graph}} \triangleq \left\{ \tilde{P} \in \triangle(\mathcal{X}_R) : \tilde{P}(r_1, r_2) \neq 0 \text{ implies } \forall d, r_2(0, d) = r_2(1, d) \right\}. \tag{10}$$

To express $H^0_{\text{cf-inter}}$, we express the interventional Markov kernels $P_{\tilde{M}}(A \mid \text{do}(S), \text{do}(D))$ in terms of $\tilde{P}$. Since counterfactual equivalence implies interventional equivalence, for all s, d, $P_M(A = 1 \mid \text{do}(S = s), \text{do}(D = d)) = P_{\tilde{M}}(A = 1 \mid \text{do}(S = s), \text{do}(D = d))$, where

$$P_{\tilde{M}}(A = 1 \mid \text{do}(S = s), \text{do}(D = d)) = \sum_{(r_1, r_2) \in \mathcal{X}_R} \mathbf{1}[r_2(s, d) = 1] \tilde{P}(r_1, r_2), \tag{11}$$

$$P_{\tilde{M}}(A = 1 \mid \text{do}(D = d)) = \sum_{s^*} \sum_{(r_1, r_2) \in \mathcal{X}_R} \mathbf{1}[r_2(s^*, d) = 1] \tilde{P}(r_1, r_2) P_{\tilde{M}}(s^*), \tag{12}$$

Subtracting (11) from (12) we get

$$\begin{aligned}
&P_{\tilde{M}}(A = 1 \mid \text{do}(S = s), \text{do}(D = d)) - P_{\tilde{M}}(A = 1 \mid \text{do}(D = d)) \\
&= \left(\sum_{(r_1, r_2) \in \mathcal{X}_R} (\mathbf{1}[r_2(0, d) = 1] - \mathbf{1}[r_2(1, d) = 1]) \tilde{P}(r_1, r_2) \right) P_{\tilde{M}}(s') = 0
\end{aligned} \tag{13}$$

for $M \in H^0_{\text{cf-inter}}$, where $s' \neq s$. Similarly,

$$\begin{aligned}
&P_{\tilde{M}}(A = 1 \mid \text{do}(S = s'), \text{do}(D = d)) - P_{\tilde{M}}(A = 1 \mid \text{do}(D = d)) \\
&= \left(\sum_{(r_1, r_2) \in \mathcal{X}_R} (\mathbf{1}[r_2(0, d) = 1] - \mathbf{1}[r_2(1, d) = 1]) \tilde{P}(r_1, r_2) \right) P_{\tilde{M}}(s) = 0.
\end{aligned} \tag{14}$$

Since both (13) and (14) hold, the response-function parameterized analogue of $H^0_{\text{cf-inter}}$ is

$$\bar{H}^0_{\text{inter}} \triangleq \left\{ \tilde{P} \in \triangle(\mathcal{X}_R) : \forall d, \sum_{(r_1, r_2) \in \mathcal{X}_R} (\mathbf{1}[r_2(0, d) = 1] - \mathbf{1}[r_2(1, d) = 1]) \tilde{P}(r_1, r_2) = 0 \right\}. \tag{15}$$

Note that both $\bar{H}^0_{\text{graph}}$ and $\bar{H}^0_{\text{inter}}$ are polyhedra in $\triangle(\mathcal{X}_R)$. Further, $\bar{H}^0_{\text{graph}} \subseteq \bar{H}^0_{\text{inter}}$. While, $\bar{H}^0_{\text{graph}}, \bar{H}^0_{\text{inter}}$ are collections of distributions, we will also refer to them as collection of response-function-parameterized SCMs.

