The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Counterfactual Identification Under Monotonicity Constraints

Aurghya Maiti, Drago Plecko, Elias Bareinboim

Causal Artificial Intelligence Laboratory, Columbia University
am5887@columbia.edu, dp3144@columbia.edu, eb@cs.columbia.edu

Abstract

Reasoning with counterfactuals is one of the hallmarks of human cognition, involved in various tasks such as explanation, credit assignment, blame, and responsibility. Counterfactual quantities that are not identifiable in the general non-parametric case may be identified under shape constraints on the functional mechanisms, such as monotonicity. One prominent example of such an approach is the celebrated result by Angrist and Imbens on identifying the Local Average Treatment Effect (LATE) in the instrumental variable setting. In this paper, we study the identification problem of more general settings under monotonicity constraints. We begin by proving the monotonicity reduction lemma, which simplifies counterfactual queries using monotonicity assumptions and facilitates the reduction of a larger class of these queries to interventional quantities. We then extend the existing identification results on Probabilities of Causation (PoCs) and LATE to a broader set of queries and graphs. Finally, we develop an algorithm, M-ID, for identifying arbitrary counterfactual queries from combinations of observational and experimental data, which takes as input a causal diagram with monotonicity constraints. We show that M-ID subsumes the previously known identification results in the literature. We demonstrate the applicability of our results using synthetic and real data.

1 Introduction

Counterfactual reasoning is essential for human cognition, underpinning various of our abilities related to understanding, credit assignment, attribution of blame and responsibility, and regret (Pearl 2000; Pearl and Mackenzie 2018; Starr 2019; Van Hoeck, Watson, and Barbey 2015). In a structure known as *Pearl Causal Hierarchy* (PCH), counterfactual knowledge resides at Layer 3, while observational and interventional knowledge corresponds to Layers 1 and 2 (Pearl and Mackenzie 2018; Bareinboim et al. 2022).

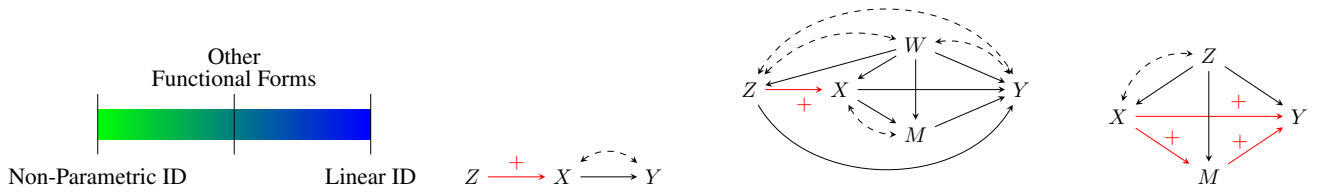
The question of non-parametric identification of causal queries from one layer of the PCH using data from another layer has received a lot of attention in the literature. Various versions of this problem have been studied extensively, from Pearl's celebrated result known as do-calculus to other more systematic, algorithmic approaches (Pearl 1995; Tian and Pearl 2002; Shpitser and Pearl 2007; Huang and Valtorta 2006; Bareinboim and Pearl 2012, 2016; Correa and Bareinboim 2017; Lee, Correa, and Bareinboim 2019; Correa and Bareinboim 2020; Lee and Bareinboim 2020; Lee, Correa, and Bareinboim 2020; Correa and Bareinboim 2024). Specifically, the problem of identifying interventional queries from observational data and the causal diagram (Layer 2 from Layer 1) has been solved by the ID algorithm from (Tian and Shpitser 2010) and from a combination of observations and experiments (Layers 1+2 to Layer 2) (Lee, Correa, and Bareinboim 2019). Similarly, the Ctf-ID algorithm from (Correa, Lee, and Bareinboim 2021) solves the problem of identifying counterfactual queries from a combination of observations and arbitrary experiments (Layer 3 from Layers 1+2). These algorithms have been shown to be sound and complete.

In contrast, the causal inference literature in econometrics and statistics has traditionally considered effect identification under parametric assumptions. A popular and well-studied case is effect identification in linear systems (Brito and Pearl 2002; Tian 2004, 2005; Chen, Pearl, and Bareinboim 2016; Chen, Kumor, and Bareinboim 2017; Kumor, Chen, and Bareinboim 2019; Kumor, Cinelli, and Bareinboim 2020; Shimizu 2014). The literature on this area has a rich past, rooted in the study of regression (Gauss 1877; Galton 1886) and instrumental variables (Wright 1928; Reiersøl 1945). This setting could be seen as the opposite of non-parametric identification, which makes no assumptions about the form of causal mechanisms, whereas the linear identification setting assumes all mechanisms (globally) to be linear (see Fig. 1a).

Interestingly, the space between the two extremes of the spectrum in Fig. 1a has received relatively less attention, and yet many interesting possibilities exist for considering effect identification under other functional form assumptions. These include examples such as additive noise models (Peters, Janzing, and Scholkopf 2011), models with local parametric assumptions (as opposed to linear models where every mechanism is assumed to be linear), shape-constrained models (assuming monotonic or convex/concave functional forms) (Imbens and Angrist 1994), and many others.

In this paper, we make an important step in this direction and study the identification of counterfactuals under local monotonicity constraints. To illustrate how such constraints can aid identification, we begin with the following example.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Identification spectrum. (b) IV Graph (c) LATE+ setting (d) Standard Fairness Graph

Figure 1: (a) Spectrum of identification settings for different functional forms. (b) IV Graph. (c) Example of a graph where LATE is identifiable, but the assumptions of LATE are not satisfied. (d) Standard fairness graph

Example 1 (Local Average Treatment Effect or LATE (Imbens and Angrist 1994))**.** *Consider the instrumental variable setting in Fig. 1b with variables X (binary), Y, and an instrument Z (binary). The Local Average Treatment Effect (LATE) is defined as the effect of X on Y within the group of units who "comply" with the treatment Z, meaning that $X = 0$ in the absence of Z and $X = 1$ in the presence of Z, written $X_{z_0} = 0, X_{z_1} = 1$. The LATE quantity can be written in counterfactual notation as:*

$$\textit{LATE} = \mathbb{E}[Y_{x_1} - Y_{x_0} \mid X_{z_0} = 0, X_{z_1} = 1]. \tag{1}$$

In the general non-parametric setting, the LATE is not identifiable (uniquely computable) from observational data. However, under the monotonicity assumption of $Z \to X$, that is if for all individuals $Z = z_1$ produces a greater or equal outcome in X than $Z = z_0$, written

$$X_{z_1}(\mathbf{u}) \geq X_{z_0}(\mathbf{u}) \tag{2}$$

for all assignments $\mathbf{u}$ of unobserved variables, the quantity can be computed as:

$$\textit{LATE} = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[X \mid Z = 1] - \mathbb{E}[X \mid Z = 0]}. \tag{3}$$

This is a seminal result, widely used in the econometrics literature (Imbens and Angrist 1994), and is part of the reason why the original authors were awarded a Nobel Prize.