From interventional equivalence (which follows as a result of counterfactual equivalence) of the response-function-parameterization, we have

$$\begin{aligned}
\mathcal{K}_{\text{cf-graph}} &= \left\{ P_{\tilde{M}}(D, A \mid \text{do}(S)) : \tilde{M} \in \bar{H}^0_{\text{graph}} \right\} \\
\mathcal{K}_{\text{cf-inter}} &= \left\{ P_{\tilde{M}}(D, A \mid \text{do}(S)) : \tilde{M} \in \bar{H}^0_{\text{inter}} \right\}.
\end{aligned}$$

We now show that $\mathcal{K}_{\text{cf-inter}} = \mathcal{K}_{\text{cf-graph}} = \mathcal{K}_{\text{IV}}$. First, notice that $\mathcal{K}_{\text{cf-inter}} \supseteq \mathcal{K}_{\text{cf-graph}}$ since $\bar{H}^0_{\text{inter}} \supseteq \bar{H}^0_{\text{graph}}$. We first show that $\mathcal{K}_{\text{cf-inter}} \subseteq \mathcal{K}_{\text{IV}}$ and then $\mathcal{K}_{\text{cf-graph}} = \mathcal{K}_{\text{IV}}$ which concludes the argument.

$\boldsymbol{\mathcal{K}_{\text{cf-inter}} \subseteq \mathcal{K}_{\text{IV}}}$: The solution function of the response-function parameterized SCM, $g_{A,D} : \mathcal{X}_S \times \mathcal{X}_R \mapsto \mathcal{X}_A \times \mathcal{X}_D$ induces a mapping from $\triangle(\mathcal{X}_R)$ which can be considered as a subset of $\mathbb{R}^{\#\mathcal{X}_R}$ to the set of Markov kernels $P_{\tilde{M}}(D, A \mid \text{do}(S))$ which can be considered to be a subset of $\mathbb{R}^{\#(\mathcal{X}_A)\times\#(\mathcal{X}_D)\times\#(\mathcal{X}_S)}$. The condition in (15) implies that for all d,

$$\sum_{r:r_2(0,d)=1} \tilde{P}(r) = \sum_{r:r_2(1,d)=1} \tilde{P}(r). \tag{16}$$

Since, $\sum_r \tilde{P}(r) = 1$,

$$\sum_{r:r_2(0,d)=0} \tilde{P}(r) = \sum_{r:r_2(1,d)=0} \tilde{P}(r). \tag{17}$$

Denote $P_{\tilde{M}}(D = d, A = a \mid \text{do}(S = s))$ by $P_{\tilde{M}}(d, a||s)$. For $P_{\tilde{M}}(d, a||s) \in \mathcal{K}_{\text{cf-inter}}$,

$$P_{\tilde{M}}(d, a||s) = \sum_{r:r_1(s)=d, r_2(s,d)=a} \tilde{P}(r).$$

Therefore, from (16),

$$\begin{aligned}
\sum_{r:r_2(0,d)=1} \tilde{P}(r) &= P_{\tilde{M}}(1, d||0) + \sum_{r:r_1(0)\neq d, r_2(0,d)=1} \tilde{P}(r) && (18)\\
&= \sum_{r:r_2(1,d)=1} \tilde{P}(r) \\
&= P_{\tilde{M}}(1, d||1) + \sum_{r:r_1(1)\neq d, r_2(1,d)=1} \tilde{P}(r). && (19)
\end{aligned}$$

From (17),

$$\begin{aligned}
\sum_{r:r_2(0,d)=0} \tilde{P}(r) &= P_{\tilde{M}}(0, d||0) + \sum_{r:r_1(0)\neq d, r_2(0,d)=0} \tilde{P}(r) && (20)\\
&= \sum_{r:r_2(1,d)=0} \tilde{P}(r) \\
&= P_{\tilde{M}}(0, d||1) + \sum_{r:r_1(1)\neq d, r_2(1,d)=0} \tilde{P}(r). && (21)
\end{aligned}$$

Since from (16),

$$\sum_r \tilde{P}(r) = \sum_{r:r_2(0,d)=0} \tilde{P}(r) + \sum_{r:r_2(0,d)=1} \tilde{P}(r) = \sum_{r:r_2(0,d)=0} \tilde{P}(r) + \sum_{r:r_2(1,d)=1} \tilde{P}(r) = 1.$$

Substituting from (20) and (19),

$$P_{\tilde{M}}(0, d||0) + \sum_{r:r_1(0)\neq d, r_2(0,d)=0} \tilde{P}(r) + P_{\tilde{M}}(1, d||1) + \sum_{r:r_1(1)\neq d, r_2(1,d)=1} \tilde{P}(r) = 1.$$

Similarly, substituting from (21) and (18),

$$P_{\tilde{M}}(0, d||1) + \sum_{r:r_1(1)\neq d, r_2(1,d)=0} \tilde{P}(r) + P_{\tilde{M}}(1, d||0) + \sum_{r:r_1(0)\neq d, r_2(0,d)=1} \tilde{P}(r) = 1.$$

This implies $P_{\tilde{M}}(0, d||0) + P_{\tilde{M}}(1, d||1) \leq 1, P_{\tilde{M}}(0, d||1) + P_{\tilde{M}}(1, d||0) \leq 1$. These are precisely the IV inequalities and they are satisfied. Therefore, $\mathcal{K}_{\text{cf-inter}} \subseteq \mathcal{K}_{\text{IV}}$.

$\boldsymbol{\mathcal{K}_{\text{cf-graph}} = \mathcal{K}_{\text{IV}}}$ follows from Theorem 17 since $M \in H^0_{\text{cf-graph}}$ implies $M \in \mathbb{M}_{\text{IV}}$. By Proposition 25, the lemma follows.

□

D.4 RELAXING THE ASSUMPTION OF NO CONFOUNDING BETWEEN S AND D

In this section, we prove that Theorem 17 and Theorem 18 hold when $\mathbb{M}_{\text{IV}+}$ and $\mathbb{M}_{\text{cf}+}$ are expanded by allowing for confounding between S and D. Denote the corresponding expanded models by $\mathbb{M}_{\text{IV}+ZX}$ and $\mathbb{M}_{\text{cf}+SD}$, respectively, where the former has structural equations of the form $Z = f_Z(U_Z, U_{ZX}), X = f_X(Z, U_X, U_{ZX}, U_{XY}), Y = f_Y(X, U_Y, U_{XY})$ where $U_Z, U_X, U_Y, U_{ZX}, U_{XY}$ are independent exogenous random variables, and the latter has structural equations of the form $S = f_S(U_S, U_{SD}), D = f_D(S, U_D, U_{SD}, U_{DA}), A = f_A(S, D, U_A, U_{DA})$ where $U_S, U_D, U_A, U_{SD}, U_{DA}$ are independent exogenous random variables. The corresponding expanded causal null hypothesis corresponding to the fairness notions are denoted by $H^0_{\text{cf-notion}+SD}$ where we use 'notion' as a placeholder for 'graph','ctrf', 'inter'.