Various extensions of the basic setting in Fig. 1b have been studied. Consider for instance the 401(k) dataset studied in (Abadie 2003), represented here by the causal diagram in Fig. 1c, with the following variables: income (W), 401(k) eligibility (Z), 401(k) participation (X), net financial assets (M), and total wealth (Y). We wish to compute the LATE of 401(k) participation (X) on total wealth (Y).

Note that eligibility (Z) cannot serve as an instrument due to potential confounders like self-employment and preferences for non-financial assets, some of which may be unobserved, represented as a bidirected edge between Z and Y. In addition, there may also be unobserved confounding such as saving proclivity, between 401(k) participation (X) and net financial assets (M). Our goal is to compute the LATE across income groups,

$$\textit{LATE}(w) = \mathbb{E}[Y_{x_1} - Y_{x_0} \mid X_{z_0} = 0, X_{z_1} = 1, w]. \tag{4}$$

Applying the classical non-parametric LATE estimator from (Imbens and Angrist 1994) conditional on the income group $W = w$,

$$\textit{classic-LATE}(w) = \frac{\mathbb{E}[Y \mid z_1, w] - \mathbb{E}[Y \mid z_0, w]}{\mathbb{E}[X \mid z_1, w] - \mathbb{E}[X \mid z_0, w]} \tag{5}$$

would, in this context, lead to incorrect conclusions. Similarly, an attempt of conditioning on M before applying the classical LATE estimator, labeled cond(M)-LATE(w), given by

$$\sum_m P(m \mid w) \frac{\mathbb{E}[Y \mid z_1, m, w] - \mathbb{E}[Y \mid z_0, m, w]}{\mathbb{E}[X \mid z_1, m, w] - \mathbb{E}[X \mid z_0, m, w]} \tag{6}$$

would also lead to an incorrect result. Thus, interestingly, state-of-the-art methods in the econometrics literature cannot solve the problem at hand. In this paper, we develop a general, graphical approach to identification under monotonicity and derive the correct identification expression

$$\frac{\sum_y y \cdot Q_1(y, w)}{\sum_y Q_1(y, w)} - \frac{\sum_y y \cdot Q_0(y, w)}{\sum_y Q_0(y, w)} \tag{7}$$

where, the weights $Q_i(y)$ are given by

$$Q_i(y, w) = \sum_{m,z} P(y \mid w, z, m, x_i) P(z \mid w) \tag{8}$$
$$\big[P(m, x_i \mid w, z_i) - P(m, x_i \mid w, z_{1-i})\big].$$

The above example is an initial indication of the usefulness of the graphical approach to causality. In previous works in econometrics (Imbens and Angrist 1994; Abadie 2003), the assumptions used for identification are hard-coded to a specific setting. In this paper, we show how to encode shape constraints into causal diagrams and then prove more general results by algorithmically leveraging the topological constraints within the graph. In doing so, we challenge the prior belief in the literature that "causal diagrams have difficulty coding shape restrictions such as monotonicity" (Imbens 2020) and demonstrate the opposite: graphical models provide a flexible and transparent language for expressing a broader set of assumptions, leading to novel identification results. Formally, our contributions are as follows:

1. We introduce a graphical encoding of *monotonicity* (Def. 2) and prove the monotonicity reduction lemma (MRL) (Lem. 2), which allows us to reduce a broad class of counterfactual queries to interventional queries.

2. Leveraging this result, we extend the identification results for LATE and Local Natural Direct Effect (LNDE) to a broader graphical context (Prop. 3 and 4). Additionally, we establish identification conditions of a generalization of the probability of necessity (PN) and probability of sufficiency (PS) that allow for conditioning on any post-treatment variable(s) (Prop. 5).
3. We then develop a sound algorithm for identifying arbitrary counterfactual queries based on the causal graph and local monotonicity constraints (Alg. 1).

Finally, in Sec. 4, we demonstrate our methods on both synthetic and real data (analyzing the data from (Abadie 2003), as mentioned in the example), showcasing their practical utility. All proofs and expressions for identification (ID expressions) are provided in Appendix B.

Preliminaries Throughout this work, we use the language of structural causal models (SCMs) (Pearl 2000). An SCM is defined as a tuple $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$, where $\mathbf{V}$ and $\mathbf{U}$ are sets of endogenous (observable) and exogenous (latent) variables, respectively. $\mathcal{F}$ is a set of functions f_{V_i}, one for each $V_i \in \mathbf{V}$, where $V_i \leftarrow f_{V_i}(\mathbf{Pa}(V_i), U_{V_i})$, with $\text{pa}(V_i) \subseteq \mathbf{V}$ and $U_{V_i} \subseteq \mathbf{U}$. $P(\mathbf{u})$ is a strictly positive probability measure over $\mathbf{U}$. Each SCM $\mathcal{M}$ is associated with a causal diagram $\mathcal{G}$ over the node set $\mathbf{V}$, where $V_i \rightarrow V_j$ if V_i is an argument of f_{V_j}, and $V_i \dashleftarrow\!\dashrightarrow V_j$ if U_{V_i} and U_{V_j} are dependent (Bareinboim et al. 2022). The potential response $Y_x(\mathbf{u})$ is the value of Y when $X = x$ for the unit $\mathbf{u}$, derived by evaluating $\mathbf{u}$ in the submodel $\mathcal{M}_x$, where equations associated with X are replaced by $X = x$.

Related Works (VanderWeele and Robins 2010) introduced strong monotonicity, where a variable has a monotonic relationship with its parents. An edge $X \rightarrow Y$ is signed positive or negative based on whether X has a positive or negative monotonic effect on Y. While they focus on providing conditions for deriving inequalities in observational distribution, our work leverages monotonicity for identifying counterfactual quantities. Identification of specific quantities like LATE (Imbens and Angrist 1994), LNDE (Yamamoto 2013), and PN/PS (Pearl 2022) have also been explored under monotonicity, often with specific distributional assumptions. Our algorithm builds on (Correa, Lee, and Bareinboim 2021), where the concepts of ctf-factors and inconsistency were introduced. A ctf-factor is a distribution of the form $P(W_{1[\mathbf{pa}_1]} = w_1, \ldots, W_{l[\mathbf{pa}_l]} = w_l)$, where each $W_i \in \mathbf{V}$ and $\mathbf{pa}_i$ are the values of parents of W_i. A ctf-factor is inconsistent if it has a single c-component and one of the following holds:

1. **(Parent-Child)** $\exists W_{\mathbf{t}} \in W_*, Z \in \mathbf{T} \cap V(\mathbf{W}_*)$ such that $z \in \mathbf{t}, z' \in \mathbf{w}_*$ and $z \neq z'$
2. **(Common Parent)** $\exists W_{i[\mathbf{t}_i]}, W_{j[\mathbf{t}_j]} \in \mathbf{W}_*$ and $T \in \mathbf{T}_i \cap \mathbf{T}_j$ such that $t \in \mathbf{t}_1, t' \in \mathbf{t}_2$ and $t \neq t'$.