For $M \in \mathbb{M}_{\text{IV}+ZX}$, pick $U \sim \text{Unif}\,[0,1]$ and a deterministic map $h : \mathcal{Z} \times [0,1] \mapsto \mathcal{U}_{ZX}$ such that $U_{ZX} = h(Z, U)$ a.s.. Note that such an h always exists for any random variable U_{ZX} taking values in a standard measurable space (see e.g. [Forré and Mooij, 2025, Corollary 2.7.7]). Define $\tilde{M} = (V = \{\tilde{Z}, \tilde{X}, \tilde{Y}\}, W = \{U_{\tilde{Z}}, U_{\tilde{X}}, U, U_{\tilde{Y}}, U_{\tilde{X}\tilde{Y}}\}, \tilde{\mathcal{X}} = \mathcal{X}_V \times \mathcal{X}_W, \tilde{f} = (f_{\tilde{Z}}, f_{\tilde{X}}, f_{\tilde{Y}}), \tilde{P})$ where a) $\forall u_{\tilde{Z}}, f_{\tilde{Z}}(u_{\tilde{Z}}) \triangleq u_{\tilde{Z}}$, b) $\forall u_{\tilde{X}}, u, u_{\tilde{X}\tilde{Y}}, \tilde{z}, f_{\tilde{X}}(u_{\tilde{X}}, u, u_{\tilde{X}\tilde{Y}}, \tilde{z}) \triangleq f_X(\tilde{z}, h(\tilde{z}, u), u_{\tilde{X}}, u_{\tilde{X}\tilde{Y}})$, c) $\forall u_{\tilde{Y}}, u_{\tilde{X}\tilde{Y}}, \tilde{x}, f_{\tilde{Y}}(u_{\tilde{Y}}, u_{\tilde{X}\tilde{Y}}, \tilde{x}) \triangleq f_Y(u_{\tilde{Y}}, u_{\tilde{X}\tilde{Y}}, \tilde{x})$, and $\tilde{P} = P_M(Z) \otimes P_X \otimes \text{Unif}\,[0,1] \otimes P_Y \otimes P_{XY}$ where $P_X, P_Y, P_{XY}, P_Z, P_{ZX}$ are marginals of P over $U_X, U_Y, U_{XY}, U_Z, U_{ZX}$ respectively. Note that $P_M(Z)$ is the push-forward of $P_Z \otimes P_{ZX}$ through f_Z thus making $\tilde{P}$ a product distribution over the exogenous random variables. By the above construction, $\tilde{M} \in \mathbb{M}_{\text{IV}+}$. For every $M \in \mathbb{M}_{\text{IV}+ZX}$, $P_M(X, Y, Z) = P_M(Z) \times P_M(X, Y \mid Z) = P_{\tilde{M}}(\tilde{Z}) \times P_{\tilde{M}}(\tilde{X}, \tilde{Y} \mid \text{do}(\tilde{Z})) \in \mathcal{P}_{\mathbb{M}_{\text{IV}+}}$. Therefore, $\{P_M(Z, X, Y) : M \in \mathbb{M}_{\text{IV}+ZX}\} = \mathcal{P}_{\mathbb{M}_{\text{IV}+}}$. Using a similar argument that replaces the labels Z, X, Y by S, D, A, $\{P_M(S, D, A) : M \in H^0_{\text{cf-notion}+SD}\} = \mathcal{P}_{\text{cf-notion}}$. This implies Theorem 17 and Theorem 18 hold for $\mathbb{M}_{\text{IV}+ZX}$ and $\mathbb{M}_{\text{cf}+SD}$ respectively.

E COMPARISON WITH EXISTING NOTIONS

E.1 WITHOUT CONFOUNDING

E.1.1 Counterfactual Fairness and Demographic Parity

We restate the counterfactual notion of fairness from Kusner et al. [2017] for the Berkeley example below.

Definition 27 (Counterfactual Fairness [Kusner et al., 2017])**.** $M \in \mathbb{M}_{\textit{no-cf}}$ *is fair if for all* s, d, $P_M(s, d) > 0$ *implies*

$$P_M\Big(A^{do(S=s')} \mid D = d, S = s\Big) = P_M\Big(A^{do(S=s)} \mid D = d, S = s\Big)$$

for $s' \neq s$.

The counterfactual fairness notion of Kusner et al. [2017] implies demographic parity for the Berkeley example without allowing for confounding.

Proposition 28. *If* $M \in \mathbb{M}_{\textit{no-cf}}$ *is counterfactually fair according to Definition 27, then* P_M *satisfies demographic parity, i.e., for all* s, s' *such that* $P_M(s), P_M(s') > 0$*,*

$$P_M(A = 1 | S = s) = P_M(A = 1 | S = s').$$

Proof. The right-hand side in Definition 27 is $P_M(A \mid D = d, S = s)$. Therefore, for s, d such that $P_M(s, d) > 0$, counterfactual fairness implies that

$$P_M\Big(A^{\text{do}(S=s')}, D = d \mid S = s\Big) = P_M(A, D = d \mid S = s).$$

Marginalizing D,

$$P_M\Big(A^{\text{do}(S=s')} \mid S = s\Big) = P_M(A \mid S = s). \tag{22}$$

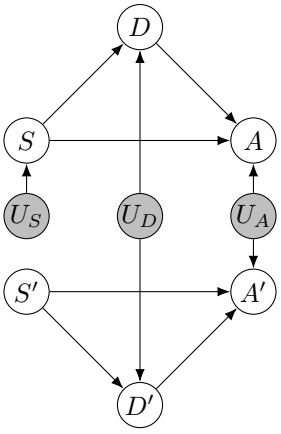


Figure 6: Causal graph of twin network $(M^{\text{twin}})^{\text{do}(S'=s')}$

By Rule 2 of do-calculus, if $P_M(S = s) > 0$,

$$\begin{aligned} P_M(A \mid S = s) &= P_M(A \mid \text{do}(S = s)), \\ &\overset{(a)}{=} P_M\Big(A^{\text{do}(S=s')} \mid S = s\Big), \\ &\overset{(b)}{=} P_{M^{\text{twin}}}(A' \mid \text{do}(S' = s'), S = s), \\ &\overset{(c)}{=} P_M(A \mid \text{do}(S = s')), \\ &\overset{(d)}{=} P_M(A \mid S = s'). \end{aligned}$$

where (a) follows from (22), (b) follows from expressing the counterfactual $P_M\Big(A^{\text{do}(S=s')} \mid S = s\Big)$ in the twin network model, (c) follows from the twin network, and (d) follows from Rule 2 of do-calculus since $P_M(s') > 0$. Therefore, counterfactual fairness implies demographic parity. □

Note that demographic parity falls prey to Simpson's paradox in the Berkeley example. The above result shows that a valid test for demographic parity is a valid test for Kusner et al. [2017]'s counterfactual fairness notion for the assumed model class, $\mathbb{M}_{\text{no-cf}}$.

E.1.2 Path-dependent Counterfactual Fairness

We next show that testing the path-dependent counterfactual fairness notion given in the appendix of Kusner et al. [2017] coincides with a conditional independence test $A \perp\!\!\!\perp S \mid D$.

Definition 29 (Path-dependent Counterfactual Fairness [Kusner et al., 2017])**.** $M \in \mathbb{M}_{\textit{no-cf}}$ *is fair if for all* s, d, $P_M(s, d) > 0$ *implies*

$$P_M\Big(A^{do(S=s',D=d)} = 1 \mid D = d, S = s\Big) = P_M\Big(A^{do(S=s,D=d)} = 1 \mid D = d, S = s\Big)$$

for $s' \neq s$.

Proposition 30.

$$\{P_M(A, D, S) : M \in \mathbb{M}_{\textit{no-cf}} \textit{ satisfies path-dependent counterfactual fairness}\} = \mathcal{P}_{\textit{no-cf-obs}}.$$

Proof. We first show that if M satisfies the path-dependent counterfactual fairness notion then $M \in H^0_{\text{no-cf-obs}}$. The path-dependent counterfactual fairness notion implies that for all s, d such that $P_M(s, d) > 0$,

$$P_M\Big(A^{\text{do}(S=s',D=d)} = 1 \mid D = d, S = s\Big) = P_M(A = 1 \mid D = d, S = s).$$

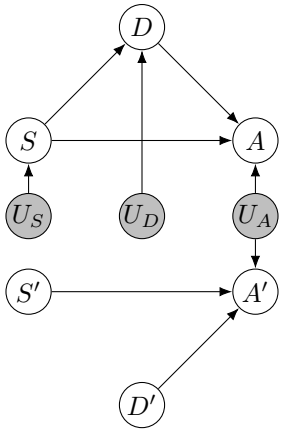


Figure 7: Causal graph of twin network $(M^{\text{twin}})^{\text{do}(S'=s',D'=d)}$ for $M \in \mathbb{M}_{\text{no-cf}}$

If for $s' \neq s$, $P_M(s', d) > 0$, then we simplify $P_M\Big(A^{\text{do}(S=s',D=d)} = 1 \mid D = d, S = s\Big)$ using the twin network in Figure 7.

$$\begin{aligned} P_M\Big(A^{\text{do}(S=s',D=d)} = 1 \mid D = d, S = s\Big) &= P_{M^{\text{twin}}}(A' = 1 \mid \text{do}(D' = d, S' = s'), D = d, S = s) \\ &\overset{(a)}{=} P_{M^{\text{twin}}}(A' = 1 \mid \text{do}(D' = d, S' = s')), \\ &\overset{(b)}{=} P_M(A = 1 \mid D = d, S = s'), \end{aligned}$$

where (a) follows since $S, D \perp\!\!\!\perp A'$ in the intervened twinned SCM and (b) follows from the twinned SCM. This implies that

$$P_M(A = 1 \mid D = d, S = s) = P_M(A = 1 \mid D = d, S = s') = P_M(A = 1 \mid D = d).$$

If instead, $P_M(s', d) = 0$, then still $P_M(A = 1 \mid D = d, S = s) = P_M(A = 1 \mid D = d)$. Therefore, $M \in H^0_{\text{no-cf-obs}}$.