(Correa, Lee, and Bareinboim 2021) also showed that such inconsistencies imply that the ctf-factors are non-identifiable (non-ID). Later, we demonstrate how monotonicity can resolve certain inconsistencies in a ctf-factor.

2 Extending Identification of Counterfactual Quantities under Monotonicity

In this section, we explore the concept of *local monotonicity* in causal graphs, where a variable depends monotonically only on its parents, and show how these local constraints can relate to global constraints. While global monotonicity assumptions are often challenging, local monotonicity is more practical and more accessible to assume.

Definition 1 (Local Monotonicity Property)**.** *Let X be a variable with a parent Z. We say that Z has a positive (negative) monotonic relationship with X if for all values z, z' of Z and for all assignment $\mathbf{pa}^-$ to $Pa(X) \setminus Z$, we have*

$$X_{z,\mathbf{pa}^-}(\mathbf{u}) \geq X_{z',\mathbf{pa}^-}(\mathbf{u}) \tag{9}$$

whenever $z > z'$ ($z < z'$). An edge $Z \rightarrow X$ is non-monotonic *if Z has neither positive nor negative monotonic relationship with X. Otherwise, it is called* monotonic*.*

A convenient aspect of such local properties is its ease of representation in a causal graph, since it is an edge property. Now, we formally define *causal diagrams with monotonicity assumptions*.

Definition 2 (Causal Graph with Monotonicity Assumptions or CGMA)**.** *A CGMA is a tuple $\langle G, M \rangle$, where*

- *G is a causal graph, with set of vertices $V(G)$, set of directed edges $E_D(G)$ and bidirected edges $E_B(G)$.*
- *$M \subseteq E_D \times \{+, -\}$ is the set of monotonicity assumptions, where if $((X, Y), +) \in M$, then X is positive monotonic on Y and if $((X, Y), -) \in M$, then X is negative monotonic on Y.*

In the following lemma, we explore the relation between two variables that do not have a parent-child relation.

Lemma 1. *If the product of signs over edges along all paths from Z to X are positive in CGMA G, then the global relation between Z and X is positive monotonic, that is, for all assignments $\mathbf{u}$ of exogenous variables and for values z, z' of Z*

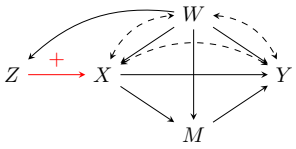
$$X_z(\mathbf{u}) \geq X_{z'}(\mathbf{u}) \tag{10}$$

whenever $z > z'$. Conversely, if any path from Z to X includes a non-monotonic edge, then there exists an SCM consistent with G where Eq. 10 does not hold.

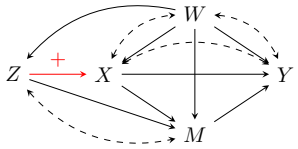
The lemma can be used to identify PoCs in CMGA shown Fig. 1d. Note that the condition is necessary to guarantee that all SCMs consistent with the CGMA satisfy Eq. 10. Hence, we remark that assuming global monotonicity, as done in previous works including (Pearl 2022), is equivalent to assuming these conditions.

In the rest of the paper, we use the term *monotonic* to mean positive monotonic unless specified otherwise. We now introduce a lemma that leverages these monotonicity constraints to simplify ctf-factors, making previously non-identifiable counterfactual events identifiable. Let $\mathbf{t_1}, \mathbf{t_2}$ denote two assignments of a set of variables $\mathbf{T}$. We say $\mathbf{t_1} \leq \mathbf{t_2}$ if for all $T \in \mathbf{T}$, we have $t \in \mathbf{t_1}, t' \in \mathbf{t_2}$ and $t \leq t'$.

Lemma 2 (Monotonicity Reduction Lemma (MRL))**.** *Let $\mathbf{T}, \mathbf{S}$ be a partition of the parents of W such that $\mathbf{T}$ and $\mathbf{S}$*



(a) LNDE setting.



(b) LNDE+ setting.

Figure 2: Examples related to local average treatment effects (LATE) and Local Natural Direct Effect (LNDE)

are the set of monotonic and non-monotonic parents of W respectively. Let $P(Y_*, W_{\mathbf{t},\mathbf{s}} = w, W_{\mathbf{t}',\mathbf{s}} = w')$ be a ctf-factor. If W is binary, then we can apply the following rules to reduce it to a simpler ctf-factor.

1. ***Simplification Rule:*** *If $\mathbf{t} \leq \mathbf{t}'$, then*
   (a) $P(Y_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 0) = P(Y_*, W_{\mathbf{t}',\mathbf{s}} = 0)$
   (b) $P(Y_*, W_{\mathbf{t},\mathbf{s}} = 1, W_{\mathbf{t}',\mathbf{s}} = 1) = P(Y_*, W_{\mathbf{t},\mathbf{s}} = 1)$
2. ***Difference Rule:*** *If $\mathbf{t} \leq \mathbf{t}'$, then*

$$\begin{aligned} &P(Y_*, W_{\mathbf{t},\mathbf{s}} = 0, W_{\mathbf{t}',\mathbf{s}} = 1) \\ &\quad = P(Y_*, W_{\mathbf{t}',\mathbf{s}} = 1) - P(Y_*, W_{\mathbf{t},\mathbf{s}} = 1) \quad (11) \\ &\quad = P(Y_*, W_{\mathbf{t},\mathbf{s}} = 0) - P(Y_*, W_{\mathbf{t}',\mathbf{s}} = 0) \quad (12) \end{aligned}$$

In practice, Eq. 11 and 12 are applied in such a way that the resulting term can be consistent. For instance, if w_1 (or w_0) appears in the ctf-expression of its children, we would use Eq. 11 (or Eq. 12). Consequently, when designing an algorithm, we will apply Rule 2 first on children and then on parents. An example of the application of MRL is shown below:

Example 2. *Consider the ctf-factor $P(M_{x_0 z} = 0, M_{x_1 z} = 1, Y_{x_1 z m_1} = 0, Y_{x_0 z m_0} = 0)$ with respect to the causal graph in Fig. 1d, where X and M are binary. We can simplify this quantity as follows:*

$$\begin{aligned} &P(M_{x_0 z} = 0, M_{x_1 z} = 1, Y_{x_1 z m_1} = 0, Y_{x_0 z m_0} = 0) \\ &= P(M_{x_0 z} = 0, M_{x_1 z} = 1, Y_{x_1 z m_1} = 0) \quad \textit{(Rule 1)} \\ &= P(M_{x_1 z} = 1, Y_{x_1 z m_1} = 0) \\ &\qquad - P(M_{x_0 z} = 1, Y_{x_1 z m_1} = 0) \quad \textit{(Rule 2)} \end{aligned}$$

Now, these simplified terms can be written as

$$\begin{aligned} &P(m_1 \mid x_1, z) P(y_0 \mid x_1, z, m_1) \\ &\qquad - P(m_1 \mid x_0, z) P(y_0 \mid x_1, z, m_1) \quad (13) \end{aligned}$$

In a later section, we will propose an algorithmic approach for applying these rules to any general ctf-factor. We will also demonstrate (in Thm. 2) that if we cannot get a set of consistent ctf-factors by application of MRL, then the ctf-factor is non-identifiable (non-ID).