Clearly, if $M \in H^0_{\text{no-cf-graph}}$, path-dependent counterfactual fairness is satisfied. However, note that, Example 14 satisfies path-dependent counterfactual fairness but does not belong to $H^0_{\text{no-cf-graph}}$. The conclusion follows from Theorem 13. □

E.2 WITH CONFOUNDING

In this section, we compare the statistical tests that result from the NDE that the notions of Nabi and Shpitser [2018] and Chiappa [2019] are based on, and Kusner et al. [2017]'s counterfactual fairness and path-dependent counterfactual fairness notions, when confounding is allowed.

E.2.1 NDE

With confounding between the mediator and the outcome, Kaufman et al. [2005] obtain bounds on the NDE for the all-binary variable case by the linear programming approach of Balke and Pearl [1997]. The resulting bounds are implied by the IV inequalities but not equivalent to them, which gives us a strictly weaker test than the IV inequalities. For completeness, we present the bounds below. The lower and upper bounds for $NDE(A; 0 \to 1)$ are

$$\max \begin{Bmatrix} P(A=0 \mid S=0)-1, \\ P(A=0 \mid D=0 \mid S=0)-P(A=1 \mid D=1 \mid S=0)+P(A=1 \mid D=0 \mid S=1)-1 \\ P(A=0 \mid D=1 \mid S=0)-P(A=1 \mid D=0 \mid S=0)+P(A=1 \mid D=1 \mid S=1)-1 \end{Bmatrix},$$
$$\min \begin{Bmatrix} 1-P(A=1 \mid S=0), \\ 1+P(A=0 \mid D=1 \mid S=0)-P(A=1 \mid D=0 \mid S=0)-P(A=0 \mid D=0 \mid S=1) \\ 1+P(A=0 \mid D=0 \mid S=0)-P(A=1 \mid D=1 \mid S=0)-P(A=0 \mid D=1 \mid S=1) \end{Bmatrix}.$$

The lower and upper bounds for $NDE(A; 1 \to 0)$ are

$$\max \begin{Bmatrix} P(A=0 \mid S=1)-1, \\ P(A=1 \mid D=0 \mid S=0)-P(A=1 \mid D=1 \mid S=1)+P(A=0 \mid D=0 \mid S=1)-1 \\ P(A=1 \mid D=1 \mid S=0)-P(A=1 \mid D=0 \mid S=1)+P(A=0 \mid D=1 \mid S=1)-1 \end{Bmatrix},$$
$$\min \begin{Bmatrix} 1-P(A=1 \mid S=1), \\ 1+P(A=0 \mid D=0 \mid S=1)-P(A=0 \mid D=1 \mid S=0)-P(A=1 \mid D=1 \mid S=1) \\ 1+P(A=0 \mid D=1 \mid S=1)-P(A=0 \mid D=0 \mid S=0)-P(A=1 \mid D=0 \mid S=1) \end{Bmatrix}.$$

Equating the NDE to 0, gives us a strictly larger null hypothesis compared to the one obtained based on the IV inequalities. This implies that the resulting statistical test is strictly weaker.

E.2.2 Counterfactual Fairness [Kusner et al., 2017]

The proof of Proposition 28 also holds when confounding is allowed since the implications of the do-calculus rules in the proof hold even in the twin network with confounding.

E.2.3 Path-dependent Counterfactual Fairness [Kusner et al., 2017]

Proposition 31. *If $M \in \mathbb{M}_{cf+}$ satisfies path-dependent counterfactual fairness then $P_M(D, A, S) \in \mathcal{P}_{\text{cf-graph}} = \mathcal{P}_{IV+}$. If $M \in H^0_{\text{cf-graph}+}$, then M satisfies path-dependent counterfactual fairness.*

Proof. We show that a model $M \in \mathbb{M}_{\text{cf}+}$ that satisfies the path-dependent counterfactual fairness notion is observationally equivalent to a model in $H^0_{\text{cf-graph}+}$. This implies that the set of observational distributions of models that satisfy the path-dependent counterfactual notion of fairness, are described by $\mathcal{P}_{\text{cf-graph}}$.

If $M \in \mathbb{M}_{\text{cf}+}$ satisfies path-dependent counterfactual fairness, then for all s, d such that $P_M(s, d) > 0$,

$$P_M\Big(A^{\text{do}(S=s',D=d)} = 1 \mid D=d, S=s\Big) = P_M\Big(A^{\text{do}(S=s,D=d)} = 1 \mid D=d, S=s\Big)$$

for $s' \neq s$. Note that the right-hand side above is $P_M(A = 1 \mid D = d, S = s)$. The counterfactual $P_M\Big(A^{\text{do}(S=s',D=d)} = 1 \mid D=d, S=s\Big)$ is given by the push-forward of $P(U_A, U|D = d, S = s)$ through $f_A(s', d, U_A, U)$. Because of the independence on the value of s', the same holds for the function

$$\bar{f}_A(d, U_A, U) = \frac{1}{2}(f_A(0, d, U_A, U) + f_A(1, d, U_A, U)).$$

Consider an SCM, $\bar{M}$, that is identical to M except for the causal mechanism of A being $\bar{f}_A$. Clearly, $\bar{M} \in H^0_{\text{cf-graph}+}$ and $P_M(S, D) = P_{\bar{M}}(S, D)$. By the above argument, for all s, d such that $P_M(s, d) > 0$, we have $P_M(A = 1 \mid D = d, S = s) = P_{\bar{M}}(A = 1 \mid D = d, S = s)$. This implies that $P_M(D, A, S) \in \mathcal{P}_{\text{cf-graph}} = \mathcal{P}_{\text{IV}+}$. Con-

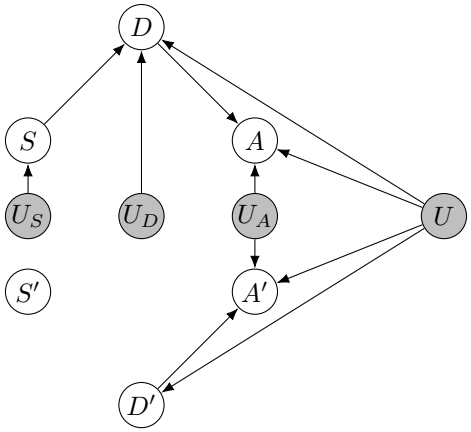


Figure 8: Causal graph of twin network $(M^{\text{twin}})^{\text{do}(S'=s',D'=d)}$ for $M \in H^0_{\text{cf-graph+}}$

versely, if $M \in H^0_{\text{cf-graph+}}$, then from the twin network of M in Figure 8, clearly $A' \perp\!\!\!\perp S' \mid S, D$ and therefore, path-dependent counterfactual fairness is satisfied.

□

Therefore, for the assumed model class, a valid statistical test for path-dependent counterfactual fairness is also a valid test for the IV inequalities from Theorem 17 and vice versa.

F ADDITIONAL RESULTS FOR BAYESIAN TESTING PROCEDURE: CUM-LAUDE DATASET AND PRIOR SENSITIVITY

We consider the dataset from Bol [2023] that contains data from 5239 PhD students in the Netherlands studying at a large Dutch university from 2011-2021. Bol [2023] observed a bias in the percentage of 'cum-laude' distinctions awarded to male PhD students (6.57%) versus female PhD students (3.68%).

As in the Berkeley example, there is data on the sex of the student, their academic field and whether they were awarded cum-laude. Unlike the Berkeley example, there are more covariates that measure additional information, including the sex composition of the dissertation committee, the sex composition of the supervisory team that includes the promoters and co-promoters. For the current analyses we don't take into account these covariates and only analyze the dataset with respect to sex, academic field and award outcome.