At first glance, the binary nature of W may seem limiting. However, note that the lemma can be applied whenever the domain can be reduced to a binary form. Consider the counterfactual query $P(X_{z_0} \leq x < X_{z_1})$ in the CGMA in Fig. 1b. We can treat any value $\leq x$ as 0 and any value $> x$ as 1. Then, by applying the difference rule, we obtain:

$$\begin{aligned} P(X_{z_0} \leq x < X_{z_1}) &= P(X_{z_0} \leq x) - P(X_{z_1} \leq x) \quad (14) \\ &= P(X_{z_1} > x) - P(X_{z_0} > x) \quad (15) \end{aligned}$$

We provide a detailed discussion of the application of MRL to such queries in the non-binary case in Appendix D. For the next sections, we will use non-identifiable to mean non-identifiable from observational distribution unless specified otherwise.

2.1 Identifying Local Effects

Identifying and estimating the Local Average Treatment Effect (LATE) has been extensively studied in previous literature (Imbens and Angrist 1994; Angrist and Imbens 1995; Frölich 2007; Heckman, Urzua, and Vytlacil 2006; Chernozhukov et al. 2018).

LATE Extensions The assumptions for identification, as proposed in these earlier works, can be restrictive in practice, with an explicit example presented in Ex. 1 related to the causal graph in Fig. 1c. In this setting, the assumption on the existence of valid instrument (Imbens and Angrist 1994) is violated since Y_{x_0} and Y_{x_1} are not independent of Z. Any attempt to apply the standard LATE formulation or conditioning on M can lead to incorrect conclusions. Interestingly, $P(Y_{x_1}, X_{z_0} = 0, X_{z_1} = 1)$ can be identified by decomposing the effect into two factors - the effect of X on M and the effect of X and M on Y. The former can be identified using Z as an instrument, and the latter is identifiable from observation. Once $P(Y_{x_1}, X_{z_0} = 0, X_{z_1} = 1)$ has been computed, we can identify the query in Eq. 4 of the introductory example, as stated in the following proposition:

Proposition 3 (LATE Extensions)**.** *LATE is identifiable in the causal graph in Fig. 1c, with local monotonicity of $Z \to X$, where X is binary. In particular, the same is given by the expression:*

$$\frac{\sum_{w,y} y \cdot Q_1(y, w) P(w)}{\sum_{w,y} Q_1(y, w) P(w)} - \frac{\sum_{w,y} y \cdot Q_0(y, w) P(w)}{\sum_{w,y} Q_0(y, w) P(w)}, \quad (16)$$

where the weight $Q_i(y, w)$ can be evaluated as follows

$$\begin{aligned} Q_i(y, w) = \sum_{m,z} &P(y \mid w, z, m, x_i) P(z \mid w) \\ &\big[P(m, x_i \mid w, z_i) - P(m, x_i \mid w, z_{1-i})\big]. \quad (17) \end{aligned}$$

Similarly, LATE is also ID in the causal graph in Fig. 2b, with the identification expression given in Appendix B.

An intuitive understanding of why LATE can be identified in Fig. 1c can be developed through the following series of observations: (i) the total effect consists of the direct effect $X \to Y$ and the indirect effect $X \to M \to Y$,

(ii) $X \to M \to Y$ requires inferring the effect $X \to M$, and $M \to Y$, (iii) for the $X \to M$ effect, Z is a valid instrument, assuming monotonicity, (iv) when considering $\{X, M\}$ jointly, the pair $\{Z, W\}$, in fact, gives a valid backdoor set for the effect of $\{X, M\}$ on Y, reflected in the ignorability statement $Y_{xm} \perp\!\!\!\perp X, M \mid Z, W$. This allows the identification of the effect of a joint intervention of $\{X, M\}$ on Y. We provide further discussion on the algebraic and graphical assumptions needed for the identification of LATE that addresses the scenarios in Fig. 1c and 2b in Appendix C.2.

Local Effects and Mediation Extensions of LATE have led to concepts like the Local Natural Indirect Effect (LNIE) and Local Natural Direct Effect (LNDE) (Yamamoto 2013). LNIE captures the part of the average treatment effect due to the mediator within the subpopulation of compliers, while LNDE represents the portion not attributable to the mediator. They are defined as follows:

$$\text{LNIE}(x) := \mathbb{E}[Y_{x,M_{x_1}} - Y_{x,M_{x_0}} \mid X_{z_0} = 0, X_{z_1} = 1]$$
$$\text{LNDE}(x) := \mathbb{E}[Y_{x_1,M_x} - Y_{x_0,M_x} \mid X_{z_0} = 0, X_{z_1} = 1]$$

It has been shown that under certain conditions, LNIE and LNDE could be estimated from the observational distribution (Yamamoto 2013). However, these assumptions are non-trivial to check in practice from observational data without the aid of a graphical structure, as many of them are independence relations in Layer 2 and 3 of PCH. In addition, they may limit the applicability to many practical scenarios. Consider the graph in Fig. 1c and 2b, which does not satisfy some of their assumptions, including *exclusion restriction* and *conditionally ignorable treatment assignment*. However, the LNDE is identifiable by the use of MRL and the graphical properties of the causal diagram.

Proposition 4 (LNDE/LNIE Extensions)**.** *LNDE and LNIE are identifiable in the graph in Fig. 2b with local monotonicity of $Z \to X$, where X is binary. In particular, LNDE(x_0) can be computed as:*

$$\frac{\sum_{w,y} y \cdot T_1(y,w)P(w)}{\sum_{w,y} T_1(y,w)P(w)} - \frac{\sum_{w,y} y \cdot T_0(y,w)P(w)}{\sum_{w,y} T_0(y,w)P(w)} \quad (18)$$

where the weights $T_i(y,w)$ can be computed as

$$\sum_{m,z} P(m \mid w, z, x_i)P(z \mid w) \Big[P(y \mid w, m, x_i, z_i)P(x_i \mid w, z_i) - P(y \mid w, m, x_i, z_{1-i})P(x_i \mid w, z_{1-i})\Big]. \quad (19)$$

The LNIE(x_0) can be computed similarly. The addition of any directed or bidirected edge to the graph makes these quantities non-ID.

Also, it should be noted that LNDE and LNIE are also identifiable in the causal graph in Fig. 1c. For further discussion on LNDE/LNIE identification, refer to Appendix C.

2.2 Queries with Post-Treatment Conditioning

In this section, we demonstrate that certain queries with post-treatment conditioning, though generally non-identifiable, can be identified under specific monotonicity assumptions. Conditioning on post-treatment variables involves computing the effect of a treatment given any of its descendants (including the treatment itself). Examples include *probability of necessity* (PN) and *probability of sufficiency* (PS).

$$\text{PN} := P(Y_{x_0} = 0 \mid x_1, y_1) \quad (20)$$
$$\text{PS} := P(Y_{x_1} = 1 \mid x_0, y_0) \quad (21)$$

Here, we identify the local monotonicity constraints needed for identifying PN and PS in graph 1d. By Lem. 1 and (Pearl 2022), we can show that these quantities are identifiable with monotonicity on $X \to Y, X \to M, M \to Y$. If either of these edges is non-monotonic, then PN/PS are non-ID. PN/PS are not the only quantities of interest with post-treatment. These quantities have been studied in several areas, including the study of fairness, in particular, V-specific effects in (Plečko and Bareinboim 2024), analyzing dangers of post-treatment bias in designing experiments for political and social science (Montgomery, Nyhan, and Torres 2018) and mitigating post-treatment bias (Blackwell et al. 2023).

Consider the standard fairness graph in Fig. 1. Let X denote the sex of the job applicant, M their PhD status, and Y the hiring decision. We might be interested in how sex influences hiring, given that the applicant has a PhD. The M-specific effect, m_1-TE$_{x_0,x_1}(y)$ can then be written as

$$m_1\text{-TE} := \mathbb{E}[Y_{x_1} - Y_{x_0} \mid m_1] \quad (22)$$

However, this quantity is not identifiable from observational distribution. Interestingly, if $X \to M$ is monotonic, we can identify m_1-TE from observational distribution. We now present the following proposition for identifying queries involving post-treatment conditioning in a causal graph:

Proposition 5 (Generalized Post-Treatment Conditioning)**.** *In the causal diagram in Fig. 1d, the following holds for any set of values y, m to Y, M and x, x' to X:*

1. *If $X \to M$ is monotonic and M is binary, then $P(Y_x \mid m, x')$ and $P(Y_x \mid m)$ are ID.*
2. *If $X \to M, X \to Y, M \to Y$ are monotonic and M, Y are binary, $P(Y_x \mid x', m, y)$ and $P(Y_x \mid x', y)$ are ID.*

If either of the required edges is non-monotonic, then the effects are non-ID whenever $x \neq x'$.

3 M-ID: Algorithmic Identification of Arbitrary Counterfactual Quantities

In previous sections, we discussed how monotonicity constraints aid in identifying well-studied counterfactual queries. However, many graphical structures and counterfactual queries remain unexplored. In this section, we propose an algorithm that identifies arbitrary counterfactual quantities from interventional and observational distributions, given a causal graph with monotonicity assumptions. Our approach extends the algorithm from (Correa, Lee, and Bareinboim 2021) to account for monotonicity constraints.

The algorithm for deriving the ID expression of a counterfactual quantity, given a CGMA, is shown in Algorithm

Algorithm 1: M-ID

Input: Causal graph with monotonicity constraints $\langle G, M \rangle$, set of counterfactual terms $\mathbf{X}_* = \mathbf{x}_*, \mathbf{Y}_* = \mathbf{y}_*$, available distributions $\mathbb{Z}$.
Output: $P(\mathbf{Y}_* = \mathbf{y}_* \mid \mathbf{X}_* = \mathbf{x}_*)$ in terms of available distributions

1: $\mathbf{d}_*, D : P(\mathbf{W}_* = \mathbf{w}_*) \leftarrow$ CTF-FACTOR$(G, \mathbf{Y}_* = \mathbf{y}_*, \mathbf{X}_* = \mathbf{x}_*)$
2: $\mathbf{v}_*, \mathbf{Q} \leftarrow$ M-REDUCE$(\mathbf{W}_* = \mathbf{w}_*, \mathbf{x}_* \cup \mathbf{y}_*, G, M)$
3: **for** $s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}$ **do**
4: $\quad \mathbf{C}_{i*} = \mathbf{c}_{i*}\ (i \in [k]) \leftarrow$ CTF-FACTORIZE$(\mathbf{T}_* = \mathbf{t}_*, G)$
5: $\quad$ **for** each $\mathbf{C}_i$ **do**
6: $\quad\quad P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i) \leftarrow$ IDENTIFY$(\mathbf{C}_i, G, \mathbb{Z})$
7: $\quad\quad P(\mathbf{C}_{i*} = \mathbf{c}_{i*}) \leftarrow P_{\mathbf{V} \setminus \mathbf{C}_i}(\mathbf{C}_i)$
8: $\quad$ **end for**
9: $\quad P(\mathbf{T}_* = \mathbf{t}_*) = \prod_i P(\mathbf{C}_{i*} = \mathbf{c}_{i*})$
10: **end for**
11: **if** M-REDUCE or IDENTIFY fails, **return** FAIL
12: $D = \sum_{\mathbf{w}_* \setminus \mathbf{v}_*} \sum_{s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}} s \cdot P(\mathbf{T}_* = \mathbf{t}_*)$
13: **return** $\sum_{\mathbf{d}_* \setminus (\mathbf{x}_* \cup \mathbf{y}_*)} D / \sum_{\mathbf{d}_* \setminus \mathbf{x}_*} D$

1. We also assume that the domain of the variables that result in inconsistency is binary. Given a conditional counterfactual query, first, M-ID obtains the ctf-factors that need to be computed using CTF-FACTOR (Line 1). Then, it reduces each ctf-factor using MRL (Line 2) if needed. For each of the reduced factors, M-ID factorizes them using CTF-FACTORIZE, based on the c-components (Line 4) in $G[V(\mathbf{W}_*)]$, which is the subgraph containing variables in $\mathbf{W}_*$. If these factors are not inconsistent, they can be written as interventional quantities, which can then be identified using IDENTIFY (Line 6), adapted from (Tian and Pearl 2002). The details of CTF-FACTOR, CTF-FACTORIZE, and IDENTIFY are provided in Appendix C.3. We now make the following claim about M-ID.