As reported by Bol [2023], unlike the Berkeley dataset, the bias among female and male cum-laude award rates does not vanish when conditioned on department. Therefore, with the assumption of no confounding between the academic field choice and the cum-laude award outcome, the conclusion of the conditional independence test implies that the data generating mechanism is unfair when assuming that no latent unprotected mediators between sex and cum-laude award rates exist. Allowing for confounding requires us to use the Bayesian testing procedure proposed in Section 5 for the IV inequalities. We have $|\mathcal{X}| = |\mathcal{X}_D| = 6$, $|\mathcal{Y}| = |\mathcal{X}_A| = 2, |\mathcal{Z}| = |\mathcal{X}_S| = 2$. We choose a flat Dirichlet prior over parameters ($\theta = \{P(d, a, s) : d \in \mathcal{X}_D, a \in \mathcal{X}_A, s \in \mathcal{X}_S\}$) in both models $\mathbb{M}_0, \mathbb{M}_1$, i.e., for $i = 0, 1, \pi(\theta|\mathbb{M}_i) = c_i \text{Dir}(1, 1, \cdots, 1)$ where c_i is a normalizing constant. The counts from the data $R_1, R_2 \cdots R_m$ are used to obtain the posterior, $P(\theta \mid R_1, R_2, \cdots R_m)$ which is also a Dirichlet distribution. Using $n = 10^6$ samples, we observe no violations of the IV inequality. Therefore, the confidence interval for the posterior probability of the cum-laude data satisfying the IV inequalities is $\left[1 - 3.69 \times 10^{-6}, 1\right]$. Hence, when allowing for confounding, on arrives at the conclusion that, given the available data and restricting the analysis to only three variables, the fairness of the data-generating mechanism is undecidable.

Prior Sensitivity: For both the Berkeley dataset and the Bol dataset, the final confidence interval is dependent on the choice of the prior. We presented the analysis with a flat Dirichlet prior, for both datasets. We find that the lower limit of the confidence interval does not change as we vary the parameter α over the interval $\left[10^{-2}, 10^5\right]$ for a Dirichlet $\text{Dir}(\alpha, \alpha, \cdots, \alpha)$ prior.

Frequentist Test of Wang et al. [2017]: The frequentist test of Wang et al. [2017] converts every IV inequality into a

one-sided association test for a 2×2 contingency table. Specifically, for fixed d, a, an IV inequality of the form

$$\Pr\left(D = d, A = a \mid S = 1\right) + \Pr\left(D = d, A = 1 - a \mid S = 0\right) \leq 1$$

is transformed into

$$\gamma^{d,a} \leq 0$$

where $\gamma^{d,a} \triangleq \Pr\left(Q^{d,a} = 1 \mid S = 1\right) - \Pr\left(Q^{d,a} = 1 \mid S = 0\right)$ where

$$Q^{d,a} = \begin{cases} \mathbf{1}\left[D = d, A = a\right] & \text{if } S = 1, \\ 1 - \mathbf{1}\left[D = d, A = 1 - a\right] & \text{if } S = 0. \end{cases}$$

Note that $Q^{d,a}$ and S are binary random variables and $\gamma^{d,a} = 0$ if and only if $Q^{d,a} \perp\!\!\!\perp S$. Further $\gamma^{d,a} \in [-1, 1]$ for all d, a.

Since the direction of the one-sided test matters, we check the sign of the difference of conditional probabilities using maximum likelihood (ML) estimates and then conduct a Pearson's chi-square test for independence. For the Berkeley data, the ML estimates of $\gamma^{d,a}$ were negative (less than -0.6) for all d, a, and the independence tests rejected the null hypothesis of independence with p-value 0.0 (i.e., less than the smallest positive number representable using double precision floating point format, i.e., $< 5 \times 10^{-324}$). We take this to be significant evidence that the null hypothesis of $\gamma^{d,a} \leq 0$ is not rejected. As noted in Wang et al. [2017], although we test for multiple IV inequalities, the Bonferroni correction is $1/2$ and does not scale as the number of IV inequalities.