Theorem 1. M-ID *is sound in identifying a counterfactual query in terms of available interventional and observational distributions, given a causal diagram and a set of monotonicity constraints.*

3.1 Monotonicity Reduction Algorithm

MRL can be applied to the ctf-factor in order to obtain a linear combination of several simplified ctf-factors. This idea is realized in Algorithm 2. The first step of M-REDUCE rewrites a variable that causes inconsistency as a summation over its domain, where the domain of V is denoted by $D(V)$ in Line 4. Then, on Line 8, it checks for any impossibility imposed by the monotonicity constraint. An impossible term is one for which the probability of it happening is 0. The conditions for impossibility are given as follows:

Definition 3 (Impossible ctf-factor)**.** *A ctf-factor is impossible if either of the following conditions holds:*

1. *There exists $w, w' \in \mathbf{w}_*$ corresponding to the same variable $W_\mathbf{t}$ and $w \neq w'$,*

Algorithm 2: M-REDUCE

Input: Ctf-factor $\mathbf{W}_* = \mathbf{w}_*$, assignments $\mathbf{Y}_* = \mathbf{y}_*$, causal graph G, monotonicity constraints M.
Output: $\mathbf{Q}$, a list of ctf-factors along with their signs.
Initialize: $\mathbf{Q} = \{+1, \mathbf{W}_* = \mathbf{w}_*\}, \mathbf{z}_* = \mathbf{y}_*, \mathbf{V} = \mathbf{V}(\mathbf{W}_*)$

1: $\mathbf{C}_{i*} = \mathbf{c}_{i*}\ (i \in [k]) \leftarrow$ FACTORIZE$(\mathbf{T}_* = \mathbf{t}_*, G)$
2: **for** $V \in \mathbf{V}$ and $s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}$ **do**
3: $\quad$ **if** $V \in C_i \setminus Y$ and causes inconsistency in $\mathbf{C}_{j*} = \mathbf{c}_{j*}$ for any j **then**
4: $\quad\quad$ Replace $(s, \mathbf{T}_* = \mathbf{t}_*)$ in $\mathbf{Q}$ with $(s, \mathbf{T}_* = \mathbf{t}_*(v))$ for all $v \in D(V)$ and update $\mathbf{z}_* = \mathbf{z}_* \cup D(V)$
5: $\quad$ **end if**
6: **end for**
7: **for** $s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}$ **do**
8: $\quad$ **if** $\mathbf{T}_* = \mathbf{t}_*$ is impossible, remove item from $\mathbf{Q}$
9: $\quad$ Apply Simplification Rule with M for all variables
10: $\quad$ **if** any condition of Lem. 6 holds, **return** FAIL.
11: $\quad$ **if** Common-Parent inconsistency exists for two variables in the same C_i for any i, **return** FAIL
12: **end for**
13: **for** V in C_i in reverse topological order in G **do**
14: $\quad$ **for** $s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}$ **do**
15: $\quad\quad$ **if** the conditions of Difference Rule are not applicable, **continue**
16: $\quad\quad$ **if** there exists $V_1 \in Ch(V), \mathbf{V}_{1\mathbf{t}_1} \in \mathbf{T}_*$, such that V, V_1 belongs to same c-component and $v \in \mathbf{t}_1$, apply Eq. 11 if $v = 1$ and Eq. 12 if $v = 0$ (if no such child exists apply either Eq. 11 or 12) to get $P(\mathbf{T}_* = \mathbf{t}_*) = P(\mathbf{T}'_* = \mathbf{t}'_*) - P(\mathbf{T}''_* = \mathbf{t}''_*)$
17: $\quad\quad$ **if** Difference Rule cannot be applied, **return** FAIL
18: $\quad\quad$ Replace $(s, \mathbf{T}_* = \mathbf{t}_*)$ with $s, \mathbf{T}'_* = \mathbf{t}'_*$ and $(-s, \mathbf{T}''_* = \mathbf{t}''_*)$
19: $\quad$ **end for**
20: **end for**
21: **for** $s, \mathbf{T}_* = \mathbf{t}_* \in \mathbf{Q}$ **do**
22: $\quad$ **if** any factors in the ctf-factorization of $\mathbf{T}_* = \mathbf{t}_*$ is inconsistent, **return** FAIL.
23: **end for**

2. *There exists $W_{\mathbf{t_1},\mathbf{s}} = w_1, W_{\mathbf{t_2},\mathbf{s}} = w_2 \in \mathbf{W}_* = \mathbf{w}_*$, such that $\mathbf{t_1} < \mathbf{t_2}$ and $w_1 > w_2$, where $\mathbf{T}$ and $\mathbf{S}$ are the set of monotonic and non-monotonic parents of W.*

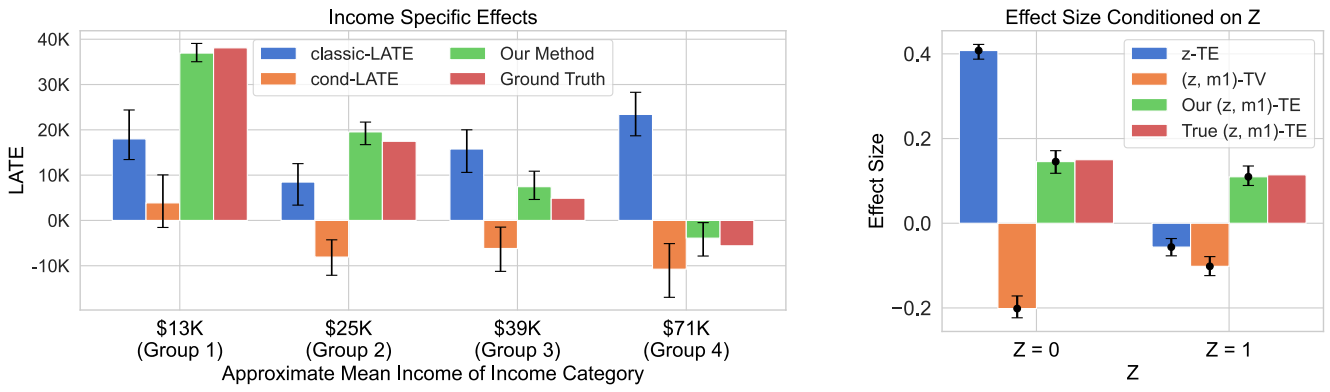
Once impossible terms have been removed, the algorithm applies Rule 1 of Lem. 2 in Line 9. After simplifying the ctf-factors, M-REDUCE checks the non-identification of a ctf-factor through the conditions of the following Lemma.

Lemma 6. *After repeated application of Rule 1 from Lem. 2, if there exists i, j ($i \neq j$) in the ctf-factor $P(\mathbf{Y}_*, W_{\mathbf{t}_1} = w_1, W_{\mathbf{t}_2} = w_2, \ldots, W_{\mathbf{t}_m} = w_m)$ such that either of the following holds:*

1. *$t \in \mathbf{t}_i, t' \in \mathbf{t}_j, t \neq t'$ for a non-monotonic parent T,*
2. *There is no total ordering between $\mathbf{t}_i, \mathbf{t}_j$,*

then the ctf-factor is non-ID.

If conditions of Lem. 6 are satisfied, the ctf-factor is immediately non-ID. Otherwise, if *Common-Parent Inconsis-*



(a) (b)

Figure 3: Experiments comparing the newly proposed method versus different baselines. (a) LATE computed in dollars on the 401(k) dataset, using the causal diagram in Fig. 1c, as discussed in Ex. 1. (b) Effects with post-treatment conditioning, based on simulations using the causal diagram in Fig. 1d. The ground truth is shown in red, the new method in green, and the baselines in blue/orange.

tency exists for two variables in the same c-component of $G[V(\mathbf{W}_*)]$, then the ctf-factor is also non-ID (Line 11). After that, M-REDUCE applies the Difference Rule of Lemma 2 in the reverse topological order so that the application does not result in any inconsistency. If at any point the condition of the Rule 2 is satisfied, but the rule cannot be applied because it will result in inconsistency, the ctf-factor is non-ID (Line 23). Finally, consider the following proposition.

Theorem 2. *If* M-REDUCE *returns* `FAIL` *on a ctf-factor, then the ctf-factor is non-ID.*

This result shows that the algorithm is complete in removing inconsistencies from a ctf-factor, or in other words, reducing a ctf-factor to a linear combination of interventional terms.

4 Experiments

In this section, we demonstrate how our method can be used in practice to identify counterfactual quantities that would otherwise be impossible to compute, and how seemingly natural choices can often lead to incorrect conclusions.

4.1 401(k) Dataset

In this section, we illustrate how naive estimation of local effects without considering graphical constraints can lead to misleading conclusions in real-world data, a problem our method effectively addresses.

To demonstrate this, we use the 401(k) dataset, which is a sample of financial data of individuals drawn from the 1991 Survey of Income and Program Participation (SIPP). The dataset comprises 9,915 observations and includes variables such as income (W), 401(k) eligibility (Z), 401(k) participation (X), net financial assets (M), and total wealth (Y). As discussed in Ex. 1, there are many unobserved confoundings, such as saving proclivity between X and M, and occupation between Z, W, and Y. The causal graph is shown in Fig. 1c. (Abadie 2003; Chernozhukov et al. 2018) have studied the LATE of 401(k) participation on net financial assets by using eligibility as an instrument. Here, we look at a more general setting (in terms of the graph and variables) and want to evaluate the LATE of participation in 401(k) (X) on total wealth (Y) for different income groups (W). Note that Z cannot be used as an instrument due to potential unobserved confounding between Z and Y, and any such attempt to ignore the causal graph is doomed to fail.

For the purpose of analysis, we discretize income, net financial assets, and total wealth into groups corresponding to quartiles, with each group represented by its mean value. Using these discretized data, we design a synthetic SCM M that matches the observational distribution of the original data. The causal graph corresponding to M is shown in Fig. 1c. We then generate 30,000 data points from this model to estimate LATE, classic-LATE, and cond-LATE. The expressions for these estimators are detailed in Ex. 1.

The bar plots of the values for the four income groups are shown in Fig. 3a. It shows that classic-LATE and cond-LATE may fail to estimate the true quantity. For example, in the low-income quartile, both methods fail to approximate the true effects by a large margin. In middle-income quartiles, conditional LATE concludes a negative impact of participation in 401(k) on total wealth despite the ground truth indicating a positive effect. On the other hand, in the high-income quartile, the classic-LATE overestimates a high positive effect, while in reality, the effect is a little negative. In contrast, our method consistently provides an estimate well-aligned with the ground truth across all income groups.

4.2 Fair Machine Learning Application

In this part, we discuss an example in the context of fair machine learning. Consider the causal graph from Fig. 1d with binary variables Z (age, 0 old, 1 young), X (sex, 0 female, 1 male), M (education, 0 low education, 1 high) and Y (income, 0 low income, 1 high). The functions and distribution in the SCM are given as follows, where U_m^1, U_m^0 are ternary

	x_0, m_0	x_0, m_1	x_1, m_0	x_1, m_1
z_0	0.1	0.8	0.9	0.6
z_1	0.4	0.9	0.1	0.8

Table 1: Distribution of $P(Y = 1 \mid z, x, m)$ in Sec. 4.2

with values from $\{a, c, n\}$. $P(Y = 1 \mid z, x, m)$ for different values of z, x and m are shown in Table 1.

$$Z = U_Z; X = U_X, \quad P(U_X = 1) = P(U_Z = 1) = 0.5$$
$$M = \begin{cases} \mathbb{1}\{U_m^1 = a\} + X \cdot \mathbb{1}\{U_m^1 = c\} & \text{if } Z = 1 \\ \mathbb{1}\{U_m^0 = a\} + X \cdot \mathbb{1}\{U_m^0 = c\} & \text{if } Z = 0 \end{cases}$$

$$P(U_m^1 = a) = 0.5, \qquad P(U_m^1 = c) = 0.3$$
$$P(U_m^0 = a) = 0.25, \qquad P(U_m^0 = c) = 0.5$$

We are interested in understanding the total causal effect of X on Y, for various subgroups of the population. We begin by computing the total effect (TE), written $\mathbb{E}[Y_{x_1} - Y_{x_0}]$, which measures the average effect of changing $x_0 \to x_1$ (female to male) across all individuals in the population. We find that TE = 0.175, which means that being male causally increases the income in the population. We then wish to look into different age groups to understand if Z modifies the effect, by computing z-specific total effect

$$z\text{-TE}_{x_0,x_1}(y) := \mathbb{E}[Y_{x_1} \mid z] - \mathbb{E}[Y_{x_0} \mid z] \quad (23)$$
$$= \mathbb{E}[Y \mid x_1, z] - \mathbb{E}[Y \mid x_0, z]. \quad (24)$$

This allows us to quantify discrimination separately for young and old populations. Furthermore, we believe that possible discrimination may be specifically different for highly educated individuals (in each age group), and we are thus also interested in computing the counterfactual quantity given by the following (z, m)-specific total effect (Plečko and Bareinboim 2024):

$$(z, m)\text{-TE}_{x_0,x_1}(y) = \mathbb{E}[Y_{x_1} \mid z, m] - \mathbb{E}[Y_{x_0} \mid z, m] \quad (25)$$

Note that this corresponds to a counterfactual question: "For a person of fixed age and education level, how would their income change if X had been equal to male, compared to had X been equal to female?" Notably, in the absence of monotonicity, this effect is not identifiable since it involves post-treatment conditioning (M comes after X causally), yet when assuming $X \to M$ monotonicity, we can recover this term.

The reader may be tempted to use the estimator

$$(z, m)\text{-TV}_{x_0,x_1}(y) := \mathbb{E}[Y \mid x_1, z, m] - \mathbb{E}[Y \mid x_0, z, m]$$

in place of (z, m)-$\text{TE}_{x_0,x_1}(y)$. However, the (z, m)-TV quantity does not equal the (z, m)-$\text{TE}_{x_0,x_1}(y)$ effect, and using it may lead to incorrect conclusions (in fact, in the graph in Fig. 1, (z, m)-TV is a measure of direct effect, not total causal effect, which also includes the indirect effect $X \to M \to Y$).

For the experiment, we sample 10000 data points from the given SCM and obtain the empirical values for the 3 quantities. Along with these estimates, we also show the ground truth (z, m)-TE as obtained from the distribution of the SCM. We also show a 95% confidence interval with bootstrapping. We present our results in Fig. 3b. More details about both experiments are in Appendix E.

5 Conclusions

This work represents a significant step towards encoding and utilizing local monotonicity constraints in a general way in a causal model. In Sec. 2, we introduced the Monotonicity Reduction Lemma that helps simplify inconsistent ctf-factors to consistent ctf-factors. We then show how this result can be used to identify local effects, such as LATE, LNDE/LNIE (Sec. 2.1), and effects with post-treatment conditioning (Sec. 2.2). In Sec. 3, we developed an efficient algorithm for identifying counterfactual quantities under monotonicity constraints (M-ID). Generalizing the standard setting found in the literature based on IVs, this algorithm takes as input an arbitrary counterfactual quantity and causal model and returns an identification expression in terms of available observational and interventional distributions. Finding a complete (sufficient and necessary) algorithm for general identification under monotonicity constraints is an important, open research question. Future works can explore monotonicity constraints for continuous variables and also additional shape constraints such as convexity and concavity to further our identification toolbox and bridge the gap. We hope this paper highlights the versatility of the graphical approach of causality in accommodating different types of shape constraints, including highly general models such as non-parametric ones, as well as linearity, additivity, and now monotonicity. This, of course, generalizes more popular and well-studied models found in the literature, including the back-door/ignorability, front-door, napkin, and instrumental variable graphs.

Acknowledgements

This research is supported in part by the NSF, DARPA, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

Abadie, A. 2003. Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2): 231–263.

Angrist, J. D.; and Imbens, G. W. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, 90(430): 431–442.

Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. 2022. On Pearl's Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 507–556. New York, NY, USA: Association for Computing Machinery, 1st edition.

Bareinboim, E.; and Pearl, J. 2012. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 113–120.

Bareinboim, E.; and Pearl, J. 2016. Causal Inference and The Data-Fusion Problem. In Shiffrin, R. M., ed., *Proceedings of the National Academy of Sciences*, volume 113, 7345–7352. National Academy of Sciences.

Blackwell, M.; Brown, J. R.; Hill, S.; Imai, K.; and Yamamoto, T. 2023. Priming bias versus post-treatment bias in experimental designs. *arXiv preprint arXiv:2306.01211*.

Brito, C.; and Pearl, J. 2002. A Graphical Criterion for the Identification of Causal Effects in Linear Models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 2, 533–540.

Chen, B.; Kumor, D.; and Bareinboim, E. 2017. Identification and Model Testing in Linear Structural Equation Models using Auxiliary Variables. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1156–1164. International Convention Centre, Sydney, Australia: PMLR.

Chen, B.; Pearl, J.; and Bareinboim, E. 2016. Incorporating Knowledge into Structural Equation Models using Auxiliary Variables. In Kambhampati, S., ed., *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 3577–3583. International Joint Conferences on Artificial Intelligence Organization.

Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; and Robins, J. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1): C1–C68.

Correa, J.; and Bareinboim, E. 2017. Causal Effect Identification by Adjustment under Confounding and Selection Biases. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 3740–3746. San Francisco, CA: AAAI Press.

Correa, J.; and Bareinboim, E. 2024. Counterfactual Graphical Models: Constraints and Inference. Technical Report R-115, Causal Artificial Intelligence Lab, Columbia University.

Correa, J.; Lee, S.; and Bareinboim, E. 2021. Nested counterfactual identification from arbitrary surrogate experiments. *Advances in Neural Information Processing Systems*, 34: 6856–6867.

Correa, J. D.; and Bareinboim, E. 2020. A Calculus for Stochastic Interventions: Causal Effect Identification and Surrogate Experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. AAAI Press.

Frölich, M. 2007. Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1): 35–75.

Galton, F. 1886. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15: 246–263.

Gauss, C. F. 1877. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, volume 7. FA Perthes.

Heckman, J. J.; Urzua, S.; and Vytlacil, E. 2006. Understanding instrumental variables in models with essential heterogeneity. *The review of economics and statistics*, 88(3): 389–432.

Huang, Y.; and Valtorta, M. 2006. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, 217–224. Arlington, Virginia, USA: AUAI Press. ISBN 0974903922.

Imbens, G. W. 2020. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4): 1129–1179.

Imbens, G. W.; and Angrist, J. D. 1994. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2): 467–475.

Kumor, D.; Chen, B.; and Bareinboim, E. 2019. Efficient Identification in Linear Structural Causal Models with Instrumental Cutsets. In *Advances in Neural Information Processing Systems*, volume 32.

Kumor, D.; Cinelli, C.; and Bareinboim, E. 2020. Efficient identification in linear structural causal models with auxiliary cutsets. In *International Conference on Machine Learning*, 5501–5510. PMLR.

Lee, S.; and Bareinboim, E. 2020. Causal Effect Identifiability under Partial-Observability. In *Proceedings of the 37th International Conference on Machine Learning (ICML-20)*.

Lee, S.; Correa, J. D.; and Bareinboim, E. 2019. General Identifiability with Arbitrary Surrogate Experiments. In *In Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*.

Lee, S.; Correa, J. D.; and Bareinboim, E. 2020. General identifiability with arbitrary surrogate experiments. In *Uncertainty in artificial intelligence*, 389–398. PMLR.

Mogstad, M.; Torgovitsky, A.; and Walters, C. R. 2019. *Identification of causal effects with multiple instruments: Problems and some solutions*. National Bureau of Economic Research.

Montgomery, J. M.; Nyhan, B.; and Torres, M. 2018. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3): 760–775.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688.

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

Pearl, J. 2022. Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 317–372.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Peters, J.; Janzing, D.; and Scholkopf, B. 2011. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12): 2436–2450.

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

Reiersøl, O. 1945. *Confluence analysis by means of instrumental sets of variables*. Ph.D. thesis, Almqvist & Wiksell.

Shimizu, S. 2014. LiNGAM: Non-Gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1): 65–98.

Shpitser, I.; and Pearl, J. 2007. What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, 352–359. Arlington, Virginia, USA: AUAI Press. ISBN 0974903930.

Starr, W. 2019. Counterfactuals. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2019 edition.

Tian, J. 2004. Identifying Linear Causal Effects. In *Proceedings of the National Conference on Artificial Intelligence*, volume 17, 346–353.

Tian, J. 2005. Identifying Direct Causal Effects in Linear Models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, 346–353.

Tian, J.; and Pearl, J. 2002. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, 567–573. USA: American Association for Artificial Intelligence. ISBN 0262511290.

Tian, J.; and Shpitser, I. 2010. On identifying causal effects. *Heuristics, Probability and Causality: A Tribute to Judea Pearl (R. Dechter, H. Geffner and J. Halpern, eds.). College Publications, UK*, 415–444.

Van Hoeck, N.; Watson, P. D.; and Barbey, A. K. 2015. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in human neuroscience*, 9: 420.

VanderWeele, T. J.; and Robins, J. M. 2010. Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(1): 111–127.

Wright, P. G. 1928. *The tariff on animal and vegetable oils*. 26. Macmillan.

Yamamoto, T. 2013. Identification and estimation of causal mediation effects with treatment noncompliance. *Unpublished manuscript*